NUMERICAL SOLUTION OF HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

by

Rosemary Anne Williamson

Churchill College, Cambridge

A dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

December 1984

Numerical Solution of Hyperbolic Partial Differential Equations

by

Rosemary Anne Williamson

An investigation of numerical techniques for solving hyperbolic partial differential equations is presented. The relevance of the terms convergence, stability, consistency and order of accuracy as applied to numerical models solving the linear equation $u_t = Lu$, where L is a linear differential operator, is discussed. With the Lax Richtmyer Equivalence theorem in mind, a review of the stability analysis for solving these equations with and without boundaries is presented.

In this dissertation we are mainly concerned with the method of lines approach for solving such equations. The semidiscretisation

$$\sum_{j=-R}^{S} f_j \frac{d}{dt} v_j(t) = \frac{1}{\Delta x} \sum_{j=-r}^{s} g_j v_j(t)$$

is adopted as an approximation to the linear equation $u_t = u_x$. By a modification of the theory of order stars the maximal accuracy of stable schemes of this kind is proved to satisfy the bound

$$p \leq \min \{ r+s+R+S , 2(r+R+1) , 2(s+S) \} .$$

By an application of the Padé theory we prove that in some cases schemes achieving this bound do exist.

We propose solving the ordinary differential system of equations obtained from the semidiscretisation by a class of two-step Runge Kutta schemes. This particular class of methods is designed to have the same number of function evaluations as a one-step method whilst obtaining the degrees of freedom associated with a two-step method. A strategy for exploiting these degrees of freedom to develop maximally efficient second- and third-order schemes compatible with the underlying semidiscretisation is described. We also discuss the design of efficient implementations for both one- and two-step Runge-Kutta schemes.

A comparison of the characteristics of conservative and dissipative semidiscretisations is performed by studying their evolution of three initial conditions under time integration by Runge-Kutta methods and the trapezoidal and midpoint rules. A group velocity analysis of the spatial discretisations is presented and it is demonstrated that for reasonable time steps the influence of the time integration is negligible. Furthermore, very good shock resolution is achieved by dissipative methods and conservative methods integrated by particular Runge-Kutta schemes. Consequently the latter models are very good candidates for inclusion in codes for solving non-linear conservation laws efficiently.

Numerical Solution of Hyperbolic Partial Differential Equations

by

Rosemary Anne Williamson

An investigation of numerical techniques for solving hyperbolic partial differential equations is presented. The relevance of the terms convergence, stability, consistency and order of accuracy as applied to numerical models solving the linear equation $u_t = Lu$, where L is a linear differential operator, is discussed. With the Lax Richtmyer Equivalence theorem in mind, a review of the stability analysis for solving these equations with and without boundaries is presented.

In this dissertation we are mainly concerned with the method of lines approach for solving such equations. The semidiscretisation

$$\sum_{j=-R}^{S} f_j \frac{d}{dt} v_j(t) = \frac{1}{\Delta x} \sum_{j=-r}^{s} g_j v_j(t)$$

is adopted as an approximation to the linear equation $u_t = u_x$. By a modification of the theory of order stars the maximal accuracy of stable schemes of this kind is proved to satisfy the bound

$$p \leq \min \{ r+s+R+S , 2(r+R+1) , 2(s+S) \} .$$

By an application of the Padé theory we prove that in some cases schemes achieving this bound do exist.

We propose solving the ordinary differential system of equations obtained from the semidiscretisation by a class of two-step Runge Kutta schemes. This particular class of methods is designed to have the same number of function evaluations as a one-step method whilst obtaining the degrees of freedom associated with a two-step method. A strategy for exploiting these degrees of freedom to develop maximally efficient second- and third-order schemes compatible with the underlying semidiscretisation is described. We also discuss the design of efficient implementations for both one- and two-step Runge-Kutta schemes.

A comparison of the characteristics of conservative and dissipative semidiscretisations is performed by studying their evolution of three initial conditions under time integration by Runge-Kutta methods and the trapezoidal and midpoint rules. A group velocity analysis of the spatial discretisations is presented and it is demonstrated that for reasonable time steps the influence of the time integration is negligible. Furthermore, very good shock resolution is achieved by dissipative methods and conservative methods integrated by particular Runge-Kutta schemes. Consequently the latter models are very good candidates for inclusion in codes for solving non-linear conservation laws efficiently.

PREFACE

The work described in this dissertation is believed to be original except where reference is explicitly made to the work of others. None of it has been or is concurrently been submitted for any degree, diploma or other qualification at any other university. The work described in Chapters Two through Four was carried out in collaboration with my supervisor, Dr. A. Iserles, and is presented with the sanction of the Board of Graduate Studies and the Degree Committee.

This dissertation describes work carried out between January 1982 and December 1984 in the Department of Applied Mathematics and Theoretical Physics at the University of Cambridge. Throughout this period my supervisor was Dr. A. Iserles and I should like to express my deep gratitude to him for his continuing encouragement and advice.

I would like to acknowledge the interest and helpful remarks of Professor M.J.D. Powell concerning the work presented in Chapters Two through Four and in particular his suggestion to use the separation theorem in the proof of Theorem 4.2.3. I also want to thank the Science and Engineering Research Council for a three-year research studentship and Churchill College and D.A.M.T.P. for assistance with travelling expenses.

Finally, I wish to express my sincere thanks to Mrs. Naomi Coyle for typing this dissertation, and to my husband for endless sympathy and support from afar.

C O N T E N T S

0. INTRODUCTION

In this dissertation we present an investigation of numerical models for the solution of *hyperbolic partial differential equations*. The particular models considered are those where the equation is approximated by *finite differences*. Such models fall into one of two categories; either *fully-* or *semi-discretised* schemes. Here we concentrate on the latter group whereby spatial and time derivatives are considered separately. This approach to solving partial differential equations is often called the *method of lines*.

Any investigation of semi-discretised (SD) models comprises two parts; first to analyse the spatial discretisation and second to consider the time integration of the resulting system of ordinary differential equations (O.D.E.). Conveniently, these two problems divide this dissertation into two parts. In the first four chapters we consider the *stability* and *order of accuracy* of the SD and in the remaining chapters the solution of the O.D.E. system is discussed. A special class of two-step Runge-Kutta methods designed to have the same number of function evaluations as a one-step method is proposed.

Chapter 1: We begin with the first chapter by reviewing some of the literature on the solution of partial differential equations by finite-difference methods. Fundamental in the analysis of such methods is the determination of their *convergence* and *order of accuracy* properties. Here we

define conditions for convergence of both fully- and semi-discretised models for the solution of the linear equation $u_t = Lu$ where L is a linear spatial differential operator. By the Lax-Richtmyer Equivalence Theorems convergence may be investigated via the *stability* of the model [La56]. With this in mind, we discuss the development of stability theory from von Neumann's Fourier analysis [Ri67] to the Kreiss matrix theorem [Kr59] and the Godunov-Ryabenki analysis of normal modes [Go64]. This discussion of the more traditional view of stability is completed by describing the Wiener-Hopf factorisation techniques for implicit models [St64b], followed by the Gustafsson Kreiss and Sundström version of stability for mixed boundary-value problems [Gu72]. Further to this the more recent group velocity interpretation of stability by Trefethen is described [Tr82a].

Finally, we consider the stability of the O.D.E. system of equations and demonstrate that absolute stability of the O.D.E. solver is not sufficient to characterise the behaviour of numerical solutions for partial differential equations. On the contrary, the location of the spectrum of the infinite dimensional Toeplitz operator describing the SD system must lie within the absolute stability region of the O.D.E. method.

Chapter 2: Here we start by defining the *order of accuracy* of a numerical scheme. From this definition it is demonstrated that determination of stability may be posed in an

approximation-theoretic framework. Further, we demonstrate a relationship between all fully-discretised (FD) and some SD schemes. As a result of this the stability and accuracy of the associated SD are inherited from the FD. Hence the maximal order of accuracy of stable FD's may be investigated via the associated SD's. We also prove that the *pole condition* for stability of an implicit SD implies a condition on the location of the zeros of the characteristic function of the SD: For stability, the SD defined by

$$\sum_{j=-R}^{S} f_j \frac{d}{dt} v_j = \frac{1}{\Delta x} \sum_{j=-r}^{s} g_j v_j \quad ,$$

where $v_j(t)$ is an approximation to the exact solution $u(x_o + j\Delta x, t)$, has a characteristic function which must have at least r zeros inside the unit circle and $(s-1)$ zeros outside the unit circle. Consequently no scheme with $s = 0$ may be stable.

Chapter 3: In this chapter we prove the major result of the first half of this dissertation. The order of accuracy, p, of a stable semi-discretisation for solving the linear conservation law is bounded by

$$p \leq \min\{ r+s+R+S,\ 2(r+R+1),\ 2(s+S) \} \quad .$$

This generalises the equivalent result proved by Iserles and Strang [Is83a] for full discretisations. Initially we describe the theory of *order stars* as introduced by Wanner,

Hairer and Nørsett [Wa78], briefly discussing some of its more recent developments and applications. A proof of the given bound follows by incorporating the zero condition into the geometric properties of the order star and then applying a combinatorial argument to derive the optimal configuration of the poles and zeros of the characteristic function. In this way we also provide an alternative proof of the bound $2(s + S)$ for FD schemes. Moreover, we prove that Padé approximations to $z^L \ln z$ are normal for particular choices of r, R, s and S.

Chapter 4: In this chapter we prove that some schemes attaining the maximal accuracy, $p = r + R + s + S$, are stable. For these schemes the bound on accuracy means that they must be sufficiently *centred*:

$$r + R \leqslant s + S \leqslant r + R + 2 ,$$

which considerably reduces the range of explicit and implicit parts that may be incorporated in the model. Our analysis concentrates on schemes derived from the appropriate Padé approximations: the *Padé schemes*. For the associated SD's proof of stability follows directly from the results of Iserles and Strang [Is83a]. Otherwise we rely on the Padé theory to determine the sign of the error constant of the approximation from which we prove that the von Neumann condition is satisfied. In addition, the geometry of the order star is used to decide which configurations of poles and zeros are possible hence proving that the pole

condition is also satisfied. Consequently we are able to prove that, as for FD's, stability occurs for the approximations lying on the three leading diagonals of the Padé tableau.

Chapter 5: In this chapter we elaborate on the discussion in Chapter 1 of the *convergence* of the numerical solution of the O.D.E. system of equations arising from the spatial discretisation. We discuss some of the advantages and disadvantages of possible classes of O.D.E. methods for solving non-linear conservation laws, motivating our decision to investigate explicit methods.

In solving a partial differential equation by the method of lines, we require that the family of eigenvalue curves of the Jacobian matrices of the semi-discretised equation, obtained as the mesh is refined, lies within the absolute stability region of the O.D.E. method. A scheme which is maximally efficient has the largest possible multiple of the spectrum of the infinite dimensional Toeplitz operator describing the SD lying inside its stability region. These schemes maximise the Courant number that may be stably used in the numerical model. Moreover, due to the varying structure of the Jacobian matrix maximally efficient schemes for one class of SD methods will not be optimal for another class.

The development of optimal schemes for parabolic equations has been extensively investigated ([Ve76], [Ho80]). We review the results for these equations and also discuss

the optimal schemes for hyperbolic equations discretised using central differences. For the former case extended stability on the negative real axis, and in the latter extended stability on the imaginary axis, is required. There are very few results in the literature applicable to hyperbolic equations with dissipative discretisations. Such SD's require maximal stability within a closed region of $\mathbb{C}^-$ that adjoins the origin. However we do discuss the application of the Comparison Theorems as introduced by Jeltsch and Nevanlinna [Je82] as well as the possible relevance of the work described by Manteuffel on Tchebychev iterations [Ma77].

Chapter 6: Here we propose a class of two-step multistage methods for solving an O.D.E. system. This particular class of methods is designed to have the same number of function evaluations as a one-step method whilst retaining most of the degrees of freedom associated with a two-step method. A strategy for using these degrees of freedom to derive schemes with extended stability regions is described. Furthermore, the optimisation problem is solved for extended intervals of stability on the imaginary axis and for extended stability within wedge-shaped regions lying in $\mathbb{C}^-$. In this way schemes suitable for integrating either conservative or dissipative SD's of hyperbolic equations in an efficient manner are obtained. We also apply the optimisation technique to one-step Runge-Kutta methods and demonstrate that, at the cost of extra storage, the incorporation of the extra step is valuable for increased efficiency.

Consequently, this particular class of methods is useful for designing efficient methods not only for partial differential equations, but also for O.D.E.'s which require stability within wedges adjoining the origin in $\mathbb{C}^-$.

Chapter 7: Here a discussion of *efficient* algorithms with *minimal storage* requirements for both one- and two-step Runge-Kutta methods is presented. First we review algorithms for the one-step methods. It is demonstrated that, by considering a specially designed scheme, a fourth-order four-stage method may be implemented with just two arrays of storage. This is a considerable improvement on the four arrays of storage usually required to implement the one-step Runge-Kutta methods. Therefore we investigate whether the class of two-step methods introduced in Chapter 6 may be implemented in a comparable manner. We derive minimal-storage algorithms for two-step schemes by generalising the one-step implementations. We show that an algorithm requiring only two arrays of storage, whilst achieving third-order accuracy with three stages, is possible. Alternatively, with less restrictions on the integration parameters, algorithms requiring only three or four arrays of storage may be used.

Finally we discuss the implementation of an *error control* mechanism for one- and two-step schemes. We demonstrate that the algorithms which need more storage are more useful in this context. For these algorithms the schemes still have some degrees of freedom available that may be used to minimise the number of additional function evaluations. We

describe a possible method of error control similar to the Runge-Kutta-Fehlberg technique. Furthermore, an algorithm using five arrays of storage suitable for the two-stage methods of order two and the three-stage methods of order two and three derived in Chapter 6 is suggested.

Chapter 8: In this final chapter a *group velocity* analysis of SD schemes is presented. We derive expressions for the group velocity of both *conservative* and *dissipative* SD's, keeping in mind that the complex nature of the group velocity for the latter schemes means that its general physical relevance is not clear. We analyse the group velocity, phase velocity and amplitude of four SD's, two of which are dissipative. A comparison of the characteristics of the SD's is then performed by studying their evolution of three initial conditions with different methods of time integration.

The numerical models considered are based on the three-point and five-point central difference formulae, as well as two third-order dissipative schemes, with integration by the midpoint and trapezoidal rules and the one and two-step Runge Kutta methods obtained in Chapter 6. By choosing suitable initial conditions, a stepfunction, a pulse and a wavepacket, the dissipative and dispersive properties of the models are predicted. Integration of the dissipative schemes shows that these predictions are insufficient. A wavepacket evolves faster than group velocity would suggest but amplitude attenuation does occur at a rate predicted by a simple amplitude analysis. Clearly, *energy velocity* must also be

considered to obtain a complete picture for these models. However, group velocity predictions for the conservative SD's are very accurate. This demonstrates that for reasonable time steps the influence of the time integration on group velocity is negligible.

Further, not only is good shock resolution achieved by dissipative schemes, but also by conservative schemes integrated by particular Runge-Kutta methods. An extensive investigation of efficient Runge-Kutta methods for conservative SD's already exists in the literature. Additionally, the numerical development of these schemes is easier than for dissipative SD's. Therefore numerical models based on the combination of conservative SD's with Runge-Kutta methods are very strong candidates for inclusion in codes designed to solve non-linear problems efficiently.

1. CONVERGENCE AND STABILITY OF PARTIAL DIFFERENTIAL EQUATIONS

1.1 Nonlinear Equations

The solution of *nonlinear partial differential equations* by numerical methods poses many problems. For example, consider the *genuine solutions* of the *nonlinear conservation laws*,

$$u_t + (g(u))_x = 0 \quad , \quad u(x,0) = \phi(x) \quad , \qquad (1.1.1)$$

where ϕ is a smooth initial condition and g is some sufficiently smooth nonlinear function of $u = u(x,t)$. It is well known that the smoothness of $\phi(x)$ does not ensure a continuous solution $u(x,t)$ for all $t > 0$ [La60b]. Unique solutions of (1.1.1) exist only under the additional assumptions that all the discontinuities of $u(x,t)$ are *shocks*: at all discontinuities $u(x,t)$ must satisfy the *Rankine Hugoniot jump* condition and the *entropy* condition [La73]. The aim is to develop numerical models whose solutions *converge* to the real solution and mimic discontinuities correctly. This is, at present, an area of active investigation by many authors; see, for example, references [Le79], [Ro81], [En80], [Ha76], where work following along the lines of that initially presented by Lax and Wendroff [La60b] is discussed.

In some cases the solution of the nonlinear model can

be shown to be *stable* by applying the *energy method* directly [En80]. In most cases, however, non-linear stability is assumed to depend on the stability of the first variation in the difference operator. Strang [St64a] proved that this assumption is justified for systems with suitably smooth initial conditions solved by a consistent numerical model. The investigation of the linearised model is thus quite valid as a guide to the solution of nonlinear equations.

In this chapter the concepts of *convergence*, *stability*, *consistency* and *order of accuracy* are explained. Further, a brief review of the determination of convergence of the numerical model for various linear partial differential equations with and without boundary conditions is presented.

1.2 Convergence, Stability and the Lax-Richtmyer Equivalence Theorems

When solving any partial differential equation by means of a numerical model the major requirement is that the solution of the numerical scheme should *converge* in some sense to the genuine solution. For this to be possible the numerical model must be a *consistent* approximation to the partial differential equation, which in turn must be a *properly-posed* problem.

It is convenient to think of the variables describing the state of a system at fixed time t as elements u(t) of a Banach space B with a norm $\| \ \|$. The norm of an operator T

is defined in the usual way by

$$\| T \| := \sup \{ \frac{\| Tg \|}{\| g \|} \mid g \neq 0 \ , \ Tg \text{ exists } , \ g \in B \} \tag{1.2.1}$$

We denote the *linear spatial differential operator* acting on elements of B by L, where L is, for simplicity, assumed independent of t. Then the initial value problem (IVP) is to find a one-parameter set of elements u(t) such that

$$\begin{aligned} u_t &= L u + f(x,t) \quad t \geqslant 0 \\ u(x,0) &= \phi(x) \quad . \end{aligned} \tag{1.2.2}$$

If boundary conditions are present we assume that they are linear homogeneous and are incorporated by assuming that the domain of L is restricted to elements satisfying these conditions. In this case we require to solve the initial boundary value problem (IBVP).

A *genuine solution* of (1.2.2) is the one-parameter set u(t) such that

i) u(t) is in the domain of L for all t in the compact interval $[\,0, \tau\,]$ and

$$\text{ii)} \quad \lim_{\Delta t \to 0} \| \frac{u(t + \Delta t) - u(t)}{\Delta t} - (L\, u(t) + f(x,t)) \| \to 0 \tag{1.2.3}$$

uniformly in t, for all t in the compact interval $[\,0, \tau\,]$ [La56].

Alternatively if we pick an element ϕ which is not in the domain of L then obviously we can not always find a genuine solution satisfying the initial conditions. However we assume that we can always approximate the initial condition as closely as required by an element in the domain of L. Thus we assume that we can define an *evolution operator* $E_0(t)$ which has domain dense in B so that for any genuine solution $u(t)$ of (1.2.2) depending uniquely on $\phi(x)$

$$u(t) = E_0(t)\,\phi(x) + \int_0^t E_0(t-s)\,f(x,s)\,ds \qquad (1.2.4)$$

[Ri67].

In addition it is desirable that the solution $u(t)$ should depend continuously on the initial value $\phi(x)$. Thus we also assume that the operator $E_0(t)$ is uniformly bounded in any compact interval $[0, \tau]$ with respect to the operator norm on B as defined by (1.2.1). These two assumptions characterise a *properly posed* problem according to the notion of Hadamard [La56].

These definitions mean that genuine solutions of properly posed problems are continuous and differentiable. However, such solutions need not exist: recall that solutions of non-linear conservations laws may be discontinuous [La73]. But as the evolution operator $E_0(t)$ is bounded with domain dense in B it has a unique bounded linear extension $E(t)$ whose domain is the entire space B and whose bound is the same as that of $E_0(t)$ [Ri67]. Thus for any properly posed problem and for arbitrary $\phi \in B$ we can interpret the

one-parameter set of elements of $u(t) \in B$ given by

$$u(t) = E(t)\,\phi(x) + \int_0^t E(t-s)\,f(x,s)\,ds \quad , \tag{1.2.5}$$

as the *generalised or weak solutions* of the IBVP. $E(t)$ is called the *generalised evolution operator*.

By (1.2.3) any genuine solution of (1.2.2) is continuous with respect to the norm of B. Furthermore, the operator $E(t)$ acting on B, satisfies the semigroup property,

$$E(t + s) = E(t)E(s).$$

Thus, applying the triangle inequality to (1.2.3) it can be shown that not only are the integral forms of the generalised solutions continuous in any compact interval $[0, \tau]$ but also that their convergence in (1.2.3) is uniform in t, $t \in [0, \tau]$, for a properly posed IBVP.

Since for a properly posed IBVP (1.2.2) the evolution operator is formally e^{Lt}, we can write the formal solution of (1.2.2) as

$$u(t) = e^{Lt}\phi(x) + \int_0^t e^{L(t-s)}\,f(x,s)\,ds \quad , \tag{1.2.6}$$

provided that f, Lf and $L^2 f$ exist and are continuous functions of t for all $t \geq 0$ [Ri67]. Henceforth we shall assume that ϕ is square integrable, since for the formal solution as

given this is sufficient.

In this dissertation we are mainly concerned with numerical models determined by finite-difference approximations to spatial-differential operators. This approach produces a system of ordinary differential equations that may then be integrated by an ordinary differential equation solver. The solution of partial differential equations in this way is often called the *method of lines*. We call the finite-difference approximation a *semi-discretisation* (SD) and abbreviate implicit and explicit schemes by ISD and ESD respectively.

Alternatively, both spatial and time derivatives may be approximated in unison by means of Taylor expansions. This produces a system of difference equations for the solution at one time level as a linear combination of solutions at preceding time levels. A numerical model obtained in this way is called a *full discretisation* (FD) and if the solution is determined by solutions at k previous levels, it is called a *k-step scheme*. Again, we abbreviate implicit and explicit FD by IFD and EFD respectively.

Assuming that $v(t)$ is an approximation to $u(t)$, $v_j(t) \approx u(j_1\Delta x_1, j_2\Delta x_2, \ldots, j_m\Delta x_m, t)$, j being a multi-index, we replace the differential equation (1.2.2) by the semi-discretisation;

$$\begin{aligned} v'(t) &= B(\Delta x)v(t) + F(t) \\ v(0) &= \Phi \quad . \end{aligned} \tag{1.2.7}$$

Here the difference operator B has matrix coefficients which are functions of the grid size $\Delta x = (\Delta x_1, \Delta x_2, \ldots, \Delta x_m)^T$. $F(t)$ and Φ are the projections of the continuous functions $f(x,t)$ and $\phi(x)$ onto the grid by the *projection operator* $P(\Delta x)$.

Consistency of the numerical model (1.2.7) is defined by requiring that the difference operator $B(\Delta x)$ approximates L uniformly for every genuine solution $u(x,t)$ and a set of initial conditions $\phi(x)$ dense within the set of square integrable functions:

$$\lim_{\Delta x \to 0} \| (P(\Delta x) L - B(\Delta x) P(\Delta x)) u(x,t) \| = 0 \qquad (1.2.8)$$

Consistency may easily be verified by expanding the product $B(\Delta x)\, v$ as a finite Taylor series. The *truncation error* involved in replacing $L u$ by this differential operator may be estimated by Taylor's theorem for sufficiently smooth functions. It can then be used to determine the convergence of the approximation to the real solution. With the same conditions as for consistency, the method is defined to be *convergent* if

$$\lim_{\Delta x \to 0} \| v(t) - P(\Delta x)\, u(t) \| = 0 \qquad (1.2.9)$$

for every t in a compact interval $[0, \tau]$.

Stability is a property of the difference approximation alone and limits the extent to which any component of the

initial function can be amplified by the numerical procedure. We say that the SD difference scheme (1.2.7) is *stable* if for every $\tau > 0$ there exists a positive constant δ such that the set,

$$\{ \| e^{t B(\Delta x)} \| \} \quad , \tag{1.2.10}$$

is uniformly bounded for all $t \in [0,\tau]$ and $0 < \Delta x < \delta$ where $\Delta x = \max_j \{\Delta x_j\}$.

From the above definitions the following theorem, fundamental to the analysis of numerical models of the linear IBVP , may be proved.

Theorem 1.2.1 The Lax-Richtmyer Equivalence Theorem for SD Methods:

Given a properly-posed partial differential equation (1.2.2) and a consistent finite-difference SD approximation (1.2.7), convergence is equivalent to stability [Go77]. □

Therefore the problem of determining under which conditions convergence of the approximation to the real solution occurs as the mesh is refined may be investigated via the stability of the numerical model.

Performing a full discretisation of (1.2.2), convergence, stability and consistency are defined in a similar manner. Let

$$v^{n+1} = C(\Delta t) v^n + F^n(\Delta t) \quad , \quad n \geq 0 \quad , \tag{1.2.11}$$

be a one-step scheme where $C(\Delta t)$ is a matrix function depending on Δx through Δt, where v^n approximates $u(n\Delta t)$, $v_j^n \approx u(j_1\Delta x_1, j_2\Delta x_2, \ldots, j_m\Delta x_m, n\Delta t)$ and F^n approximates $f(n\Delta t)$. The family of operators $C(\Delta t)$ is defined to be a convergent approximation if

$$\| (C(\Delta_j t)^{n_j} - E(t))\phi \| \to 0 \quad , \quad 0 \leq t \leq \tau \quad , \quad (1.2.12)$$

for any sequences $\{\Delta_j t, n_j\}$ such that $\Delta_j t$ tends to zero and $n_j \Delta_j t$ tends to fixed t as $j \to \infty$ for every $\phi \in B$. It is *stable* if for some $\tau^* > 0$ the family of operators,

$$\{ C(\Delta t) \}^n \quad , \quad 0 \leq \Delta t \leq \tau^* \quad , \quad (1.2.13)$$

for $0 \leq n\,\Delta t \leq \tau$ is uniformly bounded. By the *Lax-Richtmyer Equivalence Theorem* for FD methods, stability is equivalent to convergence under suitable conditions of consistency of the approximation and well posedness of the equation [La56].

In the following paragraphs we review many of the existing theories for determining the stability of SD models. Most of this theory was originally derived for FD models and then adapted later. Therefore parallel results, which will not always be quoted here, do exist for FD models. As we have already stated, it is the SD model which interests us more in this dissertation and therefore we feel this rather "back-to-front" discussion of stability is justified. Many references to the FD theory will be mentioned and where results will be required later the exact theory is stated.

1.3 The Cauchy Problem and Fourier Analysis

Determination of stability criteria is considerably simpler for the constant coefficient problems defined on a bi-infinite domain with continuous initial conditions. Such equations yield straightforwardly to a Fourier analysis. A similar analysis can be applied if the boundary conditions are periodic [Is84b].

Working in the ℓ_2 norm, the Fourier transform provides an *isometric isomorphism* between points in the solution space B and the transformation space $\tilde{B}$ [Ri67]. Consider the finite-difference scheme for solving (1.2.2) without a forcing term as follows,

$$\sum_{\beta} f_{\beta} \, v'_{k+\beta}(t) = \sum_{\beta} g_{\beta} \, v_{k+\beta}(t) \qquad (1.3.1)$$

$$v_k(0) = \phi(k\Delta x),$$

where k and β are multi-indices. We assume that the grid is regular in each direction, i.e. $\Delta x_1 = \Delta x_2 = \ldots = \Delta x_m$, and that the coefficients f_{β} and g_{β} are constants. The Fourier transform isometry then enables stability to be investigated in terms of the *characteristic function*, $h(z)$

$$h(z) = \left(\sum_{\beta} f_{\beta} \, z^{\beta}\right)^{-1} \left(\sum_{\beta} g_{\beta} \, z^{\beta}\right) \qquad (1.3.2)$$

where $z^{\beta} = z_1^{\beta_1} z_2^{\beta_2} \ldots z_m^{\beta_m}$ [Is84b].

Theorem 1.3.1

The method defined by (1.3.1) is stable if and only if there exist finite constants κ and δ such that

$$\| \exp(t\, h\, (e^{i\theta})) \| \leq \kappa \qquad (1.3.3)$$

for every $t \geq 0$, $\theta \in [0,2\pi]^m$ and $0 < \Delta x < \delta$.

This theorem follows immediately from the definition of stability, as the Fourier transformation is norm-preserving. □

Condition (1.3.3) is often called the von Neumann stability condition, although it should not be confused with the condition which we will describe in the next section that bears the same name. For the Cauchy problem as described here the two conditions are equivalent but the second condition is applicable to a wider class of partial differential equations.

In a similar way, Fourier analysis of the FD scheme

$$\sum_{\beta} b_{\beta}\, v^{n+1}_{k+\beta} = \sum_{\beta} c_{\beta}\, v^{n}_{k+\beta} \qquad (1.3.4)$$

describes stability in terms of the *characteristic function*, $a(z,\mu)$,

$$a(z,\mu) = \Big(\sum_{\beta} b_{\beta}(\mu)\, z^{\beta}\Big)^{-1} \Big(\sum_{\beta} c_{\beta}(\mu)\, z^{\beta}\Big) \qquad (1.3.5)$$

where μ is the *Courant number*, $\mu = \Delta t/(\Delta x)^K$ and K is the maximal order of space differentiation occurring in L.

Theorem 1.3.2

The method defined by (1.3.4) is stable if and only if for every $\theta \in [0,2\pi]^m$ and $0 < \Delta x < \delta$

$$\| a(e^{i\theta},\mu) \| \leq 1 \quad . \tag{1.3.6}$$

□

Extensions of Fourier analysis by means of energy inequalities to investigate equations with variable coefficients have been performed. Extra conditions of Lipschitz continuity of the defining functions are needed and the boundedness condition (1.3.6) must be somewhat strengthened [La61,62]. In the next section we consider the techniques usually employed for determining stability of IVP with variable coefficients.

1.4 The Von Neumann Condition and The Kreiss Matrix Theorem

The matrix $B(\Delta x)$ satisfies the von Neumann condition for SD methods if a constant C exists such that for every $0 < \Delta x < \delta$ and $\lambda \in \sigma(B(\Delta x))$

$$\mathrm{Re}\ \lambda < C \quad . \tag{1.4.1}$$

Obviously the von Neumann condition is necessary for stability as defined by (1.2.10). In the particular case when the matrix B is *normal*, that is it commutes with its adjoint, condition (1.4.1) is also sufficient for stability. This will be apparent from the following analysis.

Let $B(\Delta x)$ be a diagonalisable matrix, $B = Q^{-1}DQ$, where D is diagonal, with spectrum $\sigma(B)$. Then

$$\| e^{tB} \| \leq \kappa(Q)\, e^{t\lambda_{max}} , \quad \lambda_{max} = \max \{ \mathrm{Re}\,\lambda : \lambda \in \sigma(B) \}$$

follows from the definition of the matrix exponential and the diagonalisability of B. $\kappa(Q)$ is the *condition number* of Q which equals one if Q is unitary. If B is normal then Q is unitary and therefore sufficiency of (1.4.1) is immediate. However, if B is not normal, it is possible that $\kappa(Q)$ may tend to infinity faster than the exponential decays to zero as the mesh is refined; then stability does not follow.

Sufficient conditions for stability of methods defined by non-normal matrices are more difficult to determine. We define the Kronecker product of two matrices A and B which are m × m and n × n respectively as the mn × mn matrix $A \otimes B$ with elements

$$(A \otimes B)_{(K-1)n+k,(J-1)m+j} = A_{KJ}B_{kj} .$$

Decomposing a non-normal operator R as the Kronecker product

of a fixed-dimensional non-normal matrix A with an m-dimensional normal matrix D as

$$R = A \otimes D ,$$

it can easily be shown that

$$\| \exp(Rt) \| = \max_{\lambda \in \sigma(D)} \| \exp(\lambda A t) \| .$$

As usual $\sigma(D)$ represents the set of eigenvalues of D. Stability may therefore be determined by the Kreiss matrix theorem [Kr59] which gives conditions for a family of matrices to be stable as defined by (1.2.10).

Theorem 1.4.1 The Kreiss Matrix Theorem

For any family of $m \times m$ matrices $A(\omega)$ where $\omega \in \Omega$, is an arbitrary complex parameter, the following statements are equivalent:

i) There exists a constant C such that $\| \exp(A(\omega)t) \| \leq C$ for all $t \geq 0$.

ii) For some constant C_1 and all λ satisfying $\mathrm{Re}\,\lambda > 0$

$$\mathrm{Re}\,\lambda \, \| (\lambda I - A(\omega))^{-1} \| \leq C_1 .$$

iii) There exist symmetric matrices $H(\omega)$ satisfying

$$H(\omega)\, A(\omega) + A^{*}(\omega)\, H(\omega) \leqslant 0 \quad , \quad I \leqslant H(\omega)$$

and $$\| H(\omega) \| \leqslant C \quad .$$

iv) There exist matrices satisfying iii) and $K(m)$ depending only on m and not on the family $A(\omega)$ such that

$$\| H(\omega) \| \leqslant K(m) C_1$$

[Go77]. □

The most significant result here is condition ii), known as the *resolvent condition*, which comes from proving that

$$\| \exp(A(\omega)t) \| \leqslant K_1(m) \max_{\mathrm{Re}\,\lambda > 0} \mathrm{Re}\,\lambda \; \| (\lambda I - A)^{-1} \|$$

for some $K_1(m)$ [La75]. This characterisation of stability and its equivalent form for FD methods, [Kr64], is extremely useful in theoretical work. In many cases the question of stability is reduced to an estimate of the resolvent ([Kr68], [St80], [Gu72]).

An alternative technique for determining stability which has already been mentioned is the *energy method*. This method is far more general than the above techniques and relies on the determination of a norm in which the solution is demonstrated to be uniformly bounded. As application of the energy method requires proof of stability for each

method considered, it might seem that this technique is not particularly useful. However, it can be applied not only to Cauchy problems but also to mixed boundary-value problems and nonlinear equations, allowing proofs of stability in difficult cases ([En81], [St64a], [Gu72], [La60a]). Therefore it is extremely useful, even if more complicated to implement in practice.

The use of the word energy here is rather misleading. In some cases the energy method does prove that the energy of the system is conserved, but in general this is not so. Generally, the idea is to show that,

$$\frac{d}{dt} \| v(t) \|^2 \leq K \| v(t) \|^2$$

for some real constant K. Then $\| v(t) \| \leq e^{\frac{1}{2}Kt} \| v(0) \|$ and stability follows since $v(t) = e^{Bt} v(0)$.

The incorporation of boundaries further complicates the stability analysis. In the next section we consider equations defined on finite domains with homogeneous boundary conditions and for simplicity restrict the analysis to the scalar case.

1.5 Toeplitz Operators, Godunov-Ryabenki Analysis and Homogeneous Boundary Conditions

The Kreiss Matrix Theorem enables determination of

stability for non-normal operators which may be decomposed in the particular way shown. Unfortunately the incorporation of boundary conditions into R affects the normality of D. Now stability requires uniform boundedness of R as the mesh is refined. Naturally with this refinement the size of R and hence D increases. As a consequence the non-normality of R means that the eigenvalues of the finite matrices, and hence the preceding analysis, may be quite misleading as a guide to stability.

We will assume that the boundary conditions are homogeneous and that the same model is used to approximate near the boundary at each time level. Then the boundary conditions only impose a small correction on the normal matrix D which describes the internal model. Thus stability can be investigated by the the analysis of local normal modes as introduced by Godunov and Ryabenki [Go64].

Consider the family of operators $\{R(\Delta x)\}$. The point λ in the complex plane is called a point in the spectrum of the family of operators $\{R(\Delta x)\}$ if, for any positive ϵ we may choose $\Delta x_0 > 0$, such that for any Δx, $0 < \Delta x < \Delta x_0$ the inequality

$$\| (R(\Delta x) - \lambda I) v \| \leqslant \epsilon \| v \| \qquad (1.5.1)$$

is satisfied by some function v belonging to the space on which $R(\Delta x)$ is defined [Go64]. The aggregate of all such numbers λ forms the spectrum of the family of operators $\{R(\Delta x)\}$.

Obviously for any point λ not in the spectrum of $\{R(\Delta x)\}$ there exists $\Delta x < \Delta x_0$ such that $R(\Delta x) - \lambda I$ is invertible. Therefore the G-R criterion which is necessary for stability of SD methods, is that every element λ in the spectrum of $\{R(\Delta x)\}$ must have non-positive real part. With this condition imposed, the amplitudes of normal modes will either remain constant or decay in time. Therefore normal modes near the boundaries, which might have a pronounced effect on stability, are also required to remain constant or decay. Notice that the G-R criterion is in a sense an analogue for IBVP of the von Neumann condition for IVP.

Until now the implications arising from the possibility of the operator $B(\Delta x)$ being *implicit*, $B = B_1^{-1} B_2$ and $B_1 v' = B_2 v$, have not been discussed. The model is not *solvable* unless B_1 is *invertible* and therefore stability requires that in addition B_1, which depends on Δx, must be uniformly invertible as the mesh is refined. In the case of a Cauchy problem, a uniform shift in the centre of the grid to make a locally invertible B_1 uniformly invertible would be unnoticeable. However, if boundaries are incorporated, such a shift would alter the position of the boundaries and therefore the problem itself.

We restrict our analysis to the schemes in one space variable suitable for solving the *scalar* hyperbolic or parabolic equations

$$u_t = a u_x \quad , \quad \text{or} \quad u_t = a u_{xx} \quad . \qquad (1.5.2)$$

Full discretisations for solving equations (1.5.2) have been thoroughly investigated by Strang, first on the semi-infinite interval [St64b], and then on the finite interval, [St66]. He has also given an extension of the theory to the variable-coefficient case a(x): sufficient conditions for stability are the same as for the constant coefficient case with the additional requirement of Lipschitz continuity of the defining functions [St64b]. Determination of stability relies completely on a *Wiener-Hopf factorisation* of a *Toeplitz operator* to define invertibility in terms of an index condition for the operator. The extension to the SD case was carried out recently [Is83a].

We consider the SD scheme,

$$\sum_{j=-R}^{S} f_j \; v'_j(t) = \sum_{j=-r}^{s} g_j \; v_j(t) \qquad (1.5.3)$$

or its equivalent in vector notation,

$$\tilde{F}\, v' = \tilde{G} v \quad .$$

Stability requires that $\tilde{F}$ has a bounded inverse. $\tilde{F}$ is thought of as a Toeplitz matrix with the f_j on the j-th diagonal $j \in \mathbb{Z}$. Now a correspondence between doubly infinite Toeplitz matrices and functions $F(e^{i\theta})$ is established by

$$F(e^{i\theta}) = \sum f_j \, e^{ij\theta} \quad \langle === \rangle \quad \tilde{F} = T(F),\ T(F)_{jk} = f_{k-j} \qquad (1.5.4)$$

[St64]. Then the requirement that $\tilde{F}$ has a bounded inverse means that $F(z)$ must be analytic in an annulus enclosing the unit circle. Further, we apply the Wiener-Hopf technique directly on the half line $0 \leq x < \infty$. $\tilde{F}$ is factored into a product $\tilde{U}\tilde{L}$ of upper and lower triangular matrices which are themselves Toeplitz via the corresponding factorisation of F into a product of outer and inner functions.

We assume that the Toeplitz matrix $\tilde{F}$ can be factorised as a product $\tilde{U}\tilde{L}$. Then the associated Laurent polynomial $F(z)$ can be factored as two polynomials $U(z)L(z)$ where the S factors corresponding to the largest roots go into $U(z)$ and the others into $L(z)$. Then the Wiener-Hopf method depends entirely on the properties of the correspondence of $L(z)$ and $U(z)$ with the appropriate lower and upper triangular matrices $\tilde{L}$ and $\tilde{U}$. For this correspondence the following properties may be proved.

i) $F(z) = U(z)L(z)$ implies $\tilde{F} = \tilde{U}\tilde{L}$.

ii) $\tilde{U}$ is invertible only if all S roots of $U(z)$ satisfy $|z_i| > 1$ and $\tilde{U}^{-1}$ is Toeplitz.

iii) $\tilde{L}$ is invertible only if all R roots of $L(z)$ satisfy $|z_i| < 1$ and $\tilde{L}^{-1}$ is Toeplitz.

iv) $\tilde{F}$ is invertible if and only if both $\tilde{U}$ and $\tilde{L}$ are invertible.

v) $\tilde{F}^{-1}\tilde{G} = \tilde{L}^{-1}\tilde{U}^{-1}\tilde{G}$ is similar to $\tilde{H} = \tilde{U}^{-1}\tilde{G}\tilde{L}^{-1}$ which is Toeplitz.

vi) If $\mathrm{Re}\left(\frac{G(e^{i\theta})}{F(e^{i\theta})}\right) \leq 0$ for all $\theta \in [0, 2\pi]$, then $\tilde{H} + \tilde{H}^T$ is negative semi-definite [Is83a].

If the pole condition holds, that is h(z) has R poles in $|z| < 1$ and S poles in $|z| > 1$, then by conditions i) to v) the differential equation can be described in terms of the new variable $w(t) = \tilde{L}v(t)$. But then the von Neumann condition vi) implies that the equation in terms of w(t) is dissipative

$$\frac{d}{dt}(w, w) = ((\tilde{H} + \tilde{H}^T)w, w) \leq 0 \quad .$$

Thus the original problem was indeed stable and the von Neumann and pole conditions are sufficient for stability. Certainly by condition iv) if the pole condition does not hold the matrix $\tilde{F}$ can not be invertible on $0 \leq x < \infty$. Also it can be shown that if vi) is violated for some θ the solution in terms of w(t) explodes and so the original problem was unstable [Is83a]. Thus the von Neumann and pole conditions are also necessary for stability and we have the following theorem which determines convergence completely:

Theorem 1.5.1

An equation of the form (1.5.3) is stable if and only if it satisfies:

i) The von Neumann condition: $\mathrm{Re}\left(\frac{G(e^{i\theta})}{F(e^{i\theta})} \right) \leq 0$ for all θ.

ii) The pole condition: $F(z)$ has R zeros in $|z| < 1$ and S zeros in $|z| > 1$. □

We also state the equivalent result for the FD equation

$$\sum_{-\tilde{R}}^{\tilde{S}} b_j(\mu)\, v_j^{n+1} = \sum_{-\tilde{r}}^{\tilde{s}} c_j(\mu)\, v_j^n \quad . \qquad (1.5.5)$$

Theorem 1.5.2

An equation of the form (1.5.5) is stable if and only if it satisfies:

1) The von Neumann condition: $|a(e^{i\theta})| \leq 1$ for all θ.

ii) The pole condition: $Q(z,\mu)$ has $\tilde{R}$ zeros in $|z| < 1$ and $\tilde{S}$ zeros in $|z| > 1$. □

Here $a(z,\mu)$ is the characteristic function as defined by (1.3.5), and $Q(z,\mu)$ is defined by $a(z,\mu) = z^{\tilde{R} - \tilde{r}} P(z,\mu) / Q(z,\mu)$.

1.6 Gustafsson Kreiss and Sundström Theory

The extension of stability analysis to *mixed initial boundary-value problems* (IBVP) for systems of equations in one space variable occurred mainly for full discretisations. As

i) The von Neumann condition: $\mathrm{Re}\left(\frac{G(e^{i\theta})}{F(e^{i\theta})} \right) \leq 0$ for all θ.

ii) The pole condition: $F(z)$ has R zeros in $|z| < 1$ and S zeros in $|z| > 1$. □

We also state the equivalent result for the FD equation

$$\sum_{-\tilde{R}}^{\tilde{S}} b_j(\mu)\, v_j^{n+1} = \sum_{-\tilde{r}}^{\tilde{s}} c_j(\mu)\, v_j^n \quad . \qquad (1.5.5)$$

Theorem 1.5.2

An equation of the form (1.5.5) is stable if and only if it satisfies:

1) The von Neumann condition: $|a(e^{i\theta})| \leq 1$ for all θ.

ii) The pole condition: $Q(z,\mu)$ has $\tilde{R}$ zeros in $|z| < 1$ and $\tilde{S}$ zeros in $|z| > 1$. □

Here $a(z,\mu)$ is the characteristic function as defined by (1.3.5), and $Q(z,\mu)$ is defined by $a(z,\mu) = z^{\tilde{R} - \tilde{r}} P(z,\mu) / Q(z,\mu)$.

1.6 Gustafsson Kreiss and Sundström Theory

The extension of stability analysis to *mixed initial boundary-value problems* (IBVP) for systems of equations in one space variable occurred mainly for full discretisations. As

few results for semi-discretisations are given in literature, the course of events for FD is surveyed here.

In the solution of the IBVP the stability of the *interior problem* away from the boundaries is obviously crucial. We assume that the interior problem is stable in the Cauchy sense, that is for frozen coefficients and the boundaries removed to infinity. Thus we assume the von Neumann condition is satisfied and it is natural to assume the G-R criterion as well.

Initial developments were for restricted classes of problems: a scheme is said to be *dissipative* of order $2q$ if the eigenvalues $\lambda(\theta)$ of the characteristic function of the difference scheme satisfy the estimate

$$|\lambda(\theta)| < 1 - \delta|\theta|^{2q} \qquad (1.6.1)$$

for a constant $\delta > 0$, a natural number $q > 0$ and all θ, $0 \leq |\theta| \leq \pi$ [Kr66].

Imposition of dissipativity on the IVP ensures that high frequency components possibly introduced by the boundary will die away in time. Supposing that the matrices defining the IVP and its difference approximation are Hermitian, uniformly bounded and uniformly Lipschitz continuous functions in the spatial variable, dissipativity is sufficient for stability [Ri67]. Matrices which are simultaneously diagonalisable by unitary transformations are called *hyperbolic*. Hermitian matrices therefore define a hyperbolic

system and consequently hyperbolic, dissipative, IVP are stable. Kreiss [Kr66] uses this property to determine sufficient stability criteria for IBVP with hyperbolic dissipative interior schemes. Extension to non-diagonalisable systems takes place by the use of an integral relation to bound powers of the discrete time evolution operators. Stability is defined in terms of the *generalised eigenvalues* of the system [Kr68].

An alternative approach is to consider the difference operator for the IBVP as a Toeplitz operator defining the interior scheme with a finite-rank correction to account for the boundary values. By means of a Wiener-Hopf type factorisation, Osher [Os69] proves sufficient stability conditions for hyperbolic systems of equations. These are given as a "separation of zeros" criterion for the characteristic function of the interior scheme. The separation condition is weaker than dissipativity and thus the Kreiss results for diagonalisable explicit systems [Kr66] follow as a corollary of Osher's main theorem.

Development of a general theory to cater for all systems of equations, dissipative or nondissipative, diagonalisable or nondiagonalisable, is by means of a rather unusual norm for defining stability [Definition 3.3, Gu72], With this definition the Gustafsson, Kreiss and Sundström (GKS) Theorem [Theorem 5.1, Gu72] states that the difference scheme is stable if and only if there do not exist any generalised eigenvalues outside or on the perimeter of the unit

circle. The motivation for this theorem comes from the work of Kreiss, [Kr70], on the well posedness of a hyperbolic system of differential equations. Such a system is well posed only if it has no eigensolutions.

For the finite interval, stability of the two problems determined by removing one or other boundary to infinity are investigated separately. Stability of the initial boundary-value problem defined on the finite interval follows only if the corresponding left and quarter plane problems are independently stable [Theorem 5.4, Gu72].

Goldberg and Tadmor [Go78,81], give alternative stability criteria in terms of the boundary conditions. Assume that the boundary schemes are *translatory*, that is the same scheme is used at all grid points. Then defining stability as in the GKS theory, they prove that stability follows independently of the interior scheme, assumed to be Cauchy stable, if the boundary conditions are generated by an invertible stable scheme.

Strikwerda, [St80], has provided an extension of GKS theory to SD schemes in the case of hyperbolic systems of equations in one-space variable. Stability, defined with a GKS-type norm, occurs if and only if there are no eigensolutions of the difference scheme. It appears that stability of nondiagonalisable SD schemes has not been investigated.

1.7 A Group Velocity Interpretation Of Stability

For a thorough description of the *group velocity* analysis of finite-difference schemes, the reader is referred to the work of Trefethen [Tr82a, 82b,83]. Here some basic definitions are given and the main results concerning the interpretation of GKS theory are highlighted.

Consider a scalar linear partial differential equation with constant coefficients admitting solutions of the form

$$u(x,t) = e^{i(\omega t+\xi x)} \quad .$$

The *dispersion* relation for the partial differential equation is $\omega = \omega(\xi)$ for each real *wavenumber* ξ and corresponding *real frequency* ω. The speed of propagation of the solution is called the *phase speed*,

$$c(\xi) = \frac{\omega(\xi)}{\xi} \quad ,$$

whilst the speed at which energy travels is the *group speed*

$$C(\xi) = \frac{d\omega}{d\xi}(\xi) \quad .$$

Trefethen, [Tr81a], demonstrates a connection between the instability of an IBVP and the possibility that a set of waves can radiate from a boundary. This, in turn, is linked to wavenumbers with negative group velocity being supported

by the boundary schemes. Further, this is connected with the normal-mode analysis of GKS for determining the eigensolutions supportable by the scheme. Trefethen points out the existence of ℓ_2 -stable models which are unstable in the sense of GKS showing that in the case of zero group velocity the GKS theory is insufficient. There is no claim, however, that the group velocity analysis is itself complete. It is limited entirely to non-dissipative schemes and in some ways the analysis is negative demonstrating only instability rather than the traditional view of demonstrating stability.

The different approach of the group velocity technique for examining stability is, nonetheless important. The complicated GKS criteria are given physical interpretation in terms of the group velocity of parasitic waves. In realistic physical applications this approach reinforces the quite evident idea that spontaneous emission of energy from the boundary into the interior will cause instability. The relationship between GKS and group velocity is made more exact in a later work by Trefethen [Tr83].

1.8 Solution Of The Ordinary Differential System Of Equations

We showed in earlier sections that the semi-discretisation of a partial differential equation yields an *ordinary differential system of equations* $v' = Bv$. The solution of such systems of equations has been thoroughly investigated and numerous stability concepts derived imposing uniform boundedness with respect to various norms on the solution $v(t)$. A complete analysis of systems of ordinary differential equations is given in many traditional texts, for example Henrici, [He62].

Suppose that we solve the linear constant coefficient semi-discrete problem by a one-step integration formula

$$v^{n+1} = R(\Delta t B)v^n \qquad (1.8.1)$$

where $R(\Delta t B)$ is a specific matrix-valued polynomial or rational function. The scalar function $R(z)\colon \mathbb{C} \to \mathbb{C}$, for $z = \Delta t \lambda$, $\lambda \in \sigma(B)$, is called the *stability function* and determines the stability of the scheme. Then (1.8.1) is called *absolutely stable* at $z \in \mathbb{C}$, if, for this z, $|R(z)| < 1$ and A-stable if it is *absolutely stable* for all $z \in \mathbb{C}$, $\mathrm{Re}\, z < 0$. The following equivalence holds if the coefficients of the integration scheme are real constants,

$$\lambda \in \sigma(B) \quad |R(\Delta t \lambda)| < 1 \quad <===> \quad \rho[R(\Delta t B)] < 1 \quad . \qquad (1.8.2)$$

Then if the spectral radius of R is less than one for all eigenvalues of B, the sequence $\{v^n\}$ vanishes when $n \to \infty$. But how relevant is the concept of absolute stability in the analysis of partial differential equations? We need to know when absolute stability guarantees stability of the partial differential equation. In this case we require that $v^{n+1} = R^{n+1} v^0$ satisfies $\| v^{n+1} \| < \| v^0 \|$ as $n \to \infty$. So we need $\| R^{n+1} \| < 1$ and $\rho(R) < 1$ is only sufficient if R is a normal function.

Example 1.8.1

Consider solving the simple hyperbolic partial differential equation

$$\begin{aligned} u_t &= u_x \quad , \ 0 \leq x < 1, \quad t > 0 \\ u(1,t) &= 0, \quad t > 0 \\ u(x,0) &= \phi(x), \quad 0 \leq x \leq 1 \ , \end{aligned} \tag{1.8.3}$$

where $\phi(x)$ is a smooth initial condition. We perform a semidiscretisation of the spatial derivative by the standard Euler explicit method on the interval $[0,1]$ divided into m equal subintervals, $\Delta x = 1/m$. The time dependent function $v = [v_1, v_2, \ldots, v_m]^T$ then satisfies,

$$v' = Bv, \tag{1.8.4}$$

where B is the constant $m \times m$ matrix,

$$B = \frac{1}{\Delta x}\begin{bmatrix} -1 & 1 & & & 0 \\ & -1 & 1 & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ 0 & & & -1 & 1 \\ & & & & -1 \end{bmatrix}$$

Integration of this equation again by explicit Euler defines the one-step integration formula

$$v^{n+1} = R v^{n}$$

where v^n approximates $v(n\Delta t)$ and R is the $m \times m$ matrix,

$$R = \begin{bmatrix} 1-\mu & \mu & & & 0 \\ & 1-\mu & \mu & & \\ & & \cdot & \cdot & \\ & & & \cdot & \mu \\ 0 & & & 1-\mu & \mu \\ & & & & 1-\mu \end{bmatrix},$$

$\mu = \Delta t / \Delta x$ being the Courant number.

Obviously the spectral radius of R is less than one for $0 < \mu < 2$ and the eigenvalues of B lie inside the stability region of R for these bounds on μ.

However, it is stability of the complete model which is of interest when solving (1.8.3). It is elementary to show

using the energy method that the solution of the full discretisation is bounded in the ℓ_2 norm if and only if $0 \leq \mu \leq 1$. By Brenner, Thomée and Wahlbin this result can be extended to stability in ℓ_p, $1 \leq p \leq \infty$ [Br75]. This condition is far more stringent on the permissible values of μ for ℓ_p-stability and consequently absolute stability conditions must be viewed with caution in the SD framework.

Notice that if the matrix R had been normal, then $\rho(R) = \| R \|_2$ and the correct bound would have been obtained.

A stability analysis of the time integration method should be used to bound the size of the time step which can be safely used. The above example demonstrates that it is not enough to merely assert that Δt times the eigenvalue of the Jacobian matrix B should be inside the stability region for every eigenvalue. Boundedness is required as the mesh is refined or equivalently as the size of the matrix increases. Treating the system as a whole we have an FD scheme and we therefore require that the spectrum of the infinite Toeplitz operator must lie within the stability region, this being the von Neumann condition necessary for stability of FD schemes.

If the Toeplitz operator is Hermitian, then the asymptotic distribution of the eigenvalues can be determined by the theory of *equal distributions*, as defined in Grenander and Szegö (p.63, [Gr58]). We denote the eigenvalues of the

finite Toeplitz form by

$$\lambda_1^{(m)} \leqslant \lambda_2^{(m)} \leqslant \ldots \leqslant \lambda_{m+1}^{(m)}.$$

Then a consequence of a theorem (p.68, [Gr58]) is that these eigenvalues are equally distributed amongst the spectrum of the infinite Toeplitz form as $m \to \infty$. Therefore absolute stability is a safe criterion if the underlying matrix is Hermitian as in this case the eigenvalues of the finite form tend uniformly to the spectrum of the infinite form.

In many cases in the literature the Jacobian matrix has been Hermitian and therefore the problem demonstrated has not occurred. The eigenvalues of the Jacobian have been enough to bound the Courant number. In this dissertation we investigate some semi-discretisations whose Jacobians are not Hermitian and demonstrate that consideration of the eigenvalues alone can indeed be quite disastrous.

2. STABILITY, ORDER OF ACCURACY AND AN APPROXIMATION-THEORETIC PROBLEM

2.1 Introduction

During the last few decades much interest has been shown in determining the maximal accuracy of schemes for solving partial and ordinary differential equations. In some cases maximally accurate schemes can be derived. However, imposition of stability often drastically reduces the order of accuracy attainable. Alternatively, upper bounds for stability can be proved and the problem is then to find schemes which attain this bound. A further criterion is whether the maximally accurate stable schemes can be chosen to have minimal error constants.

Historically, the first success along these lines was the proof by Dahlquist [Da56,63] of stability barriers for multistep methods in the solution of ordinary differential equations. The first Dahlquist barrier states that a zero stable k-step method for the solution of the non-linear equation, $u_t = f(u)$, has accuracy bounded by $2[\frac{k+2}{2}]$ [Da56]. If *A-stability* is imposed, we have the second Dahlquist barrier that accuracy cannot be greater than 2 [Da63]. The scheme with smallest error constant is then the trapezoidal rule which is a one step method.

The introduction of the theory of order stars by Wanner, Hairer and Nørsett [Wa78] has been instrumental in proving many similar results. In this paper they generalised the Dahlquist second barrier to multistep n-derivative schemes to prove the Daniel-Moore conjecture, that an A-stable method cannot have accuracy $p > 2n$.

However, the investigation of partial differential equations is more complicated, with different results being derived according to the type of the underlying equation. Much studied is the linear hyperbolic equation in one space variable,

$$u_t = u_x \quad . \tag{2.1.1}$$

Strang [St62] derived maximally accurate explicit FD schemes (1.5.5), $p = \tilde{s} + \tilde{r}$, which he proved were stable for Courant number μ, $0 < \mu < 1$, provided that $\tilde{s} = \tilde{r}$, $\tilde{r} + 1$ or $\tilde{r} + 2$. Instability for other choices of $\tilde{r}$ and $\tilde{s}$ was not proved. The proof that the one-sided schemes, $\tilde{r} = 0$, or $\tilde{s} = 1$ have order $p \leq 2$, for stability, was proved by Engquist and Osher [En81] and Strang [St64a] respectively.

Unification of these results for the fully discrete models relies on the observation made by Iserles and Strang [Is83a] that associated with every FD scheme, is an SD scheme which has accuracy and stability properties inherited from the defining scheme. This relationship will be described in detail in Section 2.3.

Iserles [Is82], investigated the maximal order of SD schemes using an extension of the theory of order stars. From the relationship just mentioned his saturation result,

$$p \leq \min \{ r + s , 2(r + 1) , 2s \}$$

for explicit stable schemes, (1.5.3), immediately bounds the order of explicit stable schemes (1.5.5),

$$p \leq \min \{ \tilde{r} + \tilde{s} , 2(\tilde{r} + 1) , 2\tilde{s} \}$$

[Is83a]. By the same relationship, they also prove that stable implicit schemes, (1.5.5), have order bounded by

$$p \leq \min \{ \tilde{r} + \tilde{s} + \tilde{R} + \tilde{S} , 2(\tilde{r} + \tilde{R} + 1) , 2(\tilde{s} + \tilde{S}) \} .$$

Stable schemes attaining these bounds are derived in both papers.

For the explicit SD schemes, Iserles [Is82] derived an expression for the error constants of stable methods from the class $\{\tilde{r}, \tilde{s}\}$, $\tilde{r} \geq \tilde{s}$, with accuracy $2\tilde{s}$, and demonstrated that increasing the number of steps taken to the left of the origin does not improve the error constant. Increasing the number of steps taken to the right of the origin does, however, lead to a scheme with minimal error constant [Je84]. Taking an increased set of points to the right brings us arbitrarily near to the minimal error constant with a stable scheme. But the scheme with minimal error constant is unstable.

An extension of some of these results to multistep schemes was proved by Strang and Iserles [St82]. An explicit multistep stable scheme using $\tilde{s}$ points to the right and $\tilde{r}$ points to the left at each time level, has accuracy $p \leq 2(\tilde{r} + \tilde{s})$. If the variable coefficient problem, $u_t = a(x)u_x$, is being solved, then this bound is halved, $p \leq (\tilde{r} + \tilde{s})$ [Is84c].

For parabolic problems, $u_t = \omega u_{xx}$, $\omega > 0$, the only result appears to be that by Iserles [Is83a]. He considers the solution of this parabolic problem by an implicit fully discrete scheme using the same number of points, $\tilde{r}$, to the left and right at each time level. Maximal accuracy with stability is $p \leq 4\tilde{r} + 1$.

Most of these recent saturation results have been with the aid of the theory of order stars and some generalisations. The question of reconciling accuracy and stability is posed as a problem in approximation theory. We require to investigate a rational approximation to some function where the approximation must, for the sake of stability, satisfy both the von Neumann condition and the pole condition. Order star theory is crucial here for demonstrating which particular distributions of poles and zeros can occur.

In the next few sections we aim to complete the work of Iserles and Strang [Is83a]. We prove a result bounding accuracy for all stable implicit semi-discretisations. For this we must define the approximation problem completely and also derive an extra condition for stability that follows from the pole condition.

2.2 Order of Accuracy

Determining the *order of accuracy* of a numerical scheme can be interpreted as a problem in approximation theory. To justify this statement we consider the evolution equation in one space variable,

$$u_t = L\,u \ , \ u(x,0) = \phi(x) \ , \ u = u(x,t) \ , \qquad (2.2.1)$$

together with suitable boundary conditions. The operator L is a linear differential operator of the form,

$$L = \sum_{j=0}^{K} a_j D^j \qquad (2.2.2)$$

where the a_j are constant coefficients and $D = \partial/\partial x$ is the differential operator. Defining E to be the spatial shift operator, $E\,u(x_j) = u(x_j + \Delta x)$, we can reexpress (2.2.2) as

$$L = L(E) = \sum_{j=0}^{K} a_j \frac{(\ell n E)^j}{\Delta x^j} \ , \qquad (2.2.3)$$

since by operator theory $E = e^{\Delta x D}$.

The approximation of (2.2.1) by the semidiscrete system,

$$v' = B(\Delta x)\,v \ , \qquad (2.2.4)$$

is defined to be accurate of order p if $B(\Delta x)$

approximates the differential operator L with error $O(\Delta x^{p+1})$. Recalling the definition of consistency, (1.2.8), it is evident that consistency is equivalent to order of accuracy being at least one.

Assuming that the same scheme is used to solve (2.2.1) at all grid points away from the boundary, we write (2.2.4) more precisely as a system of equations, one for each $v_m(t)$,

$$\sum_{j=-R}^{S} f_j v'_{m+j}(t) = \frac{1}{\Delta x^K} \sum_{j=-r}^{s} g_j v_{m+j}(t) \qquad (2.2.5)$$

where K is the maximal order of the differential operator, $\partial/\partial x$ occurring in L .

As the operators $D_t = \partial/\partial t$ and E commute, we may write this as,

$$\left\{ D_t \sum_{j=-R}^{S} f_j E^j - \frac{1}{\Delta x^K} \sum_{j=-r}^{s} g_j E^j \right\} v_m(t) = 0 \quad .$$

The differential equation (2.2.1) defines $D_t = L$, therefore the finite-difference scheme (2.2.5) normalised by Δt, is of order of accuracy p in Δx around $z_o = 1$, $z = 1 + O(\Delta x)$ if and only if

$$\frac{\Delta t}{(\Delta x)^K} h(z) = \Delta t\, L(z) + c_{p+1}(z-1)^{p+1} + O(|z-1|^{p+2}) \quad .$$

(2.2.6)

Here $c_{p+1} \neq 0$ is a constant, $h(z)$ is the rational function,

$$h(z) = \frac{\sum_{j=-r}^{s} g_j z^j}{\sum_{j=-R}^{S} f_j z^j}$$

and we have allowed Δt and Δx to vary whilst the Courant number, $\mu = \Delta t/(\Delta x)^K$, has been held constant. Assuming that the implicit part of the operator is invertible for small Δx, the approximation $h(z)$ is well defined.

Similarly, it can be shown that the solution of (2.2.1) by the full discretisation,

$$\sum_{j=-\tilde{R}}^{\tilde{S}} b_j \, v_{m+j}^{n+1} = \sum_{j=-\tilde{r}}^{\tilde{s}} c_j \, v_{m+j}^{n}$$

is of order accuracy $\tilde{p}$ in Δx around $z = 1$ if and only if,

$$a(z,\mu) = \exp(\Delta t \, L(z)) + \tilde{c}_{\tilde{p}+1} (z-1)^{\tilde{p}+1} + \mathcal{O}(|z-1|^{\tilde{p}+2}) \, . \tag{2.2.7}$$

As above, $\tilde{c}_{\tilde{p}+1} \neq 0$ is a constant and $a(z,\mu)$ is a rational function which depends on the Courant number via the dependence of its coefficients, b_j and c_j on μ ;

$$a(z,\mu) = \frac{\sum_{j=-\tilde{r}}^{\tilde{s}} c_j(\mu) z^j}{\sum_{j=-\tilde{R}}^{\tilde{s}} b_j(\mu) z^j} \quad .$$

Having posed the problem of determination of order of accuracy as a problem in approximation theory, we are now in a position to discover the maximal accuracy obtainable of some classes of stable schemes. We describe our particular approximation problem more exactly in the next section.

2.3 The Approximation Problem and a Relationship Between Fully- and Semi-discretised Schemes

For the solution of the linear hyperbolic equation in one space variable, (2.1.1), the linear differential operator, L, is just the partial derivative $\partial/\partial x$. Thus the accuracy condition (2.2.6) is,

$$h(z) = \ln z + c_{p+1}(z - 1)^{p+1} + O(|z - 1|^{p+2}) \quad . \quad (2.3.1)$$

Necessary and sufficient conditions for stability of models such as (2.2.5) are given by Theorem 1.5.1 as conditions on the rational function $h(z)$. Therefore, determination of the maximal accuracy of a stable scheme of type (2.2.5) for solving (2.2.1), can be investigated completely through the properties of the function $h(z)$. Stated precisely the

problem is as follows:

Find the maximal order of accuracy, p, of the rational approximation, $H(z) = z^{r-R}h(z)$ to $z^{\ell}\ln z$, $\ell = r-R$, in the neighbourhood of $z = 1$, where $h(z)$ satisfies the following conditions:

(1) The von Neumann condition: $\mathrm{Re}\, h(e^{i\theta}) \leqslant 0$ for all θ,

(2) The pole condition: $h(z)$ has R poles in $|z| < 1$ and S poles in $|z| > 1$.

This problem was partially solved by Iserles and Strang [Is83a] when they solved the equivalent problem for the fully discrete scheme. In this case, by (2.2.7), the characteristic function, $a(z,\mu)$, is an approximation to z^{μ},

$$a(z,\mu) = z^{\mu} + \tilde{c}_{\tilde{p}+1}(z-1)^{\tilde{p}+1} + O(|z-1|^{\tilde{p}+2}) \tag{2.3.2}$$

Thus we seek rational functions, $A(z,\mu) = z^{\tilde{r}-\tilde{R}}a(z,\mu)$ where $a(z,\mu)$ satisfies the condition of Theorem 1.5.2.

(1) The von Neumann condition: $|a(e^{i\theta},\mu)| \leqslant 1$ for all θ,

(2) The pole condition: $a(z,\mu)$ has $\tilde{R}$ poles in $|z| < 1$ and $\tilde{S}$ poles in $|z| > 1$.

As mentioned in Section 2.1, the partial solution to

our problem arises from a relationship between fully- and semi-discretised schemes. Forming the derivative of $a(z,\mu)$ with respect to μ, evaluated at $\mu = 0$, gives a rational approximation to the logarithmic function in the neighbourhood of $z = 1$. The SD scheme defined by this approximation is called the *associated* SD scheme of the FD scheme. Each FD scheme must be associated with an SD scheme but not all SD schemes can be derived in this way.

We have the rational functions,

$$a(z,\mu) = \frac{\sum_{j=-\tilde{r}}^{\tilde{s}} c_j(\mu) z^j}{\sum_{j=-\tilde{R}}^{\tilde{S}} b_j(\mu) z^j} = \frac{c^*(z,\mu)}{b^*(z,\mu)} \qquad (2.3.3)$$

and

$$h(z) = \frac{\sum_{j=-r}^{s} g_j z^j}{\sum_{j=-R}^{S} f_j z^j} = \frac{g^*(z)}{f^*(z)} \,. \qquad (2.3.4)$$

As we require that the map defined by the FD approaches the identity as $\mu \to 0$ we have $c^*(z,0) = b^*(z,0)$. Therefore,

$$h(z) = \frac{\partial}{\partial\mu} \left[a(z,\mu) \right]_{\mu=0}$$

$$= \frac{c^*_\mu - b^*_\mu}{b^*} \Big|_{\mu=0} = \frac{c^*_\mu - b^*_\mu}{c^*} \Big|_{\mu=0} \,, \qquad (2.3.5)$$

where c^*_μ and b^*_μ denote the partial derivatives of c^* and b^* with respect to μ. By the accuracy condition (2.3.2), $h(z)$ approximates $\ell n\, z$ with order $p \geqslant \tilde{p}$. From (2.3.3) the powers of z in the numerator of $h(z)$ extend from $-\tilde{r}^*$ to $\tilde{s}^*$ and in the denominator from $-\tilde{R}^*$ to $\tilde{S}^*$ where

$$\tilde{r}^* = \max(\tilde{r}, \tilde{R}) \quad , \quad \tilde{s}^* = \max(\tilde{s}, \tilde{S}) \qquad (2.3.6)$$

$$\tilde{R}^* = \min(\tilde{r}, \tilde{R}) \quad , \quad \tilde{S}^* = \min(\tilde{s}, \tilde{S}) \quad .$$

The associated scheme therefore has $r = \tilde{r}^*$, $s = \tilde{s}^*$, $R = \tilde{R}^*$, $S = \tilde{S}^*$ and so $r \geqslant R$, $s \geqslant S$. Consequently the set of values $\{r,s,R,S\}$ which the SD scheme may take is restricted and so not all SD schemes are associated with FD schemes. As $p \geqslant \tilde{p}$, the associated scheme inherits accuracy of the FD scheme and any bound on p obviously bounds $\tilde{p}$.

The stability conditions in both cases mean that stability is also inherited. By Taylor expansion we have,

$$|a(z,\mu)|^2 = |a(z,0)|^2 + 2\mu \operatorname{Re} \frac{\partial a}{\partial \mu}\Big|_{\mu=0} + O(\mu^2) \quad .$$

Hence

$$\operatorname{Re} h(e^{i\theta}) = \lim_{\mu \to 0} \frac{|a(e^{i\theta},\mu)|^2 - 1}{2\mu} \qquad (2.3.7)$$

and one von Neumann condition implies the other. From (2.3.5) the zeros of $z^R f^*(z)$ are just the zeros of

$z^{\tilde{R}} b(z,0)$, apart from the zeros at the origin which correct for the difference in degree. One pole condition therefore follows from the other.

Iserles and Strang [Is83a] employed this relationship to determine an upper bound on $\tilde{p}$ by working with SD schemes with $r \geqslant R$ and $s \geqslant S$. They derived the bound,

$$p \leqslant \min \{ 2(r+R+1) , 2(s+S+1) , r+R+s+S \} \qquad (2.3.8)$$

for this restricted class of SD schemes. In the next chapter we extend some of the ideas of the order star theory, enabling the bound (2.3.8), with the refinement $2(s+S+1)$ to $2(s+S)$, to be derived for all SD schemes.

2.4 The Zero Condition for Stability

Iserles and Strang [Is83a] were able to derive the upper bound given by (2.3.8) quite straightforwardly, having presented a relevant order star theory. Consideration of the evolution equation (2.1.1) with initial condition, $u(x,0) = \phi(x)$, ϕ a smooth function, shows that this bound can not be optimal. Solutions of the evolution equation travel along its characteristic curves; $u(x,t) = \phi(x+t)$. We would therefore expect signals to travel from right to left. Assuming that the Courant number, $\mu = \frac{\Delta t}{\Delta x} < 1$, it may be expected that the number of points taken to the left would correspond to the number of points taken to the right, in

both explicit and implicit parts of the of the operator. Thus there should be a correspondence between $\tilde{r}+1$ and $\tilde{s}$, and $\tilde{R}+1$ and $\tilde{S}$, implying that the bound (2.3.8) is too generous.

A variant of the order star theory is required to determine the optimal bound on p. The extra condition that the geometry of the order star has to obey comes about by realising that the pole condition imposes conditions not only on the location of the poles but, in doing this, on the location of the zeros as well. This follows from the analyticity of the approximation to the logarithmic function.

Theorem 2.4.1

If the ISD scheme is stable and $p \geqslant 1$ then

$h(z)$ has at most r zeros in $0 < |z| < 1$

and at most $s - 1$ zeros in $\infty > |z| > 1$.

Proof

The desired result follows from the argument principle for meromorphic functions applied along the unit circle. However consistency requires $h(1) = 0$ and therefore the argument principle cannot be applied directly. Consider instead the function $h^*(z,\epsilon)$ defined for $0 < \epsilon \ll 1$ by

$$h^*(z,\epsilon) := h(z) - \epsilon \quad .$$

Then $h^*(z,\epsilon)$ has the same poles as $h(z)$ and its zeros are near those of $h(z)$ for small ϵ.

The von Neumann condition for stability leads to

$$\mathrm{Re}\, h^*(e^{i\theta},\epsilon) < 0 \quad \text{and thus} \quad [\arg h^*(e^{i\theta},\epsilon)]_0^{2\pi} = 0.$$

Therefore, the number of zeros and poles inside the unit circle is the same. This is true for every sufficiently small $\epsilon > 0$. Hence, by letting ϵ tend to zero, $h(z)$ has at most R zeros inside the unit circle. However at most r of these zeros lie away from the origin since $h(z)$ has a zero of order at least $(R - r)$ at the origin.

To obtain the result outside the unit circle we use the mapping $w = 1/z$ to map the outside of the unit circle inside and vice versa and apply the argument principle as above. Then $h(z)$ has at most s zeros outside the unit circle away from infinity.

However, $h(1) = 0$ and so there is at least one zero on the unit circle which must have migrated from either inside or outside of the unit circle.

To investigate which of these possibilities has occurred let z_ϵ be a zero of $h^*(z,\epsilon)$ near $z = 1$. If z_ϵ were complex it would have a conjugate also near $z = 1$

and taking the limit as ϵ tends to zero would make $z = 1$ a double zero of $h(z)$. However, $z = 1$ is only a simple zero of $h(z)$ since $h(z)$ approximates $\ln z$ in the neighbourhood of $z = 1$. Thus z_ϵ must be real for $\epsilon \ll 1$ and so we can write $z_\epsilon = 1 + \delta\epsilon + O(\epsilon^2)$ for real δ.

Now $h^*(z,\epsilon) = 0$ means that,

$$h(z_\epsilon) = \epsilon$$

and by Taylor expansion we have,

$$h(z_\epsilon) = \delta\epsilon h'(1) + O(\epsilon^2).$$

Therefore $\epsilon = \delta\epsilon h'(1) + O(\epsilon^2)$, which implies that,

$$\delta = 1/h'(1) + O(\epsilon) = 1 + O(\epsilon),$$

since $p \geq 1$ implies that $h'(1) = 1$.

Hence the expansion $z_\epsilon = 1 + \epsilon + O(\epsilon^2)$ follows and the zero at $z = 1$ has migrated from outside the unit circle. Therefore $h(z)$ has at most $s - 1$ zeros outside the unit circle. □

Observe that Theorem (2.4.1) means that no scheme with $s = 0$ may be stable.

We are now in a position to describe the order star theory necessary for the solution of the problem already

and taking the limit as ϵ tends to zero would make $z = 1$ a double zero of $h(z)$. However, $z = 1$ is only a simple zero of $h(z)$ since $h(z)$ approximates $\ln z$ in the neighbourhood of $z = 1$. Thus z_ϵ must be real for $\epsilon \ll 1$ and so we can write $z_\epsilon = 1 + \delta\epsilon + O(\epsilon^2)$ for real δ.

Now $h^*(z,\epsilon) = 0$ means that,

$$h(z_\epsilon) = \epsilon$$

and by Taylor expansion we have,

$$h(z_\epsilon) = \delta\epsilon h'(1) + O(\epsilon^2).$$

Therefore $\epsilon = \delta\epsilon h'(1) + O(\epsilon^2)$, which implies that,

$$\delta = 1/h'(1) + O(\epsilon) = 1 + O(\epsilon),$$

since $p \geqslant 1$ implies that $h'(1) = 1$.

Hence the expansion $z_\epsilon = 1 + \epsilon + O(\epsilon^2)$ follows and the zero at $z = 1$ has migrated from outside the unit circle. Therefore $h(z)$ has at most $s - 1$ zeros outside the unit circle. □

Observe that Theorem (2.4.1) means that no scheme with $s = 0$ may be stable.

We are now in a position to describe the order star theory necessary for the solution of the problem already

described. Noting that we now have the additional condition, that for stable schemes the approximation $h(z)$ must satisfy Theorem 2.4.1.

3. ORDER STARS AND A SATURATION RESULT

3.1 Order Stars

As previously stated, our aim is to show that the von Neumann condition and the pole condition imply bounds on the order of accuracy, p. We achieve this using a modification of the theory of order stars as originally introduced by Wanner, Hairer and Nørsett [Wa78].

Initially, the theory was derived as a means of determining A-acceptability of rational approximations, $R(z)$, to the exponential function. Rather than studying the stability region of the function $R(z)$, they considered areas of the complex plane defined by the function $S(z) = R(z)/e^z$. Obviously this function has the same zeros and poles as $R(z)$. The set A,

$$A := \{ z \in \mathbb{C} \mid |S(z)| > 1 \},$$

which was called the *order star*, reflects many essential properties of $R(z)$. These basic properties are described in three propositions, Propositions 2-4, [Wa78], which are proved by elementary complex analytic techniques. The shape of the order star is determined by these propositions depending on the location of the poles and zeros of R and the order of accuracy of the approximation. Imposition of

A-acceptability adds one more constraint on the order star after which the optimal configuration for stability can be worked out. Maximal accuracy can then be found by counting the number of sectors of the order star which approach the origin.

Since the publication of this innovative paper, there have been many advances in the theory. Iserles and Powell [Is81], in their investigation of A-acceptability of rational approximations interpolating the exponential, reconsidered the notion of fingers and dual fingers by introducing the idea of A -regions and D -regions. The *order property*, Proposition 3, [Wa78], was generalised for points of interpolation by $R(z)$ and, in so doing, the original proof was somewhat tightened. To exhibit the monotone behaviour of $\arg S(z)$ along oriented boundaries of A and D the original proposition concerning multiplicity was split in two.

Iserles, [Is83b], has extended the order star framework to cater for the analysis of approximations to functions which are analytic, except for isolated poles and essential singularities. The importance of the theory, along with its extension to Riemann surfaces, as a major tool in approximation problems, is clear by the number of applications that have been considered recently, for example in [Is83c,83d; Je81,82,83].

In the course of the evolution of the theory many of the terms, order star, finger, region, etc., have taken on

new definitions. We adopt the notation of [Is83b], which seems to be the most natural approach, where it is the overall picture that is called the order star rather than the set A . Therefore, the notation varies a little from that used by Iserles and Williamson, [Is84a], on which most of the following work is based.

The order star under consideration here is the one introduced by Iserles [Is82], which he calls an order star of the second kind [Is83b]. We define the function $\sigma(z)$ on the strip,

$$I := \{ z \mid |\mathrm{Im}\, z| \leqslant \pi \}$$

by,

$$\sigma(z) = h(e^z) - z \quad .$$

The essential properties of $h(z)$ are reflected in the sets

$$\begin{aligned} \mathbf{A} &= \{ z \in I : \mathrm{Re}\ \sigma(z) > 0 \} \\ \mathbf{D} &= \{ z \in I : \mathrm{Re}\ \sigma(z) < 0 \} \\ \partial &= \{ z \in I : \mathrm{Re}\ \sigma(z) = 0 \} \end{aligned} \quad , \qquad (3.1.1)$$

which together form a decomposition of the strip I. This decomposition is called the order star of σ . Connected components of $\mathbf{A}$ (and $\mathbf{D}$) are called $\mathbf{A}_0$-regions or $\mathbf{A}_\infty$-regions ($\mathbf{D}_0$-regions or $\mathbf{D}_\infty$-regions) according to whether they are bounded or unbounded.

The order stars for the following four examples, all of which are chosen to give maximum accuracy $p = r+s+R+S$ and

are derived from the relevant Padé approximations, are given in Figure 1.

a) $R = 1,\ S = 1,\ r = 0,\ s = 2, p = 4$

$$h(z) = \frac{27 - 24\,z - 3\,z^2}{\frac{1}{z} - 14 - 17\,z} .$$

It can easily be calculated that both zeros of the denominator lie inside the unit circle and that for $\theta = \pi/2$, $\mathrm{Re}\, h(e^{i\theta}) > 0$. Thus the scheme obtained satisfies neither the pole condition nor the von Neumann condition and so is unstable.

b) $R = 3,\ S = 1,\ r = 1,\ s = 1,\ p = 6$

$$h(z) = \frac{-2430/z + 1440 + 990z}{\frac{11}{z^3} - 104/z^2 + 1176/z + 2056 + 281z} .$$

It can be shown that the denominator of $h(z)$ has three zeros lying inside the unit circle, one outside the unit circle and therefore the pole condition is satisfied. However $\mathrm{Re}\, h(e^{i\theta}) > 0$ at $\theta = \pi$ so that the pole condition is violated and the scheme is unstable.

c) $R = 3,\ S = 0,\ r = 1,\ s = 0,\ p = 4$

$$h(z) = \frac{-24/z + 24}{1/z^3 - 5/z^2 + 19/z + 9}$$

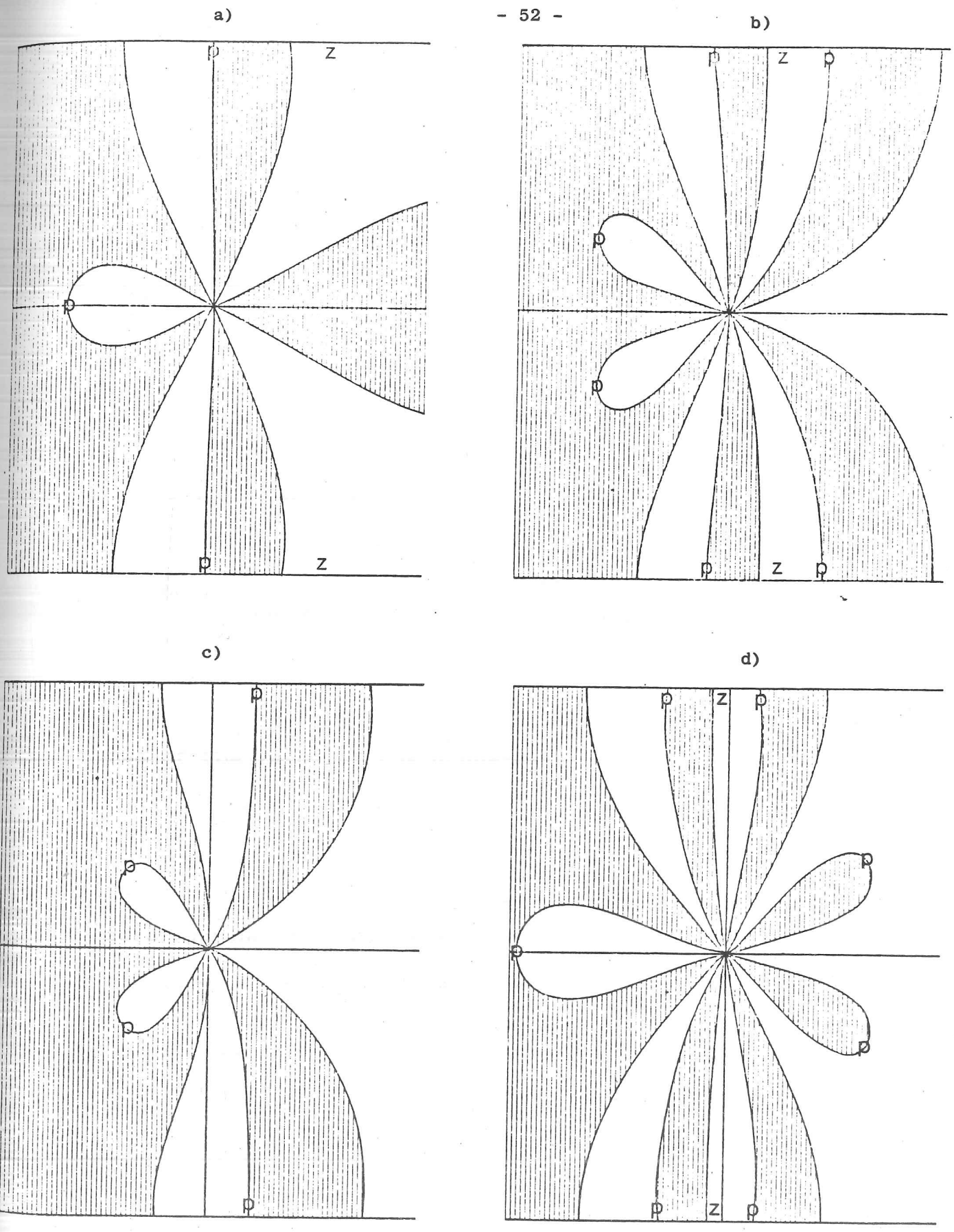


Figure 1. Examples of order stars. A is represented by the shaded area and 'p' and 'z' denote poles and zeros of h respectively.

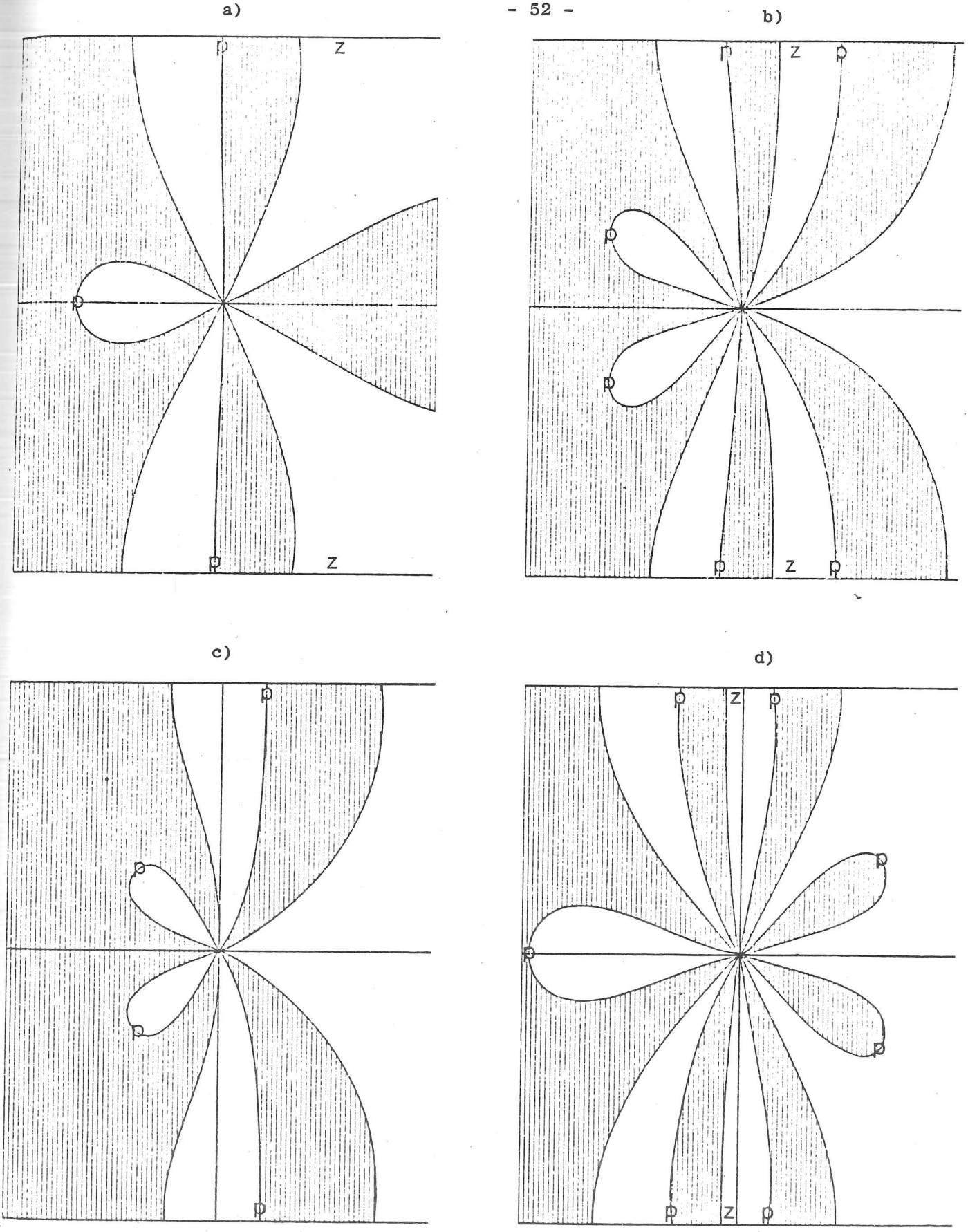


Figure 1. Examples of order stars. A is represented by the shaded area and 'p' and 'z' denote poles and zeros of h respectively.

As

$$\mathrm{Re}\, h(e^{i\theta}) = \frac{-12\,(1 - \cos\theta)^3}{65 + 11\cos\theta - 13\cos^2\theta + 9\cos^3\theta} \leqslant 0 \text{ for all } \theta ,$$

the von Neumann condition is satisfied. However it can be shown that two zeros of the denominator lie inside the unit circle and one outside. Thus the pole condition is violated and the scheme is unstable.

d) $R = 2,\ S = 3,\ r = 1,\ s = 1,\ p = 7$

$$h(z) = \frac{17190/z + 7200 - 24390z}{136/z^2 - 6029/z - 25784 - 10504z + 656z^2 - 55z^3} .$$

It can be shown that the pole condition is satisfied since two of the zeros of the denominator lie inside the unit circle and three outside. Also $\mathrm{Re}\, h(e^{i\theta}) \leqslant 0$ for all θ and so the von Neumann condition is satisfied. Therefore this scheme is stable. Note that, in addition, $h(z)$ has one zero inside the unit circle and one at $z = 1$ and thus as required it satisfies Theorem 2.4.1.

The geometric form of the order star is described in four Lemmas which are parallel to Lemmas 2.2 to 2.5 in [Is83a]. The proofs parallel those of Propositions 3.1 - 3.4 in [Is82], in which a different but analogous order star was considered. Here we just give a very brief outline of these proofs.

Lemma 3.1.1 The Order Property:

The scheme (2.2.5) (with K = 1) is of order of accuracy p only if for $z \to 0$ **A** consists of p+1 sectors of angle $\pi/(p+1)$ separated by p+1 similar sectors of **D**.

The proof of Lemma 3.1.1 follows immediately from the equation of accuracy (2.3.1) and the demonstration, as given by Iserles and Powell [Is81], that **A** and **D** do not contain sectors which are so thin that they fit between the sectors which must exist to satisfy (2.3.1). □

Lemma 3.1.2 The Pole Property:

Every pole of $\sigma(z)$ lies on ∂. Furthermore, each bounded **A** or **D** region has at least one pole of $\sigma(z)$ on its boundary. □

The proof follows from the maximum modulus principle for analytic functions or as a corollary to Lemma 3.1.5.

Lemma 3.1.3 The Essential Singularity Property:

i). If $s > S$ then for $\mathrm{Re}\, z \gg 0$ the line segment $[\mathrm{Re}\, z - i\pi, \mathrm{Re}\, z + i\pi]$ is composed of $2(s - S) + 1$ distinct intervals of **A** and **D**. If $s \leq S$ then $[\mathrm{Re}\, z - i\pi, \mathrm{Re}\, z + i\pi]$ belongs to **D** for $\mathrm{Re}\, z \gg 0$.

ii). If $r > R$ then for $\mathrm{Re}\, z \ll 0$ the line segment $[\mathrm{Re}\, z - i\pi, \mathrm{Re}\, z + i\pi]$ is composed of $2(r - R) + 1$ distinct intervals of **A** and **D**. If $r \leq R$ then $[\mathrm{Re}\, z - i\pi, \mathrm{Re}\, z + i\pi]$ belongs to **A** for $\mathrm{Re}\, z \ll 0$.

The proof of (ii) in Lemma 3.1.3 is the same as for (i) after setting $h_1(z) = h(1/z)$. For part (i) we consider $z = x + i\theta$, $x \gg 0$ and look at the dominant terms of $\mathrm{Re}\,\sigma(x + i\theta)$ as implied by the explicit form of $h(z)$. □

Lemma 3.1.4 The Stability Property

The SD scheme is stable if and only if $\mathbf{A} \cap [-i\pi, i\pi] = \phi$ and $\sigma(z)$ has R poles in $I^- := I \cap \{\mathrm{Re}\, z < 0\}$ and S poles in $I^+ := I \cap \{\mathrm{Re}\, z > 0\}$. Note that, $h(e^z)$ being periodic with period $2\pi i$, a pole lying at $x + i\pi$ and $x - i\pi$ is counted only once.

Lemma 3.1.4 is an immediate consequence of the stability conditions after transformation of the complex plane to the strip I by $w = e^z$. □

These four Lemmas can be used to derive a bound on the order of accuracy p of a stable scheme. However, we have not incorporated the extra condition that stability imposes on the location of the zeros, Theorem 2.4.1. With this in mind, the following property is evident.

Corollary of Theorem 2.4.1 The Zero Property

If the ISD is stable, then $h(e^z)$ has,

at most r zeros in I^-

at most s - 1 zeros in I^+ . □

A modification of part of Proposition 4 of [Wa78] as in [Is81] leads to a result about how the imaginary part of the

function σ changes along the boundary of ∂ . Thus we have the following Lemma, the proof of which is by examining the normal derivative of $\mathrm{Re}\,\sigma(z)$ and applying the Cauchy Riemann equations along ∂ .

Lemma 3.1.5

The imaginary part of σ decreases strictly monotonically along any part of the positively oriented boundary of an **A**-region and it increases strictly monotonically along any part of the positively oriented boundary of a **D**-region.

Proof

Let $\Sigma(z)$ be a function defined on the whole of the complex plane with the possible exception of isolated points. We apply the transformation $w = e^z$ to the complex plane and define

$$\tilde{\sigma}(z) = \ell n\ \Sigma(e^z) \text{ for } z \in I \quad .$$

Now

$$\Sigma(e^z) = |\Sigma(e^z)|\, e^{i \arg \Sigma(e^z)}$$

so that

$$\tilde{\sigma}(z) = \ell n\, |\Sigma(e^z)| + i \arg \Sigma(e^z) \quad .$$

Thus

$$\mathrm{Re}\,\tilde{\sigma}(z) = 0 \text{ if and only if } |\Sigma(e^z)| = 1$$

and

$$\mathrm{Re}\,\tilde{\sigma}(z) \gtrless 0 \text{ if and only if } |\Sigma(e^z)| \gtrless 1.$$

Therefore *A*-regions of $\Sigma(z)$ defined by

$$A = \{ z \in \mathbb{C} : |\Sigma(z)| > 1 \}$$

transform to **A**-regions of $\tilde{\sigma}(z)$, defined as in (3.1.1). Obviously *D*-regions transform similarly. Also

$$\mathrm{Im}\,\tilde{\sigma}(z) = \arg \Sigma(e^z)$$

which is by Proposition 4 of [Wa78] a strictly monotone function. Thus the desired result follows by identifying $\tilde{\sigma}$ with σ. □

The monotonicity of $\mathrm{Im}\,\sigma$ along the oriented boundary of ∂ means that between any two interpolation points of $\sigma(z)$, which necessarily belong to ∂, there must be at least one pole of $h(e^z)$ since $\mathrm{Im}\,\sigma$ vanishes at interpolation points and is unbounded at poles. This result can be further refined in a way which although strictly unnecessary to our analysis, adds to the understanding of the geometry of the order star. Let us define the *WHN number* of a region

as the number of times the boundary of the region crosses an interpolation point of the function $\sigma(z)$, counting an interpolation point the number of times that the boundary crosses it. Then the WHN number of a region is at most the number of poles of $h(e^z)$ lying on its boundary.

When we try to reconcile accuracy with stability we use the above Lemmas as in [Is83a] and count the maximum number of **A**- and **D**-regions which may reach the origin from I^- or I^+. To do this we must find the maximal number of poles which can account for two **A**-sectors adjoining the origin.

First we define a bounded portion of $\partial U(\mathbb{R} \pm i\pi)$ as a *loop* if it is a closed simple curve. By Lemma 3.1.5 poles of $h(e^z)$ and zeros of σ (i.e. interpolation points) interlace along each loop. Thus, we say that a pole of $h(e^z)$ is *efficient* if

a) it lies on $\mathbb{R} \pm i\pi$;

b) it belongs to loops all of which approach the origin;

c) there are no extra poles that lie along these loops.

Intuitively speaking, this means that as a pole lying on $\mathbb{R} \pm i\pi$ can be counted twice, an efficient pole "accounts" for two sectors of **A** that adjoin the origin. Two is the maximal number of **A**-sectors approaching the origin which can be accounted for by a single pole.

Lemma 3.1.6

The number of efficient poles N in any interval

$(x_1 \pm i\pi,\ x_2 \pm i\pi)$ is bounded by,

$$N \leqslant \min \{ Z + 1, P \}$$

where P and Z are the number of poles and least number of zeros respectively of the function $h(e^z)$ along $\mathbb{R} \pm i\pi$ in the given interval.

Proof

Because $\mathrm{Im}\,\sigma(z) = \mp \pi$ for every $z \in \mathbb{R} \pm i\pi$, the monotonicity of $\mathrm{Im}\,\sigma(z)$ as given by Lemma 3.1.5 implies that an efficient pole lies between a D-region to the left and an A-region to the right. Since $\mathrm{Re}\,\sigma(x \pm i\pi) = h(-e^x) - x$, it is clear that $h(-e^x) > x$ in any A-region, and $h(-e^x) < x$ in any D-region. Furthermore, $h(-e^x)$ becomes unbounded at the poles. Hence, due to the continuity of the function $h(-e^x)$ away from the poles, there must be a zero of $h(-e^x)$ in the interval between any two efficient poles. The bound now follows, $N \leqslant P$ being trivial. □

We apply this Lemma in the following section to reduce the theoretical limit on accuracy derived by taking into consideration only the position of the poles.

3.2 An Upper-Bound on Accuracy for Stability

Here we use the results of the previous section to obtain an upper bound on accuracy for all possible choices of R, r, S and s . Applying Lemmas 3.1.1 to 3.1.4 as in

[Is83a] leads to the bound,

$$p \leq \min \{ 2((r-R)_+ + 2R + 1),\ 2((s-S)_+ + 2S + 1),\ (m-n)_+ + 2n \}$$

where,

$$m = r + s,\ n = R + S \text{ and } (X)_+ := \max \{ X, 0 \} \quad .$$

This seems to imply that the more implicit methods, i.e. those with $R > r$ and $S > s$, will be better since for a given number of degrees of freedom, higher accuracy, with stability, will be attainable. However if $m = 0$ the above bound is extremely generous since such a method cannot be consistent, $p = 0$. A more realistic bound is derived using Lemma 3.1.6 to take into account the location of the zeros as well as the poles.

Lemma 3.2.1

a) If the ISD scheme is stable then $p \leq 2(S + s)$;

b) If the ISD scheme is stable then $p \leq 2(R + r + 1)$.

Proof

a) We bound from above the number of sectors of **A** which may reach the origin from I^+ or I^-. Let

M^+ := the number of sectors of $\mathbf{A}$ reaching the origin from I^+ ;

M^- := the number of sectors of $\mathbf{A}$ reaching the origin from I^- .

Then by Lemma 3.1.1, the order condition, it follows that

$$p + 1 = M^+ + M^-$$

$$\text{and} \quad M^- - 1 \leq M^+ \leq M^- + 1 \quad . \tag{3.2.1}$$

Also define Q^+ and Q^- as the number of sectors of $\mathbf{A}$ reaching infinity in I^+ and I^- respectively and N^+ and N^- as the number of efficient poles in I^+ and I^- respectively. By stability there are R poles in I^- and S poles in I^+ . Now each efficient pole may contribute to at most two $\mathbf{A}_0$-sectors reaching the origin and every $\mathbf{A}_0$-sector that reaches the origin must contain at least one pole along each loop on its boundary. Also if $K \geq Q^- + 1$ sectors of $\mathbf{A}_\infty$ approach the origin from I^-, say, then they must enclose among them at least $K - Q^-$ $\mathbf{D}_0$-regions, none of which may approach $\mathbb{R} \pm i\pi$. Each such $\mathbf{D}_0$-region necessarily contains a pole on its boundary, that may not be efficient. Therefore, it follows that,

$$M^+ \leq 2N^+ + (S - N^+) + Q^+$$

$$\text{and} \quad M^- \leq 2N^- + (R - N^-) + Q^- \quad . \tag{3.2.2}$$

As stability implies that there are at most r zeros in I^- and $s-1$ zeros in I^+, the bound of Lemma 3.1.6 gives,

$$N^+ \leq \min\{s, S\}$$

$$\text{and} \quad N^- \leq \min\{r+1, R\} \quad .$$

To prove part a) we consider the number of sectors reaching the origin from I^+. Note that by Lemma 3.1.5 in the case of an A_∞-region in I^+ adjoining the line $\mathrm{Im}\, z = \pm\pi$, this region must have a pole on its boundary that cannot be efficient even if it lies on the line $\mathrm{Im}\, z = \pm\pi$. Thus in this case, which by Lemma 3.1.3 may only occur if $s > S$, the number of efficient poles is bounded by,

$$N^+ \leq \min\{s, S-1\}$$

$$\text{and} \quad N^+ + Q^+ \leq \min\{s, S-1\} + (s-S)_+ + 1 = s \quad .$$

Alternatively an A_∞-region in I^+ is not bisected by the line $\mathrm{Im}\, z = \pm\pi$ and

$$N^+ + Q^+ \leq \min\{s, S\} + (s-S)_+ = s \quad .$$

Therefore by (3.2.2),

$$M^+ \leq s + s$$

and by (3.2.1),

$$p \leq 2(s + S) \quad .$$

For part b) we notice that the sign of $\mathrm{Re}(h(e^z))$ changes from minus to plus when passing through an efficient pole on the line $\mathrm{Im}\, z = \pi$, and that stability gives $\mathrm{Re}\, h(e^{ix}) \leq 0$. Thus, associated with every efficient pole in I^- is at least one zero and therefore,

$$N^- \leq \min \{ r,R \} \quad .$$

Proceeding as in part a) we obtain,

$$M^- \leq \min \{ r,R \} + R + (r - R)_+ + 1 = r + R + 1$$

and by (3.2.1),

$$p \leq 2(r + R + 1) \quad .$$

□

Note that the proof of part a) above reduces the limit from $2(S + s + 1)$ to $2(S + s)$ as mentioned in the introduction. The crucial difference in the proof (from that in [Is83a]) comes from looking at $\mathrm{Im}\, \sigma$ along the positively oriented boundary of an A-region in I^+ which reaches both

the origin and the line $\mathbb{R} \pm i\pi$.

Lemma 3.2.2

i) If $r \geq R$, $s \geq S$ or $R \geq r$, $S \geq s$, then $p \leq m + n$.

ii) If the SD scheme is stable, then $p \leq m + n$ regardless of the values of r, R, s and S .

Proof

Note immediately that for every $r \geq R$ and $s \geq S$ the result is already known [Ba75].

We proceed as in Lemma 3.2.1, to bound from above the total number K of sectors of **A** and **D** which may reach the origin. By Lemma 3.1.6 the total number of efficient poles is bounded by,

$$N \leq \min \{ r + s, R + S \} ,$$

since the total number of poles is $R + S$ and zeros in $\mathbb{C} / \{0\}$ is $r + s - 1$. Furthermore, by Lemma 3.1.3 the number of A_∞- and D_∞-sectors is $2(s - S)_+ + 2(r - R)_+ + 2$. Therefore,

$$K \leq 4N + 2(R + S - N) + 2(s - S)_+ + 2(r - R)_+ + 2$$

$$= 2(R + S) + 2N + 2(s - S)_+ + 2(r - R)_+ + 2$$

as each efficient pole can contribute to at most four bounded sectors which reach the origin and each inefficient pole to two sectors of **A** and **D** there.

Therefore, recalling the order property we must have $K = 2(p + 1)$ and the lemma is true for the choices $r \geq R$ and $s \geq S$ or $S \geq s$ and $R \geq r$.

In the other two cases, we assume stability and use the results of the proof of Lemma 3.2.1 to bound the number of sectors approaching the origin from I^+ and I^- separately.

It follows from the proof of Lemma 3.2.1 that,

$$N = N^- + N^+ \leq \min \{ r,R \} + \min \{ s,S \} \quad .$$

Hence, as before,

$$p + 1 = \frac{K}{2} \leq R + S + N + (s - S)_+ + (r - R)_+ + 1 \leq r + s + R + S + 1,$$

giving the required bound. □

Lemmas 3.2.1 and 3.2.2 combine to prove Theorem 3.2.1, which is the main result of this Chapter:

Theorem 3.2.1

If the SD scheme is stable, then $s \geq 1$ and

$$p \leqslant \min \{ r + s + R + S,\ 2(r + R + 1),\ 2(s + S) \}. \quad (3.2.3)$$

□

We shall see in the next Chapter that for some choices of r, s, R and S it is possible to derive schemes which do attain this bound and thus sometimes it is optimal. For schemes which attain the maximal accuracy, $p = m + n$, the inequality (3.2.3) requires that, as for the FD case [Is83a], they are sufficiently centred:

$$r + R \leqslant s + S \leqslant r + R + 2 \quad . \quad (3.2.4)$$

Again, stability cannot occur away from the three leading diagonals of the Padé table.

An important consequence of Lemma 3.2.2 concerns the normality of Padé approximations to $z^L \ln z$. Let $H(z) = G(z)/F(z)$ be the Padé approximation to $z^L \ln z$ at $z = 1$ where $m = r + s$ is the degree of $G(z)$, $n = R + S$ is the degree of $F(z)$ and $L = r - R$.

Corollary 3.2.1

The Padé approximations to $z^L \ln z$ are normal for,

a) $m \geqslant n + L$ and $L \geqslant 0$;

b) $n \geqslant m - L$ and $L \leqslant 0$.

Proof

By part (i) of Lemma (3.2.2), $p \leqslant m + n$. Therefore the desired result follows by Padé theory [Ba75]: since the maximal blocksize is one, necessarily $p = m + n$ and the approximations are normal. □

Note that this result is already known when $L = o$ and $m \geqslant n$ [Ba75].

Padé approximations are the natural choice of rational approximations to consider when only the degree of numerator and denominator are specified, since they use all the available degrees of freedom to satisfy the order conditions. Knowing that in the two above cases the Padé approximations are normal, we can apply the Padé theory to derive the error constants of such approximations. It is demonstrated that this is particularly useful in the following Chapter.

4. PADE SCHEMES

4.1 Introduction

In the previous chapters we demonstrated that an upper bound on the accuracy attainable by a stable scheme does exist. Here we try to decide in what sense, if any, this bound can be optimal.

For the completely explicit schemes, $R = S = 0$, attainment of the bound was demonstrated by Iserles [Is82]. Application of the Lagrange interpolation formula produced interpolatory formulae of highest order, $r + s$. The implicit case is obviously more difficult as rational, rather that polynomial, approximations are required. As previously mentioned, the natural candidates are the Padé approximations which do attain maximal accuracy for their degree. Following the notation of Iserles and Strang [Is83a], we call schemes derived from the relevant Padé approximations *Padé schemes*.

When considering fully discretised schemes approximations to z^{λ}, $\lambda = \tilde{r} - \tilde{R} + \mu$ about $z = 1$ are needed. Fortunately the polynomials $P_{M/N}$ and $Q_{M/N}$, $M = \tilde{r} + \tilde{s}$, $N = \tilde{R} + \tilde{S}$, of the [M/N] Padé approximations to z^{λ} have already been calculated as the limits of hypergeometric functions [Is79]. Thus verification of the von Neumann and pole conditions in special cases is possible:- the von

Neumann condition by explicit calculation of the difference $D^* = |Q(e^{i\theta})|^2 - |P(e^{i\theta})|^2$ and the pole condition by identification of $Q(z)$ as a multiple of a Jacobi polynomial. Thus Iserles and Strang [Is83a] were able to characterise all the stable schemes in which the number of points at the two time levels differed by at most one. These are the Padé schemes derived from approximations lying on the three central diagonals of the Padé tableau. Away from the centre of the Padé tableau, the hypergeometric identity for D^* becomes increasingly complicated, whilst the determination of the zeros of the Jacobi polynomial can no longer be treated by classical orthogonal polynomial theory. However, the case with $\tilde{R} = \tilde{S}$ was successfully resolved for all M and N.

In what follows we make much use of the results derived in the above mentioned paper to characterise stable SD methods which are associated with fully discrete methods. Also, having proved in Corollary 3.2.1 that Padé approximations are normal, we can consider other choices which are more implicit that explicit. Without calculating the Padé approximations explicitly, we can apply Padé theory to derive expressions for the error constants and order star theory to determine the location of the poles and zeros. Consequently we are able to prove stability.

4.2 Optimal Schemes

Initially we consider those schemes which can be derived by a limiting process as associated schemes and thus have $r \geqslant R$, $s \geqslant S$. Since Padé approximations to z^{λ} , $\lambda = \tilde{r} - \tilde{R} + \mu$ can be derived as a quotient of Jacobi polynomials, the Padé approximation to $z^{L} \ln z$ for $L \geqslant 0$ can be obtained as the derivative of this quotient. We have already stated that stability in the FD case is verified without calculating the approximation explicitly. As a result the connection between the stability conditions for the FD and SD cases as discussed in Section 2.3 enables us to verify stability in a similar way.

The stability conditions for the FD case can be expressed as follows.

1) The von Neumann stability condition is equivalent to

$$D^{*} = | Q(e^{i\theta},\lambda) |^{2} - | P(e^{i\theta},\lambda) |^{2} \geqslant 0, \quad 0 \leqslant \theta \leqslant 2\pi \quad ,$$

where,

$$\frac{P(z,\mu)}{Q(z,\mu)} = z^{\tilde{r}-\tilde{R}}\, a(z,\mu) = z^{\tilde{r}-\tilde{R}+\mu} + O(|z-1|^{p+1}),$$

$\lambda = \tilde{r} - \tilde{R} + \mu$ and $a(z,\mu)$ is the characteristic function of the IFD scheme.

2) The pole condition means that the Jacobi polynomial $P_N^{(\alpha,\beta)}(z)$ where $\alpha = \tilde{r} - \tilde{R} + \mu$, $\beta = \tilde{s} - \tilde{S} + \mu$ and $N = \tilde{S} + \tilde{R}$, has $\tilde{R}$ zeros inside and $\tilde{S}$ zeros outside the complex unit circle (Section 3, Theorem 3, [Is83a]).

Therefore for those SD schemes which are associated with FD schemes we have similar conditions for stability.

Lemma 4.2.1

For an SD Padé scheme with $r \geq R$ and $s \geq S$, the two requirements for stability are equivalent to,

a) the von Neumann condition: $\frac{\partial D^*}{\partial \mu}\Big|_{\mu \to 0} \geq 0$, where D^* is defined as above, and

b) the pole condition: The Jacobi polynomial, $P_n^{(\alpha,\beta)}(z)$ where $\alpha = r - R$, $\beta = s - S$ and $n = R + S$ has

R zeros in the right half plane $\mathrm{Re}\, z > 0$

and

S zeros in the left half plane $\mathrm{Re}\, z < 0$.

Proof

The proof of the von Neumann condition is from equation (2.3.7), which relates the satisfaction of this condition by SD and FD schemes, and by application of l'Hôpital's rule.

For the pole condition we saw from equation (2.3.5) that the denominator, $F(z)$ of $z^{r-R}h(z)$, could be identified with $Q(z,0)$. Thus $F(z)$ is a Möbius transformation of the required Jacobi polynomial and the result follows. □

Applications of the above Lemma leads to immediate verification of stability in the cases $m = n + 1$ and $m = n$.

Theorem 4.2.1

The only stable Padé schemes with $m = n + 1$ and $r \geq R$, $s \geq S$ are given by $\{ r = s = S = R + 1 \}$ and $\{ R = S = r = s - 1 \}$.

Proof

The choice $m = n + 1$ with $r \geq R$ and $s \geq S$ gives only two cases to consider, either $\{ r = R + 1$ and $s = S \}$ or $\{ s = S + 1$ and $r = R \}$. However by Theorem 3.2.1 these schemes must be centred since they are Padé. Thus the only possibilities are $\{ r = s = S = R + 1 \}$ and $\{ R = S = r = s - 1 \}$. Therefore we require to prove that either choice is stable. By (2.3.6) the associated FD schemes are given by,

$$s = \max \{ \tilde{s}, \tilde{S} \} \quad S = \min \{ \tilde{s}, \tilde{S} \}$$

and

$$r = \max \{ \tilde{r}, \tilde{R} \} \quad R = \min \{ \tilde{r}, \tilde{R} \} \quad .$$

Without loss of generality we can assume that $s = \tilde{s}$, $S = \tilde{S}$, $r = \tilde{r}$ and $R = \tilde{R}$. Therefore we must consider the two cases :

a) $\tilde{r} = \tilde{s} = \tilde{S} = \tilde{R} + 1$ and

b) $\tilde{R} = \tilde{S} = \tilde{r} = \tilde{s} - 1$.

Proof of stability for a) and b) is similar and hence we only present the proof for choice a). Recalling that $M = \tilde{r} + \tilde{s}$ and $N = \tilde{R} + \tilde{S}$, $M = N + 1$. Also, M is even, since $\tilde{r} = \tilde{s} = \tilde{S} = \tilde{R} + 1$. Therefore, from [Is83a],

$$D^* = \frac{(-1)^{M+1}(N!)^2 (-N-1-\mu)_M (-M+1+\mu)_M X^M}{[(M+N)!]^2} ,$$

where $X = 2(1 - \cos\theta)$.

Now forming the derivative of D^* with respect to μ and setting $\mu = 0$ we have

$$\frac{\partial D^*}{\partial \mu} \Big|_{\mu = 0} = \frac{(-1)^M M! (N!)^3 X^M}{[(M+N)!]^2} \geq 0 .$$

By Lemma 4.4 [Is83a] the polynomial $P_N^{(1,0)}(z)$ has $[N/2]$ zeros in $(0,1)$ and $[(N+1)/2]$ zeros in $(-1,0)$. The pole condition is thus automatically satisfied and the scheme is stable. □

Theorem 4.2.2

The only stable Padé schemes with $m = n$, $r = R$ and $s = S$ are given by $r = R = s = S$ and have coefficients:

$$\alpha_j := \tilde{h}_m^{-1}(\tilde{h}_j - \tilde{h}_{m-j}) \binom{m}{j}^2$$

$$\beta_j := \tfrac{1}{2}\tilde{h}_m^{-1} \binom{m}{j}^2$$

where

$$h(z) = \frac{\sum_{j=0}^{m} \alpha_j z^j}{\sum_{j=0}^{m} \beta_j z^j}$$

and

$$\tilde{h}_j := \sum_{k=1}^{j} 1/k \quad , \quad \tilde{h}_0 := 0 \quad .$$

Proof

According to [Da56] the above-mentioned coefficients give a method of accuracy $2n$. Hence, by Corollary 3.2.1, they correspond to the $[n/n]$ Padé scheme.

Now $r = R$ and $s = S$ imply that $\tilde{r} = \tilde{R}$ and $\tilde{s} = \tilde{S}$. Therefore $M = N$ and so by [Is83a] $D^* = 0$. Consequently the von Neumann condition is satisfied. For the pole condition we require the zeros of $P_{R+S}^{(0,0)}(z)$. This is a Legendre polynomial whose zeros lie in $(-1,1)$

and are symmetrically situated around the origin. By equation (3.2.4) this symmetry is consistent with stability only if $R = S$ and so the only stable scheme has $r = R = s = S$. □

It is evident that Theorems 4.2.1 and 4.2.2 are immediate corollaries of Theorems 4B and 4A of Iserles and Strang [Is83a]: stability of the FD implies stability of the SD, accuracy is inherited since both schemes are Padé, $\tilde{p} = p$, and the result follows.

In the same way we have an immediate corollary of Theorem 5 in the same paper. For the Padé schemes with $R = S$, $r \geq R$, $s \geq S$, the inequalities (3.2.4) are sufficient as well as necessary.

Now we turn to the more difficult problem of determining the stability of SD schemes which are not associated with any FD scheme. For these methods we do not know anything about the explicit form of the Padé approximation. However Padé theory and some lengthy analysis enables the investigation of the schemes lying above the diagonal in the Padé tableau which have $n = m + 1$ and $R \geq r$, $S \geq s$.

Theorem 4.2.3

The only stable Padé schemes with $n = m + 1$, $R \geq r$ and $S \geq s$ are given by $R = S = s = r + 1$ and $R = s = r = S - 1$.

Proof

The proof falls into three parts: first to show that the von Neumann condition depends on the sign of the error constant, second to find this sign and finally to examine the location of the poles and zeros using the geometric properties of the order star. As in Theorem 4.2.1 there are just two possibilities for $n = m + 1$ with $R \geqslant r$ and $S \geqslant s$ satisfying Theorem 3.2.1:

a) $r = R = s = S - 1$; and

b) $R = r + 1 = s = S$.

We will now prove that both choices are stable.

i) The von Neumann condition

For $r = R$,

$$\mathrm{Re}\, h(e^{i\theta}) = \mathrm{Re}\, \frac{F(e^{-i\theta})\, G(e^{i\theta})}{|F(e^{i\theta})|^2} ,$$

where $F(z)$ and $G(z)$ are the denominator and numerator of $H(z) = z^{r-R} h(z)$.

Therefore

$$\mathrm{Re}\, h(e^{i\theta}) \leqslant 0 \text{ if and only if } \mathrm{Re}\, F(e^{-i\theta}) G(e^{i\theta}) \leqslant 0.$$

Let c be the error constant of the approximation. Then

$$h(z) = \ln z + c(z-1)^{p+1} + O(|z-1|^{p+2}),$$

hence

$$h(e^{i\theta}) = i\theta + c(i\theta)^{2m+2} + O(\theta^{2m+3}).$$

Thus

$$\mathrm{Re}\, F(e^{-i\theta})G(e^{i\theta}) = (-1)^{m+1}c|F(e^{i\theta})|^2\,\theta^{2(m+1)} + O(\theta^{2m+3}).$$

Substituting

$$(1-\cos\theta) = \theta^2/2 + O(\theta^4) \text{ and } F(e^{i\theta}) = F(1) + O(\theta)$$

implies that

$$\mathrm{Re}\, F(e^{-i\theta})G(e^{i\theta}) = c(-1)^{m+1}2^{m+1}|F(1)|^2(1 - \cos\theta)^{m+1}$$

$$+ O((1 - \cos\theta)^{m+3/2}) \qquad (4.2.1)$$

But from (2.3.4)

$$\mathrm{Re}\,(F(e^{-i\theta})\,G(e^{i\theta})) = \mathrm{Re} \sum_{j=o}^{m} g_j e^{ij\theta} \sum_{k=o}^{n} f_k e^{-ik\theta}$$

$$= \sum_{j=o}^{m+1} r_j (1 - \cos\theta)^j$$

$$= R(1 - \cos\theta) \quad , \qquad (4.2.2)$$

where R is a polynomial in $(1 - \cos\theta)$ of degree $m + 1$.

Equating (4.2.1) and (4.2.2) implies that

$$R(1-\cos\theta) = c(-1)^{m+1}2^{m+1}|F(1)|^2(1 - \cos\theta)^{m+1}$$

where the higher order terms vanish since R has degree at most $m + 1$.

Therefore $R(1 - \cos\theta) \leq 0$ if and only if $(-1)^{m+1}c \leq 0$.

Similarly, for the second case, where $R = r + 1$, we obtain

$$\mathrm{Re}\, h(e^{i\theta}) \leq 0 \text{ if and only if } (-1)^{m+1}c \leq 0.$$

The von Neumann condition thus depends on the sign of the error constant and the parity of m.

b) <u>The error constant</u>

By Corollary 3.2.1 the approximations under consideration are normal and so Padé theory can be used to find the error constants in terms of the ratio of the determinants of two matrices whose elements are obtained from the Taylor series expansion of the underlying function [Ba75].

For $r = R$,

$$c = - \frac{\det A^{m+2}}{\det A^{m+1}},$$

where A^{m+2} is the Hankel matrix:

$$(A^{m+2})_{ij} = \begin{cases} 0 & i = j = 0 \\ \dfrac{(-1)^{i+j-1}}{i+j} & \text{otherwise, } 0 \leq i,j \leq m+1 \end{cases}$$

By using elementary row and column operations and then the Sylvester determinant identity [Ba75], the determinant of A^{m+2} can be obtained from a recurrence relation with coefficients that are Cauchy matrices X^k [Gr69, p.54]. The X^k are k-by-k matrices with

$$(X^k)_{ij} = \frac{1}{x_i + y_j} \quad , \quad x_i = y_i = i+1 \quad . \qquad (4.2.3)$$

By solving the recurrence relation we obtain,

$$\det A^{m+2} = 2\,(-1)^{m+1} \left[\sum_{j=0}^{m} \frac{[(m-j)!]^3}{(2m-2j+1)!} \right] \det X^{m+1}$$

$$= (-1)^{m+1} K_{m+2} \quad .$$

Now $x_j > x_i$ and $y_j > y_i$ for all $1 \leq i < j \leq m+1$ means that $\det X^{m+1} > 0$ [Gr69]. Therefore $K_{m+2} > 0$ and so

$$c = \frac{-(-1)^{m+1} K_{m+2}}{(-1)^m K_{m+1}} > 0 \text{ for all } m \quad .$$

The von Neumann condition is therefore satisfied for choice a): $m = r + s$ being even, $c > 0$ implies $(-1)^{m+1} c \leqslant 0$.

For $R = r + 1$,

$$c = - \frac{\det H^{m+2}}{\det H^{m+1}}$$

where,

$$(H^{m+2})_{ij} = \begin{cases} 0 & i = j = 0 \\ (-1)^{i+j+1} \tilde{h}_{i+j} & \text{otherwise, } 0 \leqslant i,j \leqslant m+1 \end{cases}$$

As before we use elementary row and column operations to express the determinant of H^{m+2} in terms of a more convenient matrix B so that,

$$\det H^{m+2} = (-1)^{m+2} \det B$$

where,

$$B_{ij} = \begin{cases} \frac{-1}{(i+j+1)(i+j+2)} & 0 \leqslant i , j \leqslant m \\ \frac{1}{j+1} & i = m+1 , 0 \leqslant j \leqslant m \\ \frac{1}{i+1} & j = m+1 , 0 \leqslant i \leqslant m \\ 0 & i = j = m+1 \end{cases}$$

It is possible to prove an explicit expression for $\det B$ inductively by finding a one stage recurrence relation for $\det B^q$ in terms of $\det B^{q+1}$. In this way we can derive an exact expression for c. However our proof only relies on finding the sign of c and so we adopt a different and more elegant technique. Consider the matrix C^{m+1} given by,

$$(C^{m+1})_{ij} = \frac{1}{(i+j+1)(i+j+2)} \quad 0 \leqslant i,j \leqslant m \quad .$$

Obviously $-C^{m+1}$ is the leading principal minor of B and all the principal minors of C^{m+1} are of the same form, namely, $C^1, C^2, \ldots, C^{m+1}$. Also it is symmetric and thus positive definite if and only if all of its principal minors are positive. Consequently we require only to demonstrate that C^{m+1} has a positive determinant for any $m \geqslant 0$ to prove that it is positive definite.

It is possible to express the determinant of C^{m+1} in terms of a new matrix D^{m+1} by letting the k-th row of the new matrix be the sum of all the rows up to and including the k-th row:

$$(D^{m+1})_{ij} = \frac{i+1}{(j+1)(i+j+2)} \quad 0 \leqslant i,j \leqslant m+1 \quad .$$

Taking a factor $(i+1)$ from the i-th row and a factor $\frac{1}{j+1}$ from the j-th column, $0 \leqslant i,j \leqslant m+1$, we see that

$$\det D^{m+1} = \det X^{m+1} \quad ,$$

where X^{m+1} is the Cauchy matrix given by (4.2.3). Recalling that the determinant of this Cauchy matrix is positive, the matrix C^{m+1} is positive definite and all of its eigenvalues are positive.

The eigenvalues of the leading principal minor of the matrix B are thus all negative. Therefore by the separation theorem for eigenvalues [Wi65, p.103], B has at most one non-negative eigenvalue. Hence B is non-singular and has a single positive eigenvalue because $x^T B x = 1.5 > 0$ for $x^T = (1,0,.....,0,1)$. The determinant of B is equal to the product of its eigenvalues, consequently,

$$\text{sgn (det B)} = \text{sgn} \left[\prod_{i=0}^{m+1} \lambda_i \right]$$

$$= (-1)^{m+1} \quad ,$$

where $\{ \lambda_i : 0 \leq i \leq m+1 \}$ are the eigenvalues of B. Therefore the matrices H^{m+1} all have negative determinants and

$$c < 0 \quad .$$

Now, $m = r + s$ being odd and $c < 0$ imply that $(-1)^{m+1} c < 0$. Thus by part i) the von Neumann condition is satisfied.

iii) The pole condition

In case a) m is even. We let $m = 2\ell$ and by Lemma 3.1.1 there are $8\ell + 4$ sectors of A and D approaching the origin of the order star. The sign of the error constant means that along the x axis,

$$h(e^x) - x = cx^{p+1} + O(x^{p+2}) > 0 \text{ for } 0<|x| \ll 1 \ .$$

Thus the x axis must bisect an A-region and as there are $2\ell + 1$ sectors in each quadrant, the y axis must bisect a D-region. Let

M^+:= the number of A -regions approaching the origin from I^+ ;

M^-:= the number of D -regions approaching the origin from I^-;

P^+:=the number of poles in I^+;

P^-:=the number of poles in I^-.

Then

$M^+ = 2\ell + 1$ and by Lemma 3.1.2 $P^+ \geq \ell + 1$;

$M^- = 2\ell$ and by Lemma 3.1.2 $P^- \geq \ell$.

It follows from $P^+ + P^- = m + 1 = 2\ell + 1$ that $P^+ = \ell + 1$ and $P^- = \ell$.

Because the pole condition requires $P^+ = S$, and $P^- = R$, the only stable configuration occurs for $R = r = s = \ell$, $S = \ell + 1$.

iii) The pole condition

In case a) m is even. We let $m = 2\ell$ and by Lemma 3.1.1 there are $8\ell + 4$ sectors of A and D approaching the origin of the order star. The sign of the error constant means that along the x axis,

$$h(e^x) - x = cx^{p+1} + O(x^{p+2}) > 0 \text{ for } 0<|x| \ll 1 \ .$$

Thus the x axis must bisect an A-region and as there are $2\ell + 1$ sectors in each quadrant, the y axis must bisect a D-region. Let

M^+:= the number of A -regions approaching the origin from I^+ ;

M^-:= the number of D -regions approaching the origin from I^-;

P^+:=the number of poles in I^+;

P^-:=the number of poles in I^-.

Then

$M^+ = 2\ell + 1$ and by Lemma 3.1.2 $P^+ \geq \ell + 1$;

$M^- = 2\ell$ and by Lemma 3.1.2 $P^- \geq \ell$.

It follows from $P^+ + P^- = m + 1 = 2\ell + 1$ that $P^+ = \ell + 1$ and $P^- = \ell$.

Because the pole condition requires $P^+ = S$, and $P^- = R$, the only stable configuration occurs for $R = r = s = \ell$, $S = \ell + 1$.

For $r + 1 = R$, the method of proof follows as above. Since m is odd, $m = 2\ell + 1$, and the only stable configuration occurs for $R = S = s = \ell + 1$, $r = \ell$. □

For a general choice of R, r, s and S there is, at present, no complete analysis. We have relied on the connection between the fully discretised and semi-discretised schemes to derive stability in a few selected cases when the scheme is more explicit. For schemes which are more implicit, we have only looked at the case when $n = m + 1$. This analysis relies heavily on the results of Padé theory to investigate the sign of the error constant of the approximation. Given the sign of the error constant, we are able to use the geometry of the order star for this particular case to demonstrate which schemes may at the same time be both von Neumann stable and satisfy the pole condition.

4.3 Conclusion

In the preceding chapters we have investigated the stability and accuracy of finite-difference methods for solving the linear conservation law, $u_t = u_x$. Through a modification of order star theory, we have completed the work begun by Iserles [Is82] and continued by Iserles and Strang [Is83a] to derive an upper bound on accuracy for all stable semi-discretisations of such equations:

$$p \leqslant \min \{ r + s + R + S,\ 2(r + R + 1),\ 2(s + S) \} \quad .$$

Furthermore, we have demonstrated for numerous choices of r, s, R and S that this bound is optimal. Certain Padé approximations achieving the bound are stable. As with the FD case, stability only occurs if the Padé scheme is sufficiently centred:

$$r + R \leqslant s + S \leqslant r + R + 2 \quad .$$

For approximations which are away from the centre of the Padé tableau, these inequalities still leave many cases to be considered. Therefore we are unable to conjecture what the right conditions for stability are, either for cases away from the three main diagonals of the Padé tableau, or for those cases which use an unbalanced number of explicit and implicit points $R \geqslant r$, $S < s$ and $R \leqslant r$, $S \geqslant s$.

Extension of this work to the solution of equations with variable coefficients causes problems. As explained in Sections 1.5.7 and 1.5.8, stability cannot be determined solely by the von Neumann and pole conditions. Instead, some kind of dissipativity needs to be imposed if stability in ℓ_2 is to be considered. Alternatively, stability in a special norm can be investigated as in the Kreiss theory. Either way, the existing approximation-theoretic techniques are unlikely to be sufficient. Certainly the bound derived here, will be an upper bound, as the von Neumann condition and pole conditions are necessary, but as Iserles [Is84c] has proved, a variable coefficient multistep discretisation has accuracy halved. Thus it is most unlikely that the

saturation result derived here is an optimal bound for variable coefficient or nonlinear problems. However it undoubtedly reduces the choices and a more educated guess about which schemes should be used, can be made.

5. NUMERICAL SOLUTION OF THE SEMI-DISCRETISED SYSTEM OF EQUATIONS

5.1 Introduction

In the second half of this dissertation we consider numerical models for solving the system of ordinary differential equations arising from the semidiscretisation of a partial differential equation. It is essential that the numerical model should be chosen in a way which is appropriate for the underlying equation and the SD. Absolute stability is not sufficient to ensure stable integration of the ordinary differential system of equations. There are properties such as *conservation*, *dissipation* and *monotonicity* which may characterise the SD system. We will briefly discuss such considerations in Chapter 8. Initially our main concern is with the overall *stability* and *efficiency* of the model.

In this chapter we will discuss the various methods of time integration available. We consider the practicalities of implementation and thus motivate further investigation into a particular class of methods. Then we review stability theory for this class of methods and discuss their optimal versions. Further to this in Chapter 6 we concentrate on a smaller class of methods. We derive optimal versions of these schemes for integrating hyperbolic partial differential equations.

In Chapter 7 we present various algorithms for implementing our chosen scheme efficiently. We also discuss a possible method of error control.

Finally in Chapter 8 we implement some of our methods. We evaluate their usefulness by considering the evolution in time of some carefully chosen initial conditions which will exhibit the characteristics of the complete numerical model.

5.2 Methods for Time Integration

Basically there are three categories of finite-difference methods which we may use to integrate our SD system. Our main requirement is for a reasonably efficient and robust method which can usefully be generalised for multidimensional and nonlinear problems.

The first category to consider is the class of *explicit* ordinary differential equation solvers: Runge Kutta formulae and the familiar linear multistep schemes based on forward differencing. Implementation of explicit methods is relatively easy, and the storage requirements for multidimensional problems are not too severe. However, efficiency of explicit methods is necessarily reduced by their conditional stability properties. For nonlinear problems, where rapid change is occurring, integrating too fast would prevent accurate observation of transient effects. Therefore restricted steps in time would not be disastrous. However, when steady state is reached, reasonable speed without

amplification of stationary discontinuities is desirable.

Extending the above class of methods to allow schemes to be defined *implicitly*, by, for example, performing backward differencing, we may obtain unconditional stability properties. Thus faster time integration would be expected and so this class might be preferable for steady-state evolution. The main disadvantage is that we require to solve large algebraic systems of equations by iterative means. In multidimensional problems the structure of these systems becomes increasingly complicated and storage requirements are very severe. Also implicit methods have the undesirable tendency to smooth discontinuities.

Finally, we might consider some kind of *splitting* method designed to avoid the excessive storage requirements of the implicit methods. Examples are the alternating direction methods and hopscotch techniques [Mi80]. The idea is to reduce a problem with complicated structure to a series of simpler problems which can be solved more efficiently. By retaining implicitness, splitting methods can also possess unconditional stability and so might be the natural candidates for investigation. However, they are not so easy to implement and accurate resolution of discontinuities can again be a problem.

It would therefore appear that in regions of rapid variation, where we do not wish to integrate too fast, we should use explicit methods as they are by far the easiest to implement. However, when an equilibrium situation is

reached, efficiency would suggest adopting some kind of implicitness. In practical applications, the resolution of discontinuities is more important than efficiency and explicit methods usually perform better in this respect. Therefore we follow the approach of many other authors and opt for investigating ways of improving the efficiency of explicit methods (cf. [Ho77], [Ve76a], [Ve76b], [La66]). Hence it is quite natural to consider Runge Kutta methods since, as we shall see, the accuracy conditions do not define the schemes completely. There is freedom left over which can be used to improve efficiency.

5.3 Stability and Accuracy of Explicit Finite-Difference Methods

Before we examine any particular class of explicit methods more carefully, it is useful to elaborate on the concepts of *stability* and *accuracy* for ordinary differential equations. We consider the ordinary differential system of equations

$$\begin{aligned} \frac{dy}{dt} &= f(y) \quad t \geq 0 \\ y(0) &= y_0 \end{aligned} \qquad (5.3.1)$$

Application of a k step m stage method for solving (5.3.1) to the linear test equation

$$y' = \lambda y \quad , \quad \lambda \in \mathbb{C}$$

yields a numerical solution $\{y^n\}$ which satisfies a recurrence relation of the form

$$\sum_{i=0}^{k} \sum_{j=0}^{m} a_{ij} z^j y^{n+i} = 0 \quad , \quad n \geq 0 \qquad (5.3.2)$$

$$a_{ko} \neq 0 \quad .$$

Here $z = \lambda \Delta t$, where Δt is the steplength and y^n is an approximation to $y(t_n) = y(n \Delta t)$.

Recurrence relations of this type represent many multistep methods, in particular Runge Kutta multistep methods, linear multistep methods and predictor corrector methods. If the coefficients a_{kj} , $1 \leq j \leq k$ all vanish, then the method is explicit.

As for partial differential equations, stability and accuracy of a method may be completely defined by properties of an algebraic function. We define the *characteristic polynomial* of the method by

$$\Phi(\alpha,z) = \sum_{i=0}^{k} \sum_{j=0}^{m} a_{ij} z^j \alpha^i \qquad (5.3.3)$$

and its *stability region* S by

$$S := \left\{ z \in \mathbb{C} \mid \begin{array}{l} \text{Roots } \alpha_j \text{ , } 1 \leq j \leq m \text{ , of } \Phi(\alpha,z) \text{ satisfy } |\alpha_j| \leq 1 \\ \text{and if } |\alpha_j| = 1 \text{ then it is a simple root} \end{array} \right\} .$$

(5.3.4)

The numerical solution $\{y^n\}$ remains bounded as $n \to \infty$ for fixed Δt and all possible initial values

$\{y^0, y^1, \ldots, y^{k-1}\}$ if and only if $z \in S$. Comparing with Section 1.9 we say that the method is *absolutely stable* at $z \in \mathbb{C}$ if $z \in S$ and the method is *A-stable* if $z \in S$ for all $z \in \mathbb{C}^- = \{ z \in \mathbb{C} \mid \mathrm{Re}\, z \leq 0 \}$. Further the method is *zero stable* if it is absolutely stable at the origin.

If the method has *order* of accuracy p and is zero stable, then the algebraic function $R(z)$ given by

$$\Phi(R(z), z) = 0$$

has a branch $R_1(z)$ which is analytic in the neighbourhood of the origin and satisfies

$$R_1(z) - e^z = c\, z^{p+1} + O(z^{p+2}) . \qquad (5.3.5)$$

It may well happen that $R_1(z)$ approximates the exponential with order q where $q > p$. In particular, this may occur for the multistep methods we consider later. We note that the principal root $R_1(z)$ of a multistep method is the analogue of the stability function for the single-step methods, and that it reduces to a polynomial if the method is both single-step and explicit.

The numerical method is said to be *convergent* if the approximate solution tends uniformly to the real solution for all initial values as the steplength tends to zero. Zero stability and consistency, $p \geq 1$, are necessary and sufficient for convergence of linear multistep methods

[He62]. Thus, as with partial differential equations, we may investigate the convergence of a method in an approximation theoretical framework via the principal root of the characteristic equation.

For multistep multistage methods explicit determination of the principal root is unlikely. However, Jeltsch and Nevanlinna have developed a powerful new theory enabling comparison of methods without evaluating $R_1(z)$ explicitly [Je81a,82,83]. We will describe some of the results of this theory in a later section of this chapter.

5.4 Extended Stability Regions

We are now in a position to describe exactly what we mean by extended stability for partial differential equations. When we are solving an ordinary differential system of equations which is an SD of a partial differential equation, the parameter λ represents a value in the spectrum of the Jacobian matrix of the system. For stability we require that λ should be inside the stability region for all points of the spectrum. Now λ is inversely proportional to the mesh size of the SD and thus the size of the stability region necessarily restricts the size of the timestep and hence the speed of integration. Therefore for maximally efficient integration of a particular SD, we need a method which has a stability region enclosing as large a multiple as possible of the set defined by the spectra of

the Jacobian matrices as Δx tends to zero. Consequently, we can distinguish different problems according to the location of the spectral set. In turn, this distinguishes different methods as being more suitable for integrating systems arising from different kinds of partial differential equations.

If the Jacobian matrix is very nearly symmetric or skew symmetric, then the location of the eigenvalues of the finite matrix characterises the problem. For parabolic equations, the Jacobian is usually nearly symmetric and thus has eigenvalues lying inside a long narrow strip around the negative axis in $\mathbb{C}^-$. Thus we require stability regions which enclose this long narrow strip. Optimal regions may be determined by mapping the inside of the unit circle onto $\mathbb{C}^-$: sufficient conditions for the roots of the stability polynomial to lie inside the unit circle can then be obtained by applying the Routh-Hurwitz criterion to the transformed equation. For restricted classes of problems these conditions are linear and optimal methods may be found by solving a linear programming problem [Ve76b].

However if the partial differential equation is hyperbolic, very often its Jacobian is nearly skew symmetric. Then the eigenvalues lie inside a narrow strip enclosing the imaginary axis and methods with extended interval of stability along the imaginary axis are desirable.

The above situation only occurs if we discretise the hyperbolic equation using central differencing. For general

differencing, the matrix loses its symmetry properties and we must consider the spectrum of its infinite Toeplitz form rather than the eigenvalues of the finite matrix (cf. Section 1.8). These spectra describe Jordan curves in the complex plane and extended stability thus requires stability regions enclosing these curves. Now, as we will describe in Chapter 6, determination of these regions requires the solution of a non-linear programming problem. This is considerably more awkward to solve numerically than the linear programming problem which may be obtained for parabolic equations. However, one often desires to add dissipation to a numerical model to prevent unrealistic amplification of errors. Consequently, the additional effort required to solve these problems is worthwhile.

Examples of two spectral curves arising from general differencing of a linear hyperbolic equation are given in Figure 2. They correspond to the following semidiscretisations

$$\text{(a)}\quad \frac{d}{dt} v_j + \frac{1}{2}\frac{d}{dt} v_{j+1} = \frac{1}{\Delta x}\left[\frac{-1}{4} v_{j-1} - v_j + \frac{5}{4} v_{j+1}\right] \tag{5.4.1}$$

$$\text{(b)}\quad \frac{1}{12}\frac{d}{dt} v_{j-1} - \frac{2}{3}\frac{d}{dt} v_j - \frac{5}{12}\frac{d}{dt} v_{j+1} = \frac{1}{\Delta x}(v_j - v_{j+1}) \,.$$

For comparison we also give the eigenvalue curves of the finite matrices. Notice that these curves lie well inside the region bounded by the corresponding Jordan curve, and

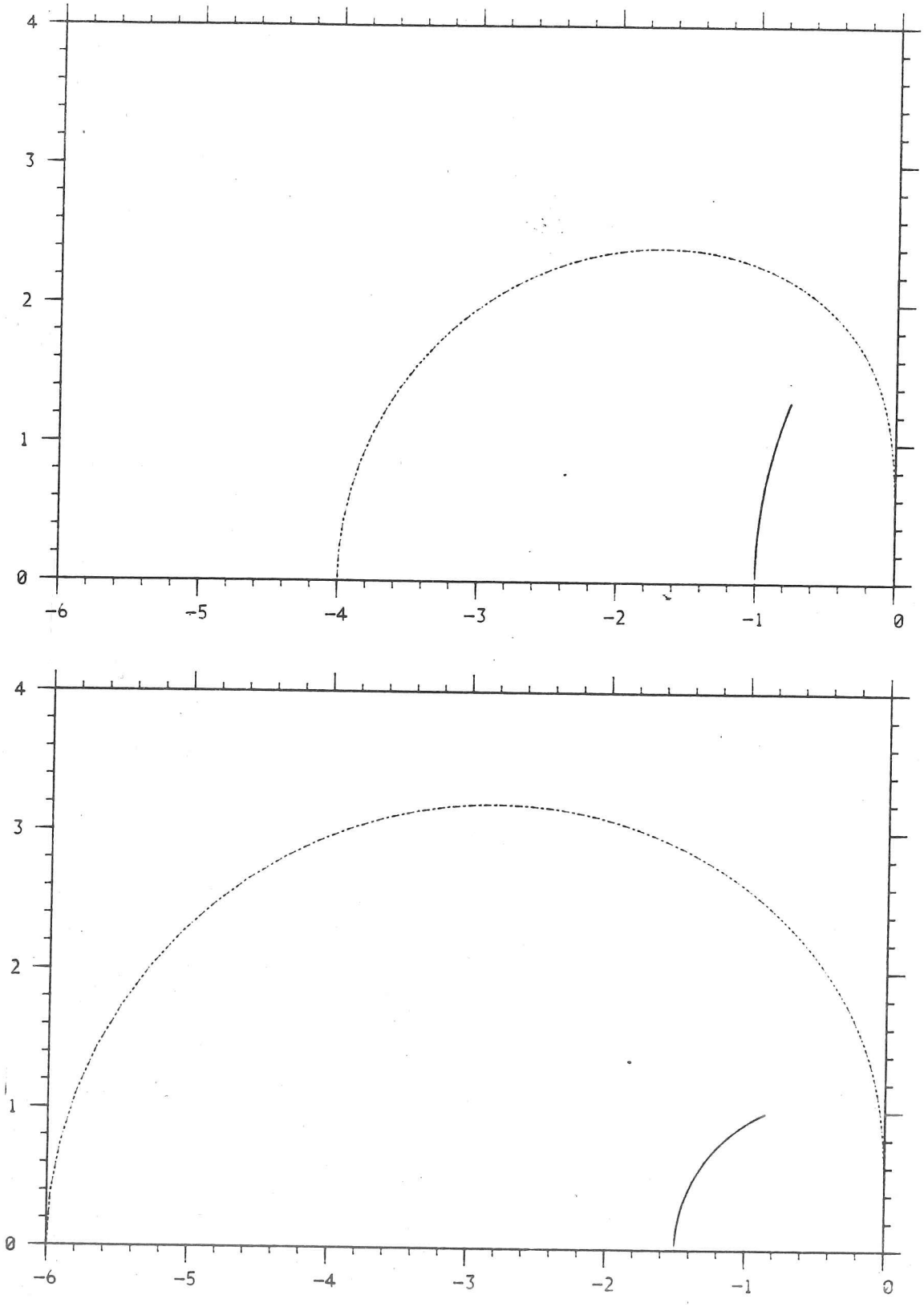


Figure 2. Examples of spectral curves and eigenvalue curves of the semi-discrete operators for cases A and B. The solid curve represents the locus of the eigenvalues and the dashed curve the locus of the spectrum of the infinite dimensional Toeplitz operator.

thus extended stability predictions, with respect to the eigenvalue curves, will be quite misleading.

We call the polynomials $\psi_i(z)$

$$\psi_i(z) := \sum_{j=0}^{m} a_{ij} z^j$$

occurring in (5.3.3) optimal (real or imaginary) stability polynomials if they give rise to a numerical method which has maximal stability interval along the negative real axis or imaginary axis respectively. In the next section we discuss optimal and nearly optimal real stability polynomials.

5.5 Extended Stability on the Negative Real Axis

Although our main concern in this dissertation is with the solution of hyperbolic partial differential equations, we present in this section a brief review of results for extended stability of parabolic problems. There has been much interest in the development of multistep multistage methods with extended negative real stability. Many of the problems encountered in the analysis also occur for stability regions extended in other ways. Thus this discussion serves to highlight some of the criteria that always need to be considered in designing such maximally efficient methods.

There is, however, a problem which arises in designing methods for parabolic and stiff equations which is of no

concern for hyperbolic equations. As we have already explained extended negative real stability is required. To obtain this, methods which use many stages are necessary. Therefore, it is possible, that there will be a considerable accumulation of round-off errors within each timestep. The extent to which these errors may destabilise the solution may be limited by analysing the *internal stability* of the method [Ho77]. The requirement that an internal stability polynomial should always be bounded by some value, usually the ratio of the maximal allowable truncation error to the machine precision, necessarily restricts the size of timestep further. Alternatively, for hyperbolioc problems the eigenvalues of the Jacobian are less widely separated and so methods with many stages are not needed. Consequently, internal instability is unlikely to dominate for reasonable timesteps.

Riha has proved that the optimal real stability polynomials of order $p = 1$ for one-step multistage methods are shifted Chebyshev polynomials [Ri72]. These polynomials have maximal interval of stability on the negative real axis $\beta_{real} = 2m^2$. The higher-order polynomials do exist [Ri72] and those with $p \leq 4$ have been constructed numerically [Ho77]. However, all these polynomials satisfy the equal ripple property, which means that they attain modulus one $m - p$ times within the stability interval. Thus despite having stability interval which increases quadratically with m, for small values of p, they cannot be practically used, particularly if eigenvalues of the Jacobian lie just off the

real axis. Also for $p > 1$ their coefficients must be derived numerically and this can lead to accumulation of rounding errors in implementations. For the second and third-order polynomials, Van der Houwen has derived analytic expressions for near optimal polynomials as special expansions in Chebyshev polynomials. The second-order polynomials have the reduced interval of stability $\beta_{real} \approx \frac{2}{3}(m^2 - 1)$ compared with $\beta_{real} \to .82m^2$ as $m \to \infty$ for the optimal schemes.

By consideration of two-step multistage methods, Verwer has derived near optimal polynomials of order $p = 1$ and $p = 2$ [Ve76a]. His second-order polynomials have $\beta_{real} \approx 1.8\ m^2$, which is an obvious improvement on the value above. However once again the coefficients must be derived numerically through an application of the equal ripple property. Thus, for large m, internal instability may lead to a deterioration in the solution.

Moving to a special class of three-step schemes without extra function evaluations, some of this harmful accumulation of errors can be reduced. For this case the polynomial coefficients are not calculated by application of the equal ripple property. Thus weak stability is no longer a problem. Instead they are calculated as a solution of a linear programming problem [Ve77]. Considerably improved stability boundaries $\beta_{real} \approx 5.15\ m^2$, and $2.29\ m^2$ for order one and two respectively are obtained. However, internal instability again becomes prevalent for large m values. It is

clear that algorithms need to be designed to reduce this error accumulation. Fortunately this course of action can be followed since the order conditions do not define the coefficients completely, and thus there is still some freedom available. Van der Houwen and Sommeijer have implemented some one-step schemes by identifying the polynomials of successive stages with shifted Chebyshev polynomials. They then employ a Richardson-type iterative method for their solution [Ho80]. In this way better internal stability properties are achieved without reducing β_{real}. Verwer extended this approach to improve the stability behaviour of his two and three-step schemes achieving $\beta_{real} \approx 5.17\ m^2$ and $\approx 2.32\ m^2$ for order one and two respectively [Ve79,82].

Sommeijer and Verwer have carried out a performance evaluation of some of these one, two and three-step Runge Kutta methods using a variable step implementation. By comparing standard algorithms with those based on Chebyshev recursions or on Jacobian linearisations, they showed how important the choice of algorithm is for successful integration. Despite having larger stability intervals, the three-step methods did not prove to to be more efficient unless linearisation of the Jacobian was performed [So80].

Higher-order accuracy up to $p = 6$ has been investigated by means of predictor corrector methods. Van der Houwen and Sommeijer considered a family of methods of this type which are constructed with restricted storage

requirements by using Chebyshev recursions [Ho83].

It should now be apparent that extended stability is not sufficient to guarantee greater efficiency. The manner of implementation is equally important. As already mentioned, we are unlikely to need so many stages in order to obtain the required speed of integration. Consequently internal stability should not play such an important role in the solution of hyperbolic problems but upper bounds on stability intervals must still be treated with caution.

5.6 Extended Stability on the Imaginary Axis

Determination of the maximal interval of stability β_{imag} is equivalent to maximising β_{imag} so that the roots of $\Phi(\alpha, iy)$ have modulus bounded by 1 in the interval $0 \leqslant iy \leqslant i\beta_{imag}$ and are simple if they attain this bound. Solving a minimax problem to determine β_{imag} means that any solution is strongly stable in the whole interval. Therefore for hyperbolic problems we do not have the problem of weak stability associated with the optimal polynomials for parabolic equations.

Determination of optimal imaginary stability polynomials was first considered by van der Houwen [Ho77]. He showed that construction of the first-order optimal imaginary polynomials can be posed as a minimax problem for a class of functions. Any polynomial which exists as a solution to this

problem is necessarily second-order accurate if it derives from a scheme with an odd number of stages, $m \geq 3$. In all cases the bound $\beta_{imag} \leq 2[m/2]$ holds. Thus measuring efficiency by scaling the bound by the number of function evaluations, m, there can be no increase in efficiency for increased m. Also, schemes with an even number of stages are obviously better candidates for increased efficiency.

The optimal imaginary polynomials of order p with m stages are denoted by $I_m^p(z)$. Then van der Houwen has shown that $I_m^p(z)$ for odd m and $p = 1$ is a sum of shifted Chebyshev polynomials:

$$I_{2k+1}^{(1)}(z) = I_{2k+1}^{(2)}(z) = T_k\left(1 + \frac{z^2}{2k^2}\right) + \frac{z}{k}\left(1 + \frac{z^2}{k^2}\right) U_{k-1}\left(1 + \frac{z^2}{2k^2}\right)$$

[Ho77]. These polynomials do attain the bound $\beta_{imag} = m - 1$. More recently it has been shown that $I_m^p(z)$ for m even are again sums of shifted Chebyshev polynomials obtainable from a recurrence relation for $I_m^p(z)$ ([Le84], [Pi83], [Ki84a]). The bound $\beta_{imag} = m - 1$ is obtained and thus, contrary to the expectation of van der Houwen's bound, schemes with an even number of stages cannot be expected to exhibit increased efficiency. Instead an odd number of stages may be preferred due to the bonus of gaining order of accuracy two. Moving to higher-order polynomials Kinnmark and Gray, [Ki84b] have found expressions for third order polynomials with $\beta_{imag} = \sqrt{((m-1)^2 - 1)}$. For large m this bound does approach $m - 1$ which suggests optimality although

this has only been proved for $m \leq 4$. Contrary to the optimal first order polynomials these polynomials are fourth-order accurate for an even number of stages. Furthermore, it has been observed that for increasing m the stability regions become increasingly slender in their far reaches and this with the linearity of β_{imag} as a function of m suggests that small m is preferable.

Moving to multistep rather than multistage methods Jeltsch has proved that any consistent linear multistep method which is stable on the imaginary axis is necessarily A-stable [Je78]. Thus it is implicit and has order at most two with the trapezoidal rule having the smallest error constant [Da63]. Further, Dekker has proved that the stability boundary for linear multistep methods of order greater than two is at most $\sqrt{3}$ [De81]. This bound is attained by the Milne-Simpson method which is of accuracy order four and implicit.

Therefore for increased efficiency, multistage rather than multistep should be preferred. If multistep multistage methods of low order cannot improve on the bound $\beta_{imag} = m - 1$, we should stay with the simplest explicit multistage methods possible or consider using an implicit scheme. For higher order, the bound $\sqrt{3}$ derived for linear multistep is easily broken by multistage methods: the Runge Kutta four-stage method of order four has $\beta_{imag} = 2\sqrt{2}$. Thus it is certainly worth considering whether by moving to multistep and multistage we might achieve yet greater

efficiency.

5.7 Comparison Theorems

So far we have not discussed the more complicated problem of determining optimal schemes for the solution of SD schemes which have spectra lying completely in $\mathbb{C}^-$. The most promising work in this direction is that of Jeltsch and Nevalinna [Je81,82,83]. They have sought ways of describing stability regions without evaluating principal roots so that comparisons between stability regions and hence methods can be made. It is well outside the scope of this dissertation to describe their work in entirety, but it is interesting to quote some of their results which are particularly relevant.

Their analysis relies on the fact that the stability and accuracy properties of a numerical model can be determined completely by an algebraic function which is the root of the characteristic equation. The principal branch of this function dominates the behaviour of the method. Therefore Jeltsch and Nevalinna have concentrated on describing a stability region qualitatively by this principal branch. As the efficiency of numerical methods can be measured in terms of the number of function evaluations, it is only useful to compare scaled stability regions. Their major result enabling comparison between explicit methods is that scaled boundaries of any two explicit methods necessarily intersect if the methods satisfy reasonable conditions which ensure

convergence.

To determine whether a method is optimal we need to find its describing algebraic function. The comparison theorems then give conditions which an optimal function must satisfy but do not guarantee existence. Thus the major problem is to find this optimal function.

Application of the theory to first-order explicit multistep multistage methods for maximal interval of stability β_{imag} yields the bound $\beta_{imag} \leq 2[\frac{m}{2}]$ of the one-step Runge Kutta. This suggests there is no gain in efficiency by incorporating more steps. The theorems are based on comparing the closure of stability regions and do not account for the appearance of branch points. They optimise stability sets that are closed intervals of the imaginary axis. The characteristic function Φ then has a factor $\Lambda(\alpha,z)$ which is of the form

$$\Lambda(\alpha,z) = \alpha^2 - 2\, i^m T_m(\frac{-iz}{m})\alpha + (-1)^m \quad . \qquad (5.7.1)$$

where $T_m(z)$ is an m^{th} degree Chebyshev polynomial. Obviously any method which has even number of stages and a factor Λ as in (5.7.1) cannot be zero-stable. For m odd, the largest stability interval is $\beta_{imag} = m \sin(\frac{\pi}{2m})$ and thus the midpoint rule which has $\beta_{imag} = 1$, is best in the scaled sense. However, by suitable construction, we can show that this bound can be broken. In particular, the two-step, three-stage method of order three which has

characteristic function

$$\Phi(\alpha,z) = \alpha^2 - (2z + \tfrac{1}{3}z^3)\alpha - 1 \qquad (5.7.2)$$

has $\beta_{imag} \approx 2.8473$. The bound has been broken by perturbing the coefficients of the characteristic function given by (5.7.1) so that the branch point no longer occurs on the pure imaginary axis.

Now the above result is not related to the accuracy of the method in any way. However, it can also be proved that there do exist linear k-step methods of order $p = k$ with an interval of stability β_{imag} where $\beta_{imag} \in [0,1]$ and $k \in \{2,3,4\}$. For $k = 1 \bmod 4$ no explicit linear k step method of order $p = k$ exists such that $\beta_{imag} > 0$ (Th5.1 [Je81]).

Extension of the theory to implicit methods can incorporate their order of accuracy since comparison is in part performed by the location of the poles of the principal root which in turn is related to accuracy [Th 2.2, Je82]. That there is a very close relationship between the location of the poles and the order of the optimal method is proved with the aid of order stars defined on Riemann surfaces. Comparison yields the bound $\beta_{imag} \leq \sqrt{3}$ for linear multistep methods of order greater than 2 already proved in a different way by Dekker [De81]. The equivalent result for two-stage methods is $\beta_{imag} \leq \sqrt{15}$ for order $p > 4$. Therefore this new theory enables proof of all existing results for stability on the imaginary axis but, more

importantly, there are results which may be useful for spectra in $\mathbb{C}^-$.

In the case that the spectral curve is a slight perturbation from a circle, some of the circle theorems are relevant. In particular, for any $r < 1$ and k there exist linear k step methods of order $p = k$ such that the disc of radius r centred at $(-r,0)$ is contained in S . An m-stage method has a disc of radius m centred at $(-m,0)$ contained in S only if the algebraic function which is the root of the characteristic equation is

$$\alpha(z) = (1 + \frac{z}{m})^m \qquad (5.7.3)$$

([Je81]). Clearly any function of this form can represent a consistent, $p = 1$, Runge Kutta method with m stages. Thus efficiency is not encouraging in the single-step multistage case.

However, the two spectral curves we are considering indicate that such curves may not be just small perturbations from a circle (cf Figure 2). Our curves are squashed more towards the imaginary axis and thus we need stability regions with boundaries that stay near this axis. Unfortunately, an explicit linear multistep method with stability region which extends further into the left half plane has a root locus curve which approaches the origin less steeply along the imaginary axis [Th 2.19, Je82]. Thus this is not encouraging for the two particular curves being considered.

So far we have not found any more theoretical results which enable determination of optimal methods in $\mathbb{C}^-$. However it may be possible to apply theory used by Manteuffel in deriving iteration methods for solving non-symmetric linear systems [Ma77]. He develops an algorithm for finding optimal iteration parameters as a function of the convex hull of the spectrum. The algorithm relies on some very useful results from complex function theory. If the spectrum of the infinite Toeplitz form lies inside a region bounded by an ellipse not containing the origin in its interior, then the unique polynomial of degree m attaining its bounds on the boundary of the region is a translated Chebyshev polynomial. The algorithm presented finds the optimal polynomial by solving a minimax problem which determines the optimal ellipse or circle bounding the spectrum. We may be able to use this algorithm to develop optimal one-step multistage schemes. Certainly it is worthwhile to investigate this work further.

For the present, we concentrate on examining the stability properties of a class of multistage two-step formulae of Runge Kutta type. We have seen what interval of stability we can achieve on the imaginary axis by allowing only one-step, and that by moving to two steps we can attain a reasonable bound for the method described by (5.7.2). Also the circle theorems have shown that by considering either multistep or multistage alone, efficiency is restricted. Thus, by allowing two steps but many stages, we may hope to attain improved efficiency. However as we have no theory to rely

on, we only obtain numerical approximations to optimal solutions.

6. A CLASS OF TWO-STEP MULTISTAGE METHODS

6.1 Introduction

In this chapter we intend to discuss some multistep multistage methods belonging to the class of so-called hybrid methods. Generally a hybrid method shares the property of Runge Kutta methods of utilising data at non-step points. We consider particular two-step members of this class which bear some similarity to those already discussed by Verwer [Ve76a,76b] and Watt [Wa67].

Our main objective is to develop efficient methods for integrating hyperbolic systems of equations. Here we concentrate on designing methods of second-order and third-order accuracy with extended regions of stability. In Chapter 7 we will consider ways of reducing storage in implementation. A method of error control is also proposed similar to a Runge-Kutta-Fehlberg scheme for one-step methods.

Equivalent one-step Runge Kutta methods are also designed. In a later chapter we compare the performance of stabilised one- and two-step Runge Kutta along with two commonly used methods of time integration. The comparison is by investigating the propagation of monochromatic and polychromatic signals as well as the propagation of discontinuities under integration by these schemes.

6.2 Two-Step Multistage Runge Kutta Formulae

We define a two-step m-stage Runge Kutta scheme for solving the system of equations $y' = f(y)$ by

$$y^{n+1} = (1 - \beta)y^n + \beta\, y^{n-1} + h \sum_{i=1}^{m} v_i k_i + h \sum_{i=1}^{m} w_i \ell_i$$

where

$$k_1 = f(y^{n-1}) \,, \quad k_i = f(y^{n-1} + h \sum_{j=1}^{i-1} \alpha_{ij} k_j)$$

(6.2.1)

$$\ell_1 = f(y^n) \,, \quad \ell_i = f(y^n + h \sum_{j=1}^{i-1} \alpha_{ij} \ell_j) \quad .$$

The vector y^n represents a numerical approximation to the analytical solution $y(t)$ at $t = t_n$ where the points t_{n+1} , t_n , t_{n-1} are the reference points of the formula, and h is the steplength, $t_{n+1} = t_n + h$, $t_{n-1} = t_n - h$.

Notice immediately that this particular class of methods is designed so that function evaluations at time t_{n-1} are the same as those taken at time t_n in the previous step. Therefore we gain the extra degrees of freedom associated with a two-step scheme *without* the need for *extra function evaluations*. Consequently we can expect to be able to design schemes which allow faster integration. However, we

can only derive benefit from these schemes if we store the necessary values from step to step. Thus we have the increased efficiency of a two-step scheme at the cost of extra storage requirements: this does not present too much of a problem. The choice of a two-step scheme which is of Runge-Kutta type means that there is still some freedom in the coefficients of the scheme after the requirements for stability have been satisfied. Hence we may use the remaining freedom to design algorithms which allow implementation of this class of method with minimal storage. We will investigate some suitable algorithms in Chapter 7. Here we are concerned with deriving efficient schemes for integrating hyperbolic equations.

First it is helpful to restate the definitions for multistep multistage schemes previously discussed in Section 5.3. In the following sections we will discuss the order of accuracy of these methods and their absolute stability properties.

The method defined by (6.2.1) is *convergent* only if for every solution, $y(t)$, of the initial-value problem $y' = f(y)$, $y(0) = y_0$, defined on the interval $t \in [0,T]$ where f is sufficiently smooth

$$\lim_{h \to 0} y^n = y(t_n) .$$

For convenience we associate the multistage method (6.2.1) with a nonlinear difference operator

$$y^{n+1} - Z(y^n , y^{n-1}) .$$

Then the method is said to be *accurate* of *order* p , at $t = t_n$, if p is the largest integer such that

$$\lim_{h \to 0} (y(t_{n+1}) - Z(y(t_n) , y(t_{n-1}))) = O(h^{p+1}) . \qquad (6.2.2)$$

If $p \geq 1$, the method is said to be *consistent*.

Let us define the polynomial $\rho_m(\alpha)$ by

$$\rho_m(\alpha) = \alpha^2 - (1 - \beta)\alpha - \beta . \qquad (6.2.3)$$

We say that the method is *zero stable* if no root of this polynomial has modulus greater than one and if a root has modulus one then it is simple.

Then the method is convergent if and only if it is zero stable and consistent [Wa67].

Immediately we see that the method is zero stable only if $-1 < \beta \leq 1$ since $\rho_m(\alpha)$ has roots $\alpha_1 = 1$ and $\alpha_2 = -\beta$. Using Taylor's theorem and expanding $y(t_{n+1})$ and $y(t_{n-1})$ about $y(t_n)$ in (6.2.2) yields

$$\lim_{h \to 0} \left[y(t_{n+1}) - Z[y(t_n), y(t_{n-1})] \right] =$$

$$h[(1 + \beta) y'(t_n) - \sum_{i=1}^{m} (v_i + w_i) f(y_n)] + O(h^2) .$$

(6.2.4)

Thus the method is consistent if and only if

$$1 + \beta = \sum_{i=1}^{m} (v_i + w_i) \qquad (6.2.5)$$

and convergent only if $\sum_{i=1}^{m} (v_i + w_i) \neq 0$ since otherwise the difference scheme might approximate a wrong differential equation. Therefore the method is convergent if and only if $-1 < \beta \leq 1$ and $1 + \beta$ is equal to the sum of weights. If β is very near -1, the convergence condition is nearly violated and therefore this situation should be avoided for accurate results.

As for linear multistep methods, it is convenient to estimate accuracy of the method by means of a normalised error constant rather than through the truncation error alone. Therefore the truncation error is normalised by the factor $\sum_{i=1}^{m} (v_i + w_i)$ which tends to zero if convergence is nearly violated. In this way we see again that we require the coefficient β not to move too close to -1 or else the normalised truncation error will be too large.

6.3 Order Conditions

Applying Taylor's theorem for several variables, we can expand the difference (6.2.2) further than in (6.2.4) to obtain order conditions. As our concern is with schemes

having third order with error control by a fourth-order method, we perform an expansion up to terms which are fourth order in h. Usually one employs the theory of elementary differentials, as introduced by Butcher [Bu62], to find the terms in this expansion. Alternatively one may use the tensor notation as described by Henrici [He62]. Here we adopt the latter approach where derivatives of f are abbreviated as follows:

$$\frac{df^i}{dy_j} = f^i_j \quad , \quad \frac{\partial^2 f^i}{dy^j dy^k} = f^i_{jk} \, ,$$

and where $y^i(t_n), f^i$ are the i^{th} components of $y(t_n)$ and f respectively. Then this expression is

$$y^i(t_{n+1}) - Z[(y(t_n), y(t_{n-1})]^i = C_1 h f^i + C_{21} h^2 f^i_j f^j +$$

$$C_{31} h^3 f^i_{jk} f^j f^k + C_{32} h^3 f^i_j f^j_k f^k + \qquad (6.3.1)$$

$$C_{41} h^4 f^i_{jk\ell} f^j f^k f^\ell + C_{42} h^4 f^i_{jk} f^j f^k_\ell f^\ell +$$

$$C_{43} h^4 f^i_j f^j_{k\ell} f^k f^\ell + C_{44} h^4 f^i_j f^j_k f^k_\ell f^\ell + o(h^5) \, .$$

It can be proved inductively that the coefficients C_{ij}, which are constants determined by the parameters of the method, are as given in Table 6.3.1. The condition $C_1 = 0$ is the consistency condition (6.2.5), whilst from the definition of order of accuracy it is apparent that the scheme has order 2 if in addition $C_{21} = 0$, order 3 if $C_{32} = C_{31} = 0$ as well and order 4 if the coefficients C_{41}, C_{42}, C_{43} and C_{44} are also zero.

$$C_1 = 1+\beta - \sum_{i=1}^{m} (v_i + w_i)$$

$$C_{21} = \frac{1-\beta}{2} - \sum_{i=2}^{m} \sum_{j=1}^{i-1} \alpha_{ij} (v_i + w_i) + \sum_{i=1}^{m} v_i$$

$$C_{31} = \frac{1+\beta}{6} - \frac{1}{2} \left[\sum_{i=1}^{m} v_i \left(\sum_{j=1}^{i-1} \alpha_{ij} - 1 \right)^2 + \sum_{i=2}^{m} w_i \left(\sum_{j=1}^{i-1} \alpha_{ij} \right)^2 \right]$$

$$C_{32} = \frac{1+\beta}{6} - \sum_{i=3}^{m} \sum_{j=2}^{i-1} \sum_{k=1}^{j-1} (v_i + w_i) \alpha_{ij} \alpha_{jk} + \sum_{i=2}^{m} \sum_{j=1}^{i-1} v_i \alpha_{ij} - \frac{1}{2} \sum_{i=1}^{m} v_i$$

$$C_{41} = \frac{1-\beta}{24} - \frac{1}{6} \sum_{i=1}^{m} v_i \left(\sum_{j=1}^{i-1} \alpha_{ij} - 1 \right)^3 - \frac{1}{6} \sum_{i=2}^{m} w_i \left(\sum_{j=1}^{i-1} \alpha_{ij} \right)^3$$

$$C_{42} = \frac{1-\beta}{8} - \sum_{i-1}^{m} v_i \left(\sum_{j=1}^{i-1} \alpha_{ij} - 1 \right) \left(\sum_{j=2}^{i-1} \sum_{k=1}^{j-1} \alpha_{ij} \alpha_{jk} - \sum_{j=1}^{i-1} \alpha_{ij} + \frac{1}{2} \right)$$

$$- \sum_{i=3}^{m} w_i \left(\sum_{j=1}^{i-1} \alpha_{ij} \right) \left(\sum_{j=2}^{i-1} \sum_{k=1}^{j-1} \alpha_{ij} \alpha_{jk} \right)$$

$$C_{43} = \frac{1-\beta}{24} - \frac{1}{2} \sum_{i=2}^{m} \sum_{j=1}^{i-1} \alpha_{ij} \left(w_i \left(\sum_{k=1}^{j-1} \alpha_{jk} \right)^2 + v_i \left(\sum_{k=1}^{j-1} \alpha_{jk} - 1 \right) \right)^2 + \frac{1}{6} \sum_{i=1}^{m} v_i$$

$$C_{44} = \frac{1-\beta}{24} - \sum_{i=4}^{m} \sum_{j=3}^{i-1} \sum_{k=2}^{j-1} \sum_{\ell=1}^{k-1} \alpha_{ij} \alpha_{jk} \alpha_{k\ell} (v_i + w_i) + \sum_{i=3}^{m} \sum_{j=2}^{i-1} \sum_{k=1}^{j-1} \alpha_{ij} \alpha_{jk} v_i$$

$$- \frac{1}{2} \sum_{i=2}^{m} \sum_{j=1}^{i-1} \alpha_{ij} v_i + \frac{1}{6} \sum_{i=1}^{m} v_i \quad .$$

Table 6.3.1 Order conditions for a two-step m stage scheme.

It is immediately evident that the order conditions comprise a set of nonlinear equations in the coefficients of the numerical scheme. Theoretically, the maximal order of a scheme can be determined by solving these equations. However, even for a relatively small number of stages, nonlinearity makes solution of the equations difficult. As for one-step Runge Kutta formulae, it is likely that an m stage formula cannot necessarily be made m^{th}-order accurate. As the number of stages increases, the number of conditions required to be satisfied increases faster than the number of degrees of freedom and a large degree of dependence is necessary if m^{th} order is to be obtained.

Comparing with the maximal order of the usual one-step schemes, there is improved accuracy for schemes with up to three stages. However, even for just four stages the equations are increasingly complicated to solve and we merely conjecture that it seems very likely that fifth order is attainable. Whether sixth order is attainably by a five-stage scheme is doubtful. A comparison between the maximal order of one and two-step schemes and the number of degrees of freedom is given in Table 6.3.2. We see immediately that the number of parameters in a two-step scheme is the same as for a one-step scheme with an extra function evaluation. It is therefore fairly reasonable to expect at least an increase by one in maximum order attainable.

No. of Stages	No. of Steps	Maximal order p	No. of Conditions	No. of Parameters
1	1	1	1	1
1	2	2	2	3
2	1	2	2	3
2	2	3	4	6
3	1	3	4	6
3	2	4	8	10
4	1	4	8	10
4	2	(5)	17	15
5	1	4	8	15
5	2	(6)	37	21

Table 6.3.2 Maximal order of one and two-step schemes where the bracketed expressions are conjectures for maximal order.

The extra order attainable by the schemes with few stages, $m \leq 3$, justifies the investigation of this particular class of methods. Without any extra function evaluations but possibly a little extra computational complexity, an extra degree of order is achievable.

$$v_1 = \frac{\beta - 1}{2} + \frac{5 - \beta}{12\alpha} + \frac{1}{6\delta\alpha^2}[\alpha - (\gamma + \delta)]$$

$$w_1 = \frac{3 + \beta}{2} + \frac{\beta - 5}{12\alpha} + \frac{1}{6\delta\alpha^2}[\gamma + \delta - \alpha]$$

$$v_2 = \frac{\beta - 5}{12\alpha} + \frac{\gamma + \delta}{6\delta\alpha^2}$$

$$w_2 = \frac{5 - \beta}{12\alpha} - \frac{\gamma + \delta}{6\delta\alpha^2}$$

$$v_3 = \frac{-1}{6\delta\alpha} \qquad w_3 = \frac{1}{6\delta\alpha}$$

$$\beta = 5 + \frac{2}{\delta\alpha^2}[(\gamma + \delta)^2 - \alpha(\gamma + 3\delta)]$$

Table 6.3.3 Coefficients of three-stage, fourth-order schemes (Here, for clarity, we have made the following changes in notation; $\alpha_{21} = \alpha$, $\alpha_{31} = \gamma$, $\alpha_{32} = \delta$.)

In Table 6.3.3 we give the coefficients of a fourth-order, three-stage scheme as an example of the solution of the equations in Table 6.3.1. Notice that even with fourth order there is still some freedom in the computation available. However as we shall see in the next section, none of these parameters can be used to increase stability. For fourth order, the remaining degrees of freedom can only be used to determine the $\{\alpha_{ij}\}$ which, in turn, determine the intermediary points of the calculation. Nonetheless, as we describe in Chapter 7, this flexibility proves to be very useful in the design of error control schemes which use a minimal number of extra function evaluations.

Even though the scheme given in Table 6.3.3 does attain fourth order, we suggest caution in its use for a global integration. Since $w_3 = -v_3$ and $w_2 = -v_2$ we do not have positivity of coefficients and so we increase the likelihood of cancellations in rounding errors occurring during calculation. However this is not a deterrent in its implementation as a means of error control of a lower-order scheme where we only need to integrate locally.

6.4 Absolute Stability

The absolute stability of a method is investigated, as in Section 5.3, by applying it to the linear test model

$$y' = \lambda y \quad , \quad \lambda \in \mathbb{C} \quad . \tag{6.4.1}$$

Putting $z = h\lambda$ we obtain the recursion relation (5.3.2) for a two-step scheme

$$y_{n+1} = S(z)y_n + P(z)y_{n-1} \tag{6.4.2}$$

(cf. [Ve76a]). Here the polynomials $S(z)$ and $P(z)$ are polynomials of degree m whose coefficients can be shown to be defined by the coefficients of the scheme as follows:

$$S(z) = \sum_{i=o}^{m} s_i z^i \quad , \quad P(z) = \sum_{i=o}^{m} p_i z^i$$

$$s_o = 1 - \beta \quad , \quad s_1 = \sum_{i=1}^{m} w_i$$

$$s_i = \sum_{k_1=i}^{m} \sum_{k_2=i-1}^{k_1-1} \cdots \sum_{k_i=1}^{k_{i-1}-1} \alpha_{k_1 k_2} \cdots \alpha_{k_{i-1} k_i} w_{k_i} \quad 2 \leqslant i \leqslant m$$

$$p_o = \beta \quad , \quad p_1 = \sum_{i=1}^{m} v_i \qquad (6.4.3)$$

$$p_i = \sum_{k_1=i}^{m} \sum_{k_2=i-1}^{k_1-1} \cdots \sum_{k_i=1}^{k_{i-1}-1} \alpha_{k_1 k_2} \cdots \alpha_{k_{i-1} k_i} v_{k_i} \quad 2 \leqslant i \leqslant m \ .$$

Notice that the nature of the method chosen, whereby the intermediary points of the calculation are the same at subsequent steps, means that the coefficients $\{s_i \ , \ p_i \mid 1 \leqslant i \leqslant m\}$ are similar, differing only by the weights of the method $\{v_i \ , \ w_i\}$.

As explained in Section 5.3, the recurrence relation has the characteristic equation

$$\alpha^2 - S(z)\alpha - P(z) = 0 \quad ; \quad z = h\lambda \qquad (6.4.4)$$

the roots of which determine the properties of the method. Consistency of order p requires that one root of (6.4.4), the principal root, is an approximation to the exponential function, $\exp(z)$, of order p .

It is convenient to express the consistency conditions of the method in terms of the polynomial coefficients. Since the principal root does approximate exp(z) to order p we may do this by substituting the Taylor expansion for exp(z) up to terms of order p in Equation 6.4.4. Then for accuracy up to order $p = 3$, we have the following conditions

$$
\begin{aligned}
& p_o + s_o = 1; \\
& s_1 + p_1 + s_o = 2 \qquad \text{order } 1; \\
& s_2 + s_1 + \frac{s_o}{2} + p_2 = 2 \qquad \text{order } 2; \\
& s_3 + s_2 + \frac{s_1}{2} + \frac{s_o}{6} + p_3 = \frac{4}{3} \qquad \text{order } 3;
\end{aligned}
\tag{6.4.5}
$$

The first condition of (6.4.5) is satisfied automatically by the choice already made, $s_o = 1 - \beta$, $p_o = \beta$. If order $p = 2$ is required, then the conditions here are sufficient; however if order $p = 3$ is required, then we see by Table 6.3.1 that there is an extra condition. The conditions here only correspond to $C_1 = 0$, $C_{21} = 0$ and $C_{32} = 0$. In addition we require $C_{31} = 0$ which we cannot obtain from the characteristic equation. However, it is useful to obtain an expression for C_{31} in terms of the coefficients $\{ s_i, p_i \}$. Obviously we may do this by finding expressions for $\{ v_i, w_i \}$ from Equations 6.4.3 and substituting them into that for C_{31}. For a two or three-stage scheme this gives

$$(p_2 + s_2)\alpha + \left(\frac{p_3 + s_3}{\delta\alpha}\right)\left((\gamma + \delta)^2 - \alpha(\gamma + 3\delta)\right) = 0 \quad .$$

(6.4.6)

Then the solution of (6.4.6) with (6.4.5) for a two-stage scheme completely defines the coefficients $\{s_i , p_i\}$ in terms of one free variable as

$$s_o = 1 - p_o \ , \ s_1 = \frac{3 + p_o}{2} \ , \ s_2 = \frac{5 - p_o}{12} \ ,$$

$$p_1 = \frac{p_o - 1}{2} \ , \ p_2 = -s_2 \quad .$$

(6.4.7)

However for a three-stage scheme, equations (6.4.5) are sufficient to determine the number of degrees of freedom available for stability. Equation (6.4.6) merely provides a restriction on the $\{\alpha_{ij}\}$ once optimal coefficients $\{s_i , p_i\}$ have been calculated. Notice that once stability is determined, the scheme is not completely predicted: some of the $\{\alpha_{ij}\}$ are still free. We stress again that this is the property of Runge Kutta methods which makes them so advantageous for an analysis of this sort. The extra available degrees of freedom may be used either to reduce storage requirements or for designing a particularly efficient method of implementation.

We now turn to the major problem which will concern us in the next section, the maximal efficiency of the method.

As in Section 5.2, we define the stability region of the method by

$$S = \left\{ z \in \mathbb{C} \;\middle|\; \begin{array}{l} \text{Both roots } \alpha_i(z) \text{ of (6.4.4) satisfy} \\ |\alpha_i| \leq 1 \text{ and if } |\alpha_i| = 1 \text{ then it is simple} \end{array} \right\} . \tag{6.4.8}$$

We say the method is absolutely stable at a point z if $z \in S$. Recall that, stability of the numerical model of the partial differential equation requires that the infinite spectrum of the Jacobian multiplied by the steplength $h = \Delta t$ should lie inside S . For an efficient scheme we therefore require as large a multiple as possible of this spectrum to lie inside S . Since the values of the spectrum are proportional to $\frac{1}{\Delta x}$, where Δx is the largest grid size in the spatial discretisation, this is equivalent to finding the maximal Courant number $\mu = \frac{\Delta t}{\Delta x}$ for which the partial differential equation can be stably integrated. Consequently we will now discuss how we may calculate the coefficients of efficient schemes and hence find maximal Courant numbers.

6.5 The Stability Problem

As previously mentioned, we wish to construct stabilised schemes allowing maximal Courant number, μ , subject to the order conditions (6.4.5). For hyperbolic systems being solved by conservative SD's the general problem is to find the maximal interval of stability on the imaginary axis.

However, for systems with spectra which lie in $\mathbb{C}^-$, no general solution is sufficient for all problems. Either we must find a solution for each semidiscretisation employed or solve a general problem to give a set of approximate solutions. As explained in Section 5.7, it is only application of the maximum modulus principle which makes it feasible to find any approximate solution at all.

We rely on finding a region in $\mathbb{C}^-$ within which the roots of the characteristic equation have modulus less than one and attain modulus one on its boundary. By the maximum modulus principle we only need to determine the roots of the characteristic equation along the boundary of the region. It is not sensible to find the roots of the characteristic equation at several points on $\mathbb{C}^-$ and then evaluate their moduli. Instead we need criteria by which we can determine whether roots at a given point will satisfy the stability property.

Recall that for points lying on the real axis, the Routh Hurwitz criterion provides sufficient conditions to determine stability and thus maximising μ is a linear programming problem [Ve76b]. However, we are interested in the more complicated problem where the characteristic equation cannot be transformed to an equation with real coefficients. In this case sufficient conditions for stability at a given point are provided by the Cohn-Schur criterion. For the quadratic equation (6.4.4) this criterion takes the form of two nonlinear inequalities,

$$\text{i)} \quad |P(z)| \leq 1 \text{ ; and}$$
$$\text{ii)} \quad \big|\, |P(z)|^2 - 1 \,\big| \geq |\bar{S}(z)P(z) + S(z)|. \qquad (6.5.1)$$

If strict inequality holds the roots of the quadratic equation lie inside the unit circle [Mi71]. Therefore evaluation of maximal μ is a nonlinear programming problem subject to the linear constraints (6.4.5) along with the condition for zero stability $-1 < \beta \leq 1$.

We will describe a solution of this nonlinear programming problem in the next section. First a few cases for which solutions are known exactly are discussed.

Theorem 6.5.1

The maximal interval of stability attainable on the imaginary axis by a two-step third-order scheme is given by $\beta_{imag} = 1$. The schemes which achieve this value for β_{imag} have characteristic function

$$\Phi(\alpha,z) = \alpha^2 - \left(\frac{4}{5} + \frac{8z}{5} + \frac{2}{5}z^2\right)\alpha - \left(\frac{1}{5} - \frac{2}{5}z - \frac{2}{5}z^2\right) \quad . \qquad (6.5.2)$$

Proof

We saw in Section 6.4 that the coefficients of the polynomials $S(z)$ and $P(z)$ for a two-step two-stage scheme can be obtained from relations (6.4.7). Substitution of these coefficients into conditions (6.5.1) and putting $z = iy$ yields

i) $\frac{1}{144}(5 - p_o)^2 y^4 + \frac{1}{12}(1 + p_o)(3 + p_o)y^2 - (1 - p_o^2) \leq 0$

ii) $y^2 \leq 24 \frac{(1 - p_o^2)}{(5 - p_o)^2}$.

Here condition ii) implies condition i) for $|p_o| < 1$. The maximal value of y satisfying ii) is $y = 1$ and occurs for $p_o = 1/5$. Thus $\beta_{imag} = 1$ and the characteristic equation is of the given form after substituting $p_o = 1/5$ in equations (6.4.7). □

We now refer to the result of Jeltsch and Nevanlinna which was discussed in Section 5.7: for an m-stage scheme either $I_m \subset S$ or $\overline{I}_m = \overline{S}$ and the characteristic polynomial has the factor (5.6.1) where $I_r = \{ iy \mid |y| \leq r \}$. By (5.7.1) the two-step three-stage schemes with $\overline{I}_3 = \overline{S}$ are defined by the characteristic equation

$$\alpha^2 - (2z + \frac{8}{27} z^3)\alpha - 1 = 0 \qquad (6.5.3)$$

and thus have order-of-accuracy two. Nevertheless, the largest value of r such that $I_r \subset S$ is $r = 1.5$. This is because the comparison theorems do not cater for the occurrence of branch points and (6.5.3) has a branch point at $z = 1.5i$. Fortunately, as mentioned in Section 5.7, we can break the bound on r if we remove the condition of closure. The third-order schemes defined by (5.7.2) have $r \approx 2.8473$. These schemes have the advantage that their coefficients are known exactly and their stability bound is

nearly optimal . Implementation of these schemes requires that we solve for the parameters $\{ v_i, w_i \}$ from the polynomial coefficients. It is preferable that coefficients are known exactly so that there is no loss in accuracy when solving for the parameters since the stability regions are sensitive to perturbations in the polynomial coefficients particularly if the zero-stability parameter β_{imag} becomes close to -1. Therefore we have not attempted to find any better approximation numerically.

For spectra lying in $\mathbb{C}^-$, we have the results of the circle theorems which may give approximately optimal solutions for spectral curves being small perturbations from circles. By (5.7.3) the one-step Euler method described by $\alpha(z) = 1 + z$ is the most efficient one-step m-stage method of first order for the circle of radius 1 centred at (-1,0) lying inside S . Thus for higher-order methods, no discs of radius m can lie inside S , but the largest such disc is not known. Also for two-step m-stage methods, $m > 1$, the largest disc lying inside S is not determined.

Consequently the existing theory in the literature is not sufficient to determine the maximally efficient schemes for integrating hyperbolic equations. In general we must resort to numerical techniques for solving the nonlinear problem described by the Cohn-Schur criterion. A technique for finding optimal, or near optimal, schemes for arbitrary domains lying within $\mathbb{C}^-$ is described in the next section.

6.6 A Solution Technique

Here we give a method for solving the non-linear equations (6.5.1). An optimal scheme with stability region completely enclosing a domain Ω is sought. Heuristically the idea is to perform a discretisation along the boundary of Ω, i.e. define points z_j, $1 \leq j \leq N$ which become dense along the boundary $\partial\Omega$, as $N \to \infty$. Thus we have $2N$ non-linear constraints on the variables $\{p_i , s_i \quad 0 \leq i \leq m\}$ by requiring that the Cohn-Schur criterion is satisfied at each point z_j. Solving the consistency conditions (6.4.6) and (6.4.5) enables these constraints to be expressed in terms of M independent variables where M varies with the number of stages and order imposed. As the points z_j depend on the Courant number μ via the equation defining the boundary of Ω, we have $M + 1$ independent variables. Then adding the condition for zero stability, we may state the optimisation problem as follows:

$$\text{maximise } \mu$$

subject to

$$\begin{aligned} &-1 < p_o \leq 1 \\ &|P(z_j)| \leq 1 \quad 1 \leq j \leq N \\ &|\bar{S}(z_j)P(z_j) + S(z_j)| \leq |1 - |P(z_j)|^2| \quad 1 \leq j \leq N \end{aligned} \tag{6.6.1}$$

where $P(z_j)$ and $S(z_j)$ are functions of $X = (\mu, X_1, \ldots, X_M)^T$

and $\{X_i \mid 1 \leq i \leq M\}$ is a subset of $\{p_i, s_i \mid 0 \leq i \leq m\}$.

A solution to this problem was found by using the NAG library routine E04VBF. This routine attempts to find a minimum of a function of several variables subject to general constraints as described by (6.6.1). It uses a sequential augmented Lagrangian method and solves the minimisation subproblems by a modified Newton method.

The approach we employed was to solve the problem for a very small value of N , say $N = 5$, and then successively increase the value of N using each preceding solution as a starting point for the next problem. Generally we found that the solution failed to converge for $N > 20$, although in the case of $M = 1$ convergence for values of N up to $N = 40$ could be achieved. At each stage we checked the feasibility of the solution by seeing whether the coefficients did indeed generate a stable scheme. This was necessary because at convergence the norm of the residual of the active constraints is required to be a minimum. Thus several of the constraints may be satisfied as equalities and hence some of the zeros with modulus one may not be simple. In this way we also found that in some cases where convergence had apparently not occurred because the residuals were too large, the non-converged solutions were in fact stable and hence near-optimal. In such cases a minimum was predicted because the other convergence criterion which measured the difference of the gradient vector and a multiple of the Jacobian of the active constraints was small.

Therefore our solutions are only approximations to the optimal solutions. However, experiments show that they are reasonable. In many cases trying to impose the convergence criteria more exactly caused the programme to converge to solutions with $p_o = -1$. This is not acceptable since then convergence of the solution of the numerical scheme to the solution of the differential equation is not guaranteed. When this did occur, a suboptimal solution was obtained by restricting the value of p_o to some smaller interval such as $-.9 \leq p_o \leq 1$. The programme then converged to the solution with p_o on its lower bound confirming that in fact the optimal solution would not be zero stable.

We recall that the Cohn-Schur criterion does not ensure that roots lying on the unit circle are simple but that if strict inequality is imposed, all roots do lie inside the unit circle. However the optimisation procedure relies on the constraints being achievable as equalities and thus we cannot guarantee the stability of a solution. To avoid this difficulty, we did attempt to find a solution to the damped problem where the roots are required to lie on or inside a circle of radius less than one (cf. Verwer [Ve76b]). This approach did not seem to be particularly beneficial, since allowing no roots to have modulus one produced radically reduced Courant numbers. It was somewhat better to stay with the rather *ad hoc* approach of finding approximate solutions by carefully examining non-converged results.

In the course of our investigations, various other

attempts at a solution technique were tried. The method and routine used here gave the best results of those considered. However it is likely that with the development of new routines for solving nonlinear optimisation problems, better results can be achieved. Certainly the schemes developed here will be good candidates with which to begin any other iterative procedure.

In view of our discussion of the normalised error constant in Section 6.2 it would be beneficial to restrict p_o to some value away from -1. As yet the consequences of this course of action have not been fully investigated but we do intend to continue the experiments at a later date.

6.7 Application Of The Technique To Some Specific Problems

Our original intention has always been to design a stable numerical method of third-order accuracy for solving hyperbolic systems of partial differential equations. It is unlikely that any method of time integration with order of accuracy greater than three would be very useful since efficiency is likely to be further reduced. Also, the accuracy of the time integration should be compatible with that of the spatial discretisation which, even if of order three in regions where the solution is smooth, will certainly have reduced accuracy in the neighbourhood of discontinuities. In regions where the solution is rapidly varying, consideration of too many stages in the solution process might lead

to unnecessary accumulation of errors. Therefore, for the present, we only consider the design of two-step schemes with two or three stages and order of accuracy two and three. The schemes of order two are developed for the sake of comparison. We also develop equivalent one-step schemes to see whether there is any benefit in taking the extra step into account.

No of steps	No. of stages	Order p	Free variables
2	2	2	$\{p_o, p_1, p_2\}$
2	2	3	$\{p_o\}$
2	3	2	$\{p_o, p_1, p_2, p_3, s_3\}$
2	3	3	$\{p_o, p_1, p_2, p_3\}$
1	2	2	—
1	3	2	$\{r_3\}$
1	4	2	$\{r_3, r_4\}$
1	3	3	—
1	4	3	$\{r_4\}$
1	5	3	$\{r_4, r_5\}$

Table 6.7.1 Free variables of optimised Runge Kutta schemes.

We summarise the schemes considered and their available degrees of freedom for stability in Table 6.7.1. Here the one-step m-stage schemes are described by the stability polynomial $R(z) = \sum_{i=o}^{m} r_i z^i$.

In the previous section we described a solution technique for a general domain Ω to lie completely inside the stability region. Here we consider four such regions as well as an interval on the imaginary axis. As the stability regions are symmetric about the real axis, we only need to consider points on the portion of the boundary of Ω lying above the real axis.

For systems with spectra in the left half complex plane, we consider Ω to be a wedge-shaped region subtending angle 2α at the origin and bounded by a smooth curve. This is similar to finding schemes with maximal $A(\alpha)$-stability for solving ordinary differential equations (O.D.E.). Therefore the results presented here are equally relevant for O.D.E. solvers as well. If the bounding arc of the wedge is not particularly parabolic in shape, the maximal Courant number obtainable is predominantly determined by the angle of the wedge rather than this arc. The optimal solutions for any bounding arcs will be good approximations to optimality for any solution required in a wedge of similar angle.

We decided to test our solution technique with four wedges. The choice of two of these wedges was guided by the loci of the eigenvalues of the Jacobian matrices describing the SD's given by Equations 5.4.1. By theorem (4.2.1) these are the only third-order implicit semidiscretisations which attain maximal order. The eigenvalues of their Jacobians of size $n+1$ are given by the following equations:

$$\text{i)} \quad \lambda_j = \frac{1}{4}[(\cos^2\theta_j - 4) \pm i\cos\theta_j (28 - \cos^2\theta_j)^{\frac{1}{2}}] \quad 1 \leq j \leq n+1$$

$$\text{ii)} \quad \lambda_j = \frac{6[(\cos^2\theta_j - 4) \pm i\cos\theta_j (13 - \cos^2\theta_j)^{\frac{1}{2}}]}{16 + 5\cos^2\theta_j} \quad 1 \leq j \leq n+1$$

$$\theta_j = \frac{j\pi}{n+1} \quad , \quad 1 \leq j \leq n+1 \quad .$$

The angle, α , which the loci of the eigenvalues subend at the origin is 49.1° and 60° respectively. For these eigenvalue curves μ is maximised so that $\mu\lambda_j \in S$, $1 \leq j \leq n+1$, and so that the ray from $\mu\lambda_{n+1}$ to the origin lies completely in S. The other wedges we considered have bounding curves which are arcs of circles and subtend angles 45° and 90° at the origin. As before, discretisation of the boundary Ω involves discretising along the radius and the bounding arc of the wedge. As it is the radius which is in some sense being maximised, a solution is found more efficiently if a denser distribution of points is taken further away from the origin.

However, as we have already explained, to find optimal schemes for these particular SD's we should have optimised with respect to the spectra of their infinite Toeplitz forms. Thus we have not obtained the results originally required. Nonetheless, we do have optimal schemes for four different wedges and have demonstrated that our solution technique is successful. Clearly maximally stable multistage methods, for any wedge with angle α can be found in this

way.

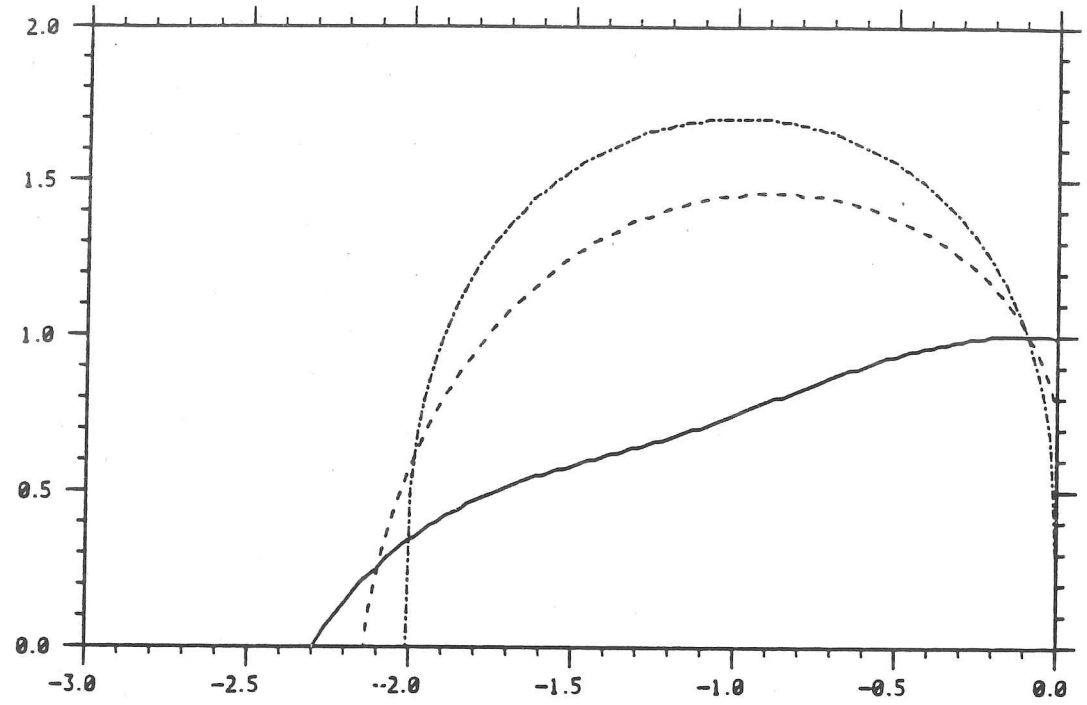
Before describing our results further we briefly mention the application of the technique to optimal schemes for conservative SD's for hyperbolic equations. For such systems optimisation merely requires maximising μ such that the interval $[0, i\mu]$ lies in the stability region. Note that we have already given analytic solutions for the third-order two-step schemes and that some solutions for the one-step schemes are widely known. The remaining problems are solved by the optimisation technique applied by discretising the interval $[0, i\mu]$.

As mentioned in the previous section, we often had problems with p_o becoming close to -1 . In cases where this occurred, we experimented with fixing it away from -1 at values of -0.8 , -0.9 and -0.95 . Therefore we derived several possible schemes in each case. Every possible solution was checked for its validity and in some cases approximations made to the coefficients to bring them into a more compact form. If a solution could be found with coefficients in closed form, then this was the solution that we preferred. Thus the solution of the optimisation problem was in many cases just a starting point from which schemes could be determined by various methods of trial and error. Our solution technique is therefore totally heuristic in form, requiring some formal optimisation and some judgement.

No.of stages m	No.of Steps k	Order p	μ_1	μ_2	μ_I	$\frac{\mu_1}{m}$	$\frac{\mu_2}{m}$	$\frac{\mu_I}{m}$
2	1	2	1.33	1.33	0	.665	.665	—
3	1	2	1.88	2.31	2	.627	.770	.667
4	1	2	3.33	3.33	$2\sqrt{2}$	.833	.833	.707
3	1	3	1.68	1.64	$\sqrt{3}$	.560	.847	.577
4	1	3	2.06	2.39	$2\sqrt{2}$	.515	.598	.707
5	1	3	3.39	3.29	$\sqrt{15}$	.678	.823	.775
2	2	2	2.26	2.27	1.98	1.130	1.135	.99
3	2	2	3.39	3.75	2.93	1.130	1.25	.977
2	2	3	1.33	1.41	1.00	.665	.705	.500
3	2	3	1.71	2.01	2.84	.570	.670	.945

Table 6.7.2 Predicted CFL numbers

In Table 6.7.2 we give the predicted values of maximal Courant numbers attainable. Scaling by the number of stages allows direct comparison between predicted efficiency of different schemes. Here μ_1 , μ_2 and μ_I are the maximal Courant numbers for curves i) and ii) and those on the imaginary axis respectively. In Figure 3 we show some of the optimal stability regions. Note that the maximal radii of the wedges based on the eigenvalue curves are actually larger than the Courant numbers as these have been scaled by $\frac{\sqrt{21}}{6}$ and $\frac{2}{3}$ respectively.



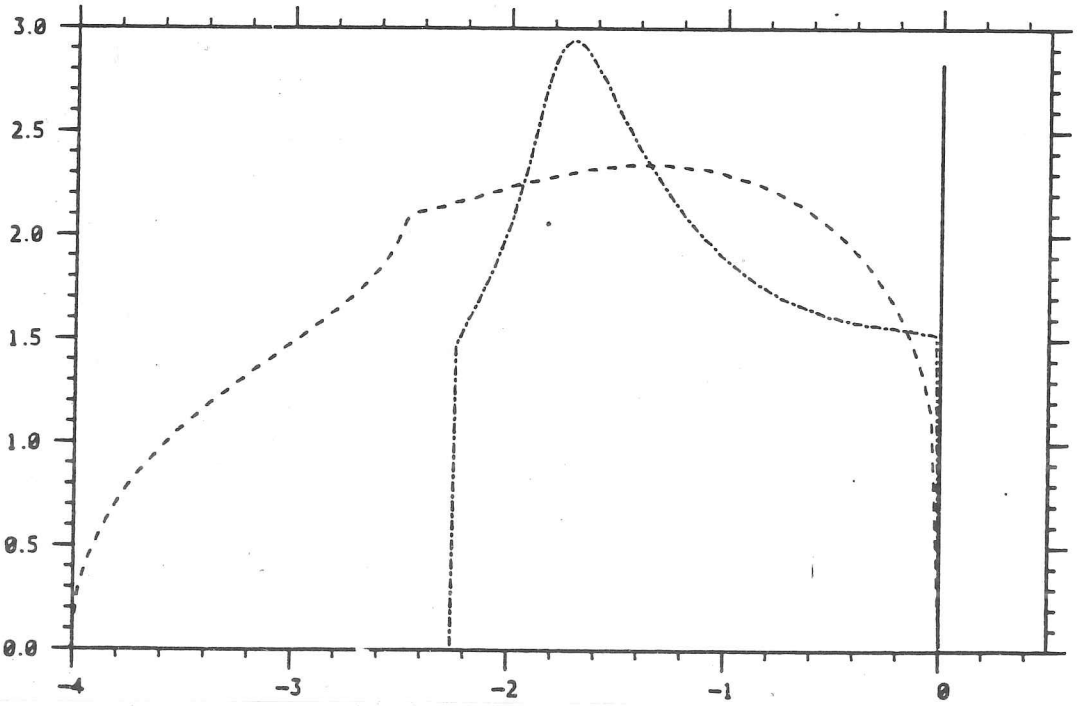


Figure 3. Examples of stability regions of Runge Kutta methods. Schemes with maximal interval of stability on the imaginary axis are represented by the solid lines. Schemes with maximal stability in wedges with semi-angles 49.1° and 60° are represented by the dotted and dashed lines and the dashed lines respectively.

The coefficients obtained as solutions to the optimisation problems are given in Appendix A. In Appendix B we solve for the integration parameters of the schemes in terms of the free variables of the optimisation problem. Thus particular schemes can be derived from the equations in Appendix B.

Although these predictions are not all biased in favour of one class of methods or another, it does appear that the second-order two-step schemes might be particularly useful. Advancing in time with a scaled Courant number greater than one seems particularly promising. However for the dissipative SD's these values are rather too optimistic. As the spectra of the infinite Toeplitz forms vary markedly from those of the finite matrices the Courant numbers we can actually expect are considerably reduced. The expected values are given in Table 1 in Appendix C.

We have shown that this particular class of two-step Runge-Kutta methods can be optimised in a way which produces schemes with improved efficiency within a given domain in $\mathbb{C}^-$. The technique has also been demonstrated to apply equally successfully to one-step Runge-Kutta. Nevertheless, the particular SD's which have been considered are so strongly dissipative that no optimisation will produce any markedly improved efficiency. Instead the procedure should be applied to some SD's possessing less dissipativity. Further, this class of two-step multistage methods may prove to be useful as an explicit O.D.E. solver with improved

stability in wedges within $\mathbb{C}^-$.

In a later chapter we discuss the application of these Runge Kutta methods for the solution of general semi-discretisations. Although we have not gained the efficiency that we had hoped for, the complete numerical model does possess some rather useful properties suggesting that further investigation would be beneficial. Before describing these applications in detail, we discuss various implementation details concerning reduced storage requirements and possible ways of error control.

7. IMPLEMENTATION OF RUNGE KUTTA METHODS

7.1 Efficient Algorithms for One-Step m-Stage Runge Kutta

A resurgence in interest in Runge Kutta methods for integrating hyperbolic equations was initiated recently by Jameson [Ja82]. However, he did not investigate the possibility of designing algorithms with reduced storage, focusing on the implementation of the standard four-stage fourth-order Runge Kutta method without performing any linearisation of the terms. As a result his algorithm is sufficiently general to be applicable to the integration of any nonlinear system of equations. Also, he noted that although fourth-order accuracy might be necessary in the transient stage where rapid change is occurring, the order might be reasonably sacrificed for faster integration when an asymptotic state is reached. Thus we would like to know whether a reduction in order might not only allow larger Courant numbers, but also a reduction in the four arrays of storage which his algorithm requires.

Recall that the main advantage of Runge Kutta schemes is that having obtained the stability polynomial of a given scheme, there is still some freedom available for determining the integration parameters. This freedom may be utilised in a number of ways. Here we discuss algorithms which have minimal storage requirements and allow error control

without too many extra function evaluations. Alternatively, we could have considered minimisation of the local or global error constants, but note that in some cases we have already taken account of the local error by specifying that p_o is not too close to -1. Initially we review some of the algorithms available for efficient implementation of one-step m-stage Runge Kutta.

As we wish to make comparison to Jameson's algorithm, we describe it here, albeit in the notation adopted by Pike and Roe [Pi83]. All the operations of data smoothing and flux balancing are incorporated into a spatial operator $B_{\Delta x}$. Then $\Delta t\, B_{\Delta x}$ acting on an array $W^{(i)}$ is denoted by $z\,W^{(i)}$ where the $W^{(i)}$ denote the generic storage units of the calculation. Additionally, the arrays $W^{(n)}$ and $W^{(n+1)}$ denote y^n and y^{n+1} respectively. Assignment of information to an array is represented by an arrow. With this notation the algorithm is

$$
\begin{aligned}
W^{(1)} &\leftarrow\!- W^{(n)} + \tfrac{1}{2}\, z\, W^{(n)} \\
W^{(2)} &\leftarrow\!- W^{(n)} + \tfrac{1}{2}\, z\, W^{(1)} \\
W^{(3)} &\leftarrow\!- W^{(n)} + z\, W^{(2)} \\
W^{(n+1)} &\leftarrow\!- \tfrac{1}{3}(W^{(3)} + 2W^{(2)} + W^{(1)} - W^{(n)}) + \tfrac{1}{6} z\, W^{(3)} .
\end{aligned}
\qquad (7.1.1)
$$

Jameson's Algorithm

Then, following from the definition of one-step schemes as defined in Appendix B, Jameson's algorithm corresponds to the scheme which has

$$\theta_4 = \frac{1}{6} \ , \ \theta_3 = \frac{1}{3} \ , \ \theta_2 = \frac{1}{3} \ , \ \theta_1 = \frac{1}{6} \ ,$$

$$\lambda_{43} = 1 \ , \ \lambda_{32} = \frac{1}{2} \ , \ \lambda_{21} = \frac{1}{2} \ , \ \lambda_{ij} = 0 \ , \ j \neq i-1 \ .$$

Pike and Roe noticed that if the equation to be solved is linear, then any rearrangement of the stability polynomial is valid. For instance, the stability polynomial can be written in nested form and then interpreted to give an algorithm which is neater but with the same storage requirements. For the four-stage scheme of Jameson, the algorithm becomes

$$\begin{aligned} W^{(1)} &\leftarrow\!- W^{(n)} + \tfrac{1}{4} z W^{(n)} \\ W^{(2)} &\leftarrow\!- W^{(n)} + \tfrac{1}{3} z W^{(1)} \\ W^{(3)} &\leftarrow\!- W^{(n)} + \tfrac{1}{2} z W^{(2)} \\ W^{(n+1)} &\leftarrow\!- W^{(n)} + z W^{(3)} \ . \end{aligned} \qquad (7.1.2)$$

Further, realising that subsequent stages can be thought of as intermediate storage locations and generalising to an arbitrary number of stages, we have an algorithm requiring only three arrays of storage:

$$\begin{aligned} W^{(1)} &\leftarrow\!- W^{(n)} + c_1 z W^{(n)} \\ W^{(2)} &\leftarrow\!- W^{(n)} + c_2 z W^{(1)} \\ W^{(1)} &\leftarrow\!- W^{(n)} + c_3 z W^{(2)} \ . \end{aligned} \qquad (7.1.3)$$

Pike-Roe Algorithm

Here the last two steps are repeated until c_m appears on

the right-hand side and the coefficients c_i are those of the nested polynomial $c_{m-i} = \frac{r_{i+1}}{r_i}$, $0 \leqslant i \leqslant m-1$ [Pi83]. This algorithm is fourth order for linear equations. However, it can be applied for nonlinear equations with a reduction in order if we set

$$\theta_m = 1 \ , \ \theta_j = 0 \ , \ 0 \leqslant j \leqslant m-1 \ ,$$

$$\lambda_{ji} = 0 \quad \text{if} \quad i \neq j-1 \ , \ \lambda_{j,j-1} = c_{j-1} = r_{m-j+2}/r_{m-j+1}.$$

Substituting these coefficients in the consistency conditions for one step schemes, as given in Appendix B, it is clear that for nonlinear equations the algorithm is of second order because $c_{m-1} = \frac{1}{2}$.

Immediately, we question whether by some other rearrangement of the linear algorithm, fourth order might be obtained for nonlinear equations. Iserles adopted an algorithm not based on the nested form of the polynomial $R(z)$ but one with integration parameters of the scheme incorporated as follows:

$$\begin{aligned} W^{(1)} &\leftarrow\!- \quad z\,W^{(n)} \\ W^{(2)} &\leftarrow\!- \quad W^{(n)} + \theta_1 W^{(1)} \\ W^{(1)} &\leftarrow\!- \quad z\,(W^{(n)} + c_1 W^{(1)}) \\ W^{(2)} &\leftarrow\!- \quad W^{(n)} + \theta_2 W^{(1)} \ , \end{aligned} \tag{7.1.4}$$

Iserles's Algorithm I

[Is84d]. Again, the last two steps are repeated until θ_m occurs on the right-hand side. Hence the scheme is similar

to the one obtained from the nested polynomials without the restriction $\theta_j = 0$, $1 \leqslant j \leqslant m-1$. We need to solve for the coefficients $c_i = \lambda_{i+1,i}$ and θ_i in a way that achieves maximal order. Iserles demonstrates that in fact a four-stage order-four method can be achieved in this way.

A further refinement in the definition of the scheme enables Iserles's algorithm (7.1.4) to be implemented with just two arrays of storage. The intermediary points y_i^n are redefined by,

$$y_i^n = y^n + h \sum_{j=1}^{i-1} \theta_j \ell_j + h e_{i-1} \ell_{i-1} \qquad (7.1.5)$$

giving an algorithm where storage of y^n is not needed so that the information in $W^{(n)}$ is overwritten during the calculation. Thus (7.1.4) becomes

$$\begin{aligned} W^{(1)} &\leftarrow\!- \quad z W^{(n)} \\ W^{(n)} &\leftarrow\!- \quad W^{(n)} + \theta_1 W^{(1)} \\ W^{(1)} &\leftarrow\!- \quad z (W^{(n)} + e_1 W^{(1)}) \\ W^{(n)} &\leftarrow\!- \quad W^{(n)} + \theta_2 W^{(1)} \quad . \end{aligned} \qquad (7.1.6)$$

Iserles's Algorithm II

Here the last two steps are repeated until θ_m appears on the right hand side. If, in addition, the restriction $\theta_j = 0$, $1 \leqslant j \leqslant m-1$ is imposed, as in the non-linear version of (7.1.3), the increment in $W^{(n)}$ need only be made once and the algorithm is then computationally equivalent to (7.1.3). The algorithm can be made fourth order for a four-stage

scheme if the restriction is slightly relaxed so that $\theta_j = 0,\ 1 \leqslant j \leqslant m-2$ [Is84d]. Therefore these algorithms enable higher order to be obtained with reduced storage whilst maintaining efficiency comparable to those of either Pike and Roe or Jameson.

It is therefore clear that with judicious choices of integration parameters, storage may be reduced whilst maintaining high order of accuracy. At the same time, the number of operations can be radically reduced. Working with higher dimensional problems these criteria are very important and thus implementation of the normal Runge Kutta methods with one of the algorithms described here should be considered. In the following section we investigate whether our class of two-step m-stage formulae can be implemented in a comparable way.

7.2 Efficient Algorithms For Two-Step m-Stage Runge Kutta

As a result of the analysis of the one-step schemes it seems unlikely that implementation of the particular class of two-step formulae will be possible with just two arrays of storage. These formulae were chosen to be efficient in terms of function evaluations and, to take advantage of this, information calculated at one step must be stored for use at subsequent steps. Thus at least one array of storage is required by the very structure of the formulae.

Following the ideas of Pike and Roe [Pi83] for one-step

formulae, we can base the implementation of two-step methods on the nested form of the stability polynomials. It is always assumed that there has already been integration by a suitable one-step method and that the current solution is in array $W^{(n)}$. Also, all the information about the preceding step necessary to continue integration is assigned to array $W^{(1)}$. The arrays $W^{(2)}$ and $W^{(3)}$ are generic units of storage. Once the information in $W^{(n)}$ and $W^{(1)}$ has been used these arrays also become intermediary storage units of the calculation being used to accumulate the information ready for the next step, $W^{(n)}$ on exit being y^{n+1}. Then Pike and Roe's algorithm applied to two-step schemes takes the form,

$$\begin{aligned} W^{(2)} &\leftarrow\!- (1 - \beta)\, W^{(n)} + W^{(1)} \\ W^{(1)} &\leftarrow\!- \beta\, W^{(n)} \\ W^{(3)} &\leftarrow\!- z\, W^{(n)} \\ W^{(n)} &\leftarrow\!- W^{(n)} + c_1\, W^{(3)} \\ W^{(3)} &\leftarrow\!- z\, W^{(n)} \end{aligned} \qquad (7.2.1a)$$

Two-step Pike-Roe Algorithm I

When the last two steps have been performed (m-1) times so that c_{m-1} has appeared, $W^{(1)}$ is incremented for use at the next step and $W^{(n)}$ is incremented to contain y^{n+1},

$$W^{(1)} \leftarrow\!- W^{(1)} + v_m W^{(3)}$$

$$W^{(n)} \leftarrow\!- W^{(2)} + w_m W^{(3)} \ . \qquad (7.2.1b)$$

Thus (7.2.1a) with (7.2.1b) define an algorithm which we

describe as the first two-step version of Pike and Roe's algorithm. The coefficients c_i are defined as those of the nested polynomial by

$$c_{m-i} = \alpha_{m-i+1,\, m-i} = \frac{s_{i+1}}{s_i} = \frac{p_{i+1}}{p_i} \quad , \quad 1 \leqslant i \leqslant m-1 \, ,$$

$$c_m = w_m \quad .$$

The other integration parameters, except for v_m , are zero. For $m \geqslant 2$ stages, this algorithm can achieve third order but is not convergent for $m = 1$ since $p_o = -1$. Also a three-stage method cannot be made fourth-order accurate.

Imposing $p_o = 0$, gives a two-step version of Iserles's first algorithm which can be implemented using just two arrays of storage. On exit from the algorithm after one integration, all the information necessary to integrate further is contained in $W^{(1)}$ and thus the algorithm starts as follows:

$$\begin{aligned} W^{(n)} &\leftarrow\!- \quad W^{(n)} + v_m W^{(1)} \\ W^{(1)} &\leftarrow\!- \quad z\, W^{(n)} \\ W^{(n)} &\leftarrow\!- \quad W^{(n)} + c_1 W^{(1)} \quad . \end{aligned} \tag{7.2.2}$$

Two-step Iserles Algorithm I

The last two steps continue as before until $c_m = w_m$ appears on the right-hand side. This time no additional assignments are needed and the algorithm restarts. For

$m \geq 3$, we can achieve order of accuracy at least three.

The three-stage scheme is completely defined by the above algorithm if it has order of accuracy $p = 3$. This occurs as the extra third-order condition,

$$(s_2 + p_2)\alpha + \frac{(s_3 + p_3)}{\delta \alpha^2} ((\gamma + \delta)^2 - \alpha(\gamma + 3\delta)) = 0$$

and conditions $\gamma = 0$ and $c_{m-i} = \frac{p_{i+1}}{p_i}$ mean that necessarily $p_2^2 = 2p_1 p_3$. Thus p_2 , p_1 and p_3 can be evaluated from the consistency conditions as

$$p_2 = (1 - s_1)(\frac{3}{2} - s_1) \quad p_3 = (1 - s_1)(\frac{7}{6} + s_1^2 - 2s_1)$$

$$p_1 = (1 - s_1) \quad \text{where} \quad s_1 = \frac{1}{2}(1 \pm \frac{2}{\sqrt{6}}) \ .$$

Substituting in the value for s_1 and evaluating only the necessary coefficients p_1 , c_1 , and c_2 gives

$$v_3 = p_1 = \frac{1}{2}(1 \mp \frac{2}{\sqrt{6}}) \quad c_1 = \frac{1}{2}(1 \mp \frac{1}{\sqrt{6}}) \quad c_2 = (1 \mp \frac{1}{\sqrt{6}}) \ .$$

Thus there are no degrees of freedom available for the development of stabilised formulae. Further, it can be shown that the maximal interval of stability on the imaginary axis is obtained for the choice $s_1 = \frac{1}{2}(1 + \frac{2}{\sqrt{6}})$ and is $\beta_{imag} \approx 1.8110$. This is away from the optimal bound $\beta_{imag} = 3$

and consequently this algorithm is of limited applicability for the solution of hyperbolic SD equations.

Alternatively, we may return to Pike and Roe's algorithm (7.2.1) and relax the conditions on the coefficients. The algorithm is then similar regarding the amount of storage but requires more operations. However, we still insist that $\alpha_{ij} = 0$, $j \neq i-1$. Writing $c_i = \alpha_{i+1,i}$ the algorithm is

$$
\begin{aligned}
W^{(2)} &\leftarrow\!- W^{(n)} \\
W^{(3)} &\leftarrow\!- z W^{(n)} \\
W^{(n)} &\leftarrow\!- (1 - \beta) W^{(n)} + W^{(1)} + w_1 W^{(3)} \\
W^{(1)} &\leftarrow\!- \beta W^{(2)} + v_1 W^{(3)} \\
W^{(2)} &\leftarrow\!- W^{(2)} + c_1 W^{(3)} \\
W^{(3)} &\leftarrow\!- z W^{(2)} \\
W^{(n)} &\leftarrow\!- W^{(n)} + w_2 W^{(3)} \\
W^{(1)} &\leftarrow\!- W^{(1)} + v_2 W^{(3)} \quad .
\end{aligned}
\tag{7.2.3}
$$

Two-step Pike-Roe Algorithm II

Again the last four steps are repeated until v_m has appeared on the right-hand side. With this algorithm fourth-order accuracy for $m \geq 3$ can be achieved since the only restriction is $\alpha_{ij} = 0$, $j \neq i-1$.

It is also interesting to consider an algorithm similar to Iserles's second version. The intermediary points can be defined as in (7.1.5) by

$$y_i^n = y^n + h \sum_{j=1}^{i-1} (w_j \ell_j) + h e_{i-1} \ell_{i-1}$$

and

$$y_i^{(n-1)} = y^{(n-1)} + h \sum_{j=1}^{i-1} w_j k_j + h e_{i-1} k_{i-1} \quad . \tag{7.2.4}$$

Then, with $v_i = 0$, $1 \leqslant i \leqslant m-1$, an algorithm which uses only three arrays of storage is as follows:

$$\begin{aligned}
W^{(n)} &\leftarrow\!- \quad (1 - \beta) W^{(n)} + W^{(1)} \\
W^{(1)} &\leftarrow\!- \quad \beta W^{(n)} \\
W^{(2)} &\leftarrow\!- \quad z W^{(n)} \\
W^{(n)} &\leftarrow\!- \quad W^{(n)} + w_1 W^{(2)} \\
W^{(2)} &\leftarrow\!- \quad z (W^{(n)} + e_1 W^{(2)}) \\
W^{(n)} &\leftarrow\!- \quad W^{(n)} + w_2 W^{(2)} \quad .
\end{aligned} \tag{7.2.5a}$$

Two-step Iserles Algorithm II

The last two steps are repeated until w_m has appeared and then $W^{(1)}$ is incremented,

$$W^{(1)} \leftarrow\!- \quad W^{(1)} + v_m W^{(2)} \quad . \tag{7.2.5b}$$

Alternatively, the first step of the algorithm could be,

$$W^{(n)} \leftarrow\!- \quad (1 - \beta) W^{(n)} + W^{(1)} + v_m W^{(2)} \quad . \tag{7.2.6}$$

Clearly, imposing $\beta = 0$ allows the method to be implemented with just two arrays of storage. Although a three-stage scheme cannot be made fourth-order accurate with this choice of coefficients, this algorithm is very encouraging. It is

certainly much more than we had hoped for and suggests that possibly optimisation of schemes designed for this algorithm should be performed.

Therefore the most promising approach here, is to base algorithms either on Iserles first algorithm or its refinement (7.2.6) which needs less storage. If there is no need for order four, with only a few stages, the latter should be recommended. Unfortunately, the very properties that make it so desirable here remove some freedom when we try to design an error-control mechanism. We note that for most of our formulae the polynomial coefficients are not known in closed form. This may be a disadvantage for the algorithms based on the integration parameters of the schemes since the stability regions of the optimised methods are very sensitive to perturbations of the polynomial coefficients and so they must be calculated to the greatest accuracy possible to prevent instability arising. This instability would not be a problem for the nested form algorithms. This, along with the reduced number of operations required by these algorithms, suggests a preference for Pike and Roe's first version. However, we shall see that, in fact, it is beneficial to have the extra degrees of freedom of the second version for the design of error control mechanisms. Nonetheless, if no error control is required the two-step versions of Iserles's second algorithm should be preferred since they only need either two or three arrays of storage.

At present the two-step formulae have not been tested

on any large scale practical problems. For the test problems considered the integration has been started using a one-step Runge Kutta method chosen with comparable stability and order of accuracy properties. In this way it is ensured that the starting mechanism does not introduce unnecessary errors in the initial conditions. The method of steplength adjustment has until now been very crude. If instability is evident in the solution then the integration has been started with a less optimistic Courant number. Obviously this is not a sensible course of action in practical problems and thus we have begun to investigate a method for automatic steplength adjustment and error control. We review our ideas in the next section, although they are of a very preliminary nature and have not yet been implemented.

7.3 A Strategy for Error Control of Two-Step Schemes

Here we consider the design of higher-order two-step methods which may be used as part of a numerical scheme with automatic error control. The main requirement is that this should be performed as cheaply and as efficiently as possible,ie. we desire not only minimal extra function evaluations and operations but also minimal storage.

A method of error control very similar to that proposed by Verwer [Ve80] for three-step Runge-Kutta methods is suggested. Here we only describe the major differences between our method and that presented by Verwer. For more details

concerning implementation the reader is referred to this reference. His method forms estimates of the local error by forming approximations to the derivatives occurring in the Taylor expansions. Instead, the method presented here is based on the assumption that given two formulae from the same class of methods, one with higher order than the other, then the difference between solutions at a given time is an approximation to the error in the lower-order formula. This is based on the Runge-Kutta-Fehlberg technique usually used for error control of one-step methods.

If the estimate of the error does not satisfy a suitable error tolerance then the stepsize must be reduced. It is not desirable to implement the error check too often as this would be costly, due to the extra function evaluations required to restart the process. It is suggested that Verwer's technique of stepsize adjustment is adopted whereby he uses the root formula to predict a stepsize which is then scaled down to produce a conservative estimate for the new step. Similarly, if after a set number of steps no adjustment has been necessary the same formula can be used to predict an increased stepsize. Note that any increased step must always be bounded by that predicted as maximal for a linear equation being integrated by the particular scheme.

As already mentioned, each time that a new step is adopted or when the integration starts, the first step must be evaluated in some way. This may be by interpolation or using a one-step method. Verwer uses interpolation at points

during the integration and a one-step Runge-Kutta method to start the process. We suggest that, if it is not too costly, a one-step method should be used at all times. Clearly the one-step method should be chosen with stability properties comparable to the main integrator and also a similar maximal Courant number. Then, instabilities which may arise by a too frequent use of interpolation should not be a problem. Obviously the advantages and disadvantages of any error mechanism can only be determined by a detailed performance evaluation. Verwer's model has been proven whilst the modifications suggested here have not yet been tested.

Some two-step schemes compatible with our original methods and designed for maximal efficiency are now presented. The higher-order scheme, which is a member of the same class of formulae as discussed previously, is defined by:

$$y_e^{n+1} = (1 - \beta^*) y^n + \beta^* y^{n-1} + h \sum_{i=1}^{m} (v_i^* k_i + w_i^* \ell_i) \quad . \quad (7.3.1)$$

Here we are requiring that the intermediary points of the calculation are the same in both cases. Thus, the coefficients $\{\alpha_{ij}\}$ are the same in both formulae and if we can obtain higher order with no extra stages, then the error control does not require any extra function evaluations. The difference $\| y_e^{n+1} - y^{n+1} \|$ will be an approximation to the error in y^{n+1}.

In general the order conditions and the coefficients $\{\alpha_{ij}\}$ do not define the coefficients of the error scheme (7.3.1) completely. There is still some freedom available

which can be used to limit storage and reduce computer operations. Before considering these points, we concentrate on the stability of scheme (7.3.1). As we are not integrating globally with this scheme, its global stability is not crucial. However, we have already mentioned that the normalised error constant of a consistent method is the truncation error normalised by $1 + \beta^*$ and thus has a magnitude determined by the zero stability parameter β^* . It is desirable that the order of magnitude of the errors of both schemes is comparable. Therefore, wherever we are able to do so we define $\beta^* = \beta$. Indeed, this is also useful for reducing the number of computer operations.

For the three-stage schemes of order $p \leqslant 3$ and the two-stage scheme of order $p = 2$, we can implement the error control scheme with just one additional array of storage. Having decided at a particular step that an error check is to be made, we use that step to assign the information,

$$\beta^* y^{n-1} + h \sum_{i=1}^{m} v_i^* k_i \quad ,$$

to $W^{(4)}$. Then at the next step we have no need to store information for subsequent steps and so we can just add the information

$$(1 - \beta^*) y^n + h \sum_{i=1}^{m} w_i^* \ell_i$$

to that already present in $W^{(4)}$. We may reduce the number of operations if we can make as much as possible of $\beta^* y^{n-1} + h \sum_{i=1}^{m} v_i^* k_i$ the same as $\beta y^{n-1} + h \sum_{i=1}^{m} v_i k_i$. We therefore use these criteria to define the coefficients of the higher-order scheme completely.

We give, for example, the algorithm for a two-stage scheme based on the two-step version of Pike and Roe's second algorithm with error control incorporated:

$$
\begin{aligned}
& W^{(2)} \leftarrow W^{(n)} \\
& W^{(3)} \leftarrow z\,W^{(2)} \\
& W^{(n)} \leftarrow (1-\beta)W^{(n)} + W^{(1)} + w_1 W^{(3)} \\
& W^{(1)} \leftarrow \beta W^{(2)} + v_1 W^{(3)} \\
& W^{(4)} \leftarrow \beta W^{(2)} + v_1^{*} W^{(3)} \\
\Gamma \quad & \left[\begin{aligned}
& W^{(2)} \leftarrow W^{(2)} + c_1 W^{(3)} \\
& W^{(3)} \leftarrow z\,W^{(2)} \\
& W^{(n)} \leftarrow W^{(n)} + w_2 W^{(3)} \\
& W^{(1)} \leftarrow W^{(1)} + v_2 W^{(3)} \\
& W^{(4)} \leftarrow W^{(4)} + v_2^{*} W^{(4)}
\end{aligned}\right. \\
& W^{(2)} \leftarrow W^{(n)} \\
& W^{(3)} \leftarrow z\,W^{(2)} \\
& W^{(n)} \leftarrow (1-\beta)W^{(n)} + W^{(1)} + w_1 W^{(3)} \\
& W^{(1)} \leftarrow \beta W^{(2)} + v_1 W^{(3)} \\
& W^{(4)} \leftarrow W^{(4)} + (1-\beta)W^{(2)} + w_1^{*} W^{(3)} \\
\Lambda \quad & \left[\begin{aligned}
& W^{(2)} \leftarrow W^{(2)} + c_1 W^{(3)} \\
& W^{(3)} \leftarrow z\,W^{(2)} \\
& W^{(n)} \leftarrow W^{(n)} + w_2 W^{(3)} \\
& W^{(1)} \leftarrow W^{(1)} + v_2 W^{(3)} \\
& W^{(4)} \leftarrow W^{(4)} + w_2^{*} W^{(4)} \quad .
\end{aligned}\right.
\end{aligned}
$$

For clarity we have not included any loops but obviously for more stages they occur at Γ and Λ. On exit from this algorithm the difference $y_e^{n+1} - y^{n+1}$ is represented by $W^{(4)} - W^{(n)}$. A routine to determine the next course of action is entered and we then return to the original algorithm (7.2.3) for some predetermined number of steps. Notice that this algorithm uses only five arrays of storage and that setting $v_i = v_i^*$ would remove an appreciable number of operations if the number of stages is high.

We use this algorithm because the freedom existing in the coefficients allows error control for schemes with few stages. In Appendix B we give these coefficients for both three-stage schemes and the two-stage second-order scheme. As explained above, we have used the freedom in each case to set $\beta = \beta^*$ and $v_i = v_i^*$ $1 \leqslant i \leqslant j$ for j as large as possible. For a two-stage scheme, no fourth-order method exists and so error control cannot be implemented without resort to extra function evaluations. Alternatively, we could use a three-step fourth-order scheme with two function evaluations. We have not investigated this possibility as yet and so in Appendix B we give coefficients for a compatible three-stage third-order scheme designed with the same aims as above.

Clearly we must explain why we have decided to base our error control schemes on this version of Pike and Roe's algorithm rather than the first version or either of those of Iserles. We already noted that with Iserles's first

algorithm we cannot have order three and less than four stages. Therefore it is not directly useful here, but, with increased number of stages it may be preferable. The first version of Pike and Roe's algorithm which is based on using polynomials in nested form yields a third-order two-stage scheme with $p_o = -1$. For the third-order three-stage scheme, the extra third-order condition of the error control schemes removes some of the freedom of the original scheme. The second version of Iserles's algorithm is particularly promising if no error control is being imposed. However, as with its one-step counterpart, the structure of the formula imposes extra constraints when we try to derive two schemes which are compatible. So again, this algorithm is not suitable in our framework.

It is not our purpose in this dissertation to investigate one-step schemes and after all, well-tried methods of error control for one-step Runge Kutta already exist. However, having surveyed some algorithms for both one- and two-step methods, and demonstrated that in the two-step scheme not all the algorithms are practicable for error control, we must make some comments about similar problems for the one-step method.

Pike and Roe's algorithm can only attain order two and is thus immediately put to one side unless we are interested in a low level of accuracy. If we want error control without extra function evaluations, then the structure of the defining formula for Iserles's second algorithm imposes

that the lower-order formula has the same coefficients $\{\theta_i\}$ as the error control method and it must also be disregarded. We are then left with Pike and Roe's algorithm as the only choice, thus four storage locations are required. As for the two-step methods, the available degrees of freedom of the algorithm may be used to minimise computer operations. Obviously if we are already integrating with $p = m$, extra function evaluations cannot be avoided if error control is incorporated.

8. AN EVALUATION OF SOME FINITE-DIFFERENCE SCHEMES

8.1 Dissipation, Dispersion, Phase Velocity and Group Velocity

Throughout this dissertation we have concentrated on the design of stable numerical models suitable for the integration of hyperbolic partial differential equations. However, any numerical model or exact solution also exhibits other characteristics which we have, so far, not discussed in detail. Before examining the results of our experiments which have been designed to highlight some of these properties we present a *group velocity* analysis of the SD model. Our presentation is similar to that in Vichnevetsky and Bowles [Vi82]. For a more physical interpretation of group velocity the reader may refer to one of the more standard texts on wave analysis, for example Brillouin [Br53].

We concentrate on the error intrinsic in the form of the semi-discretisation,

$$\frac{d}{dt} v_j(t) = B_{\Delta x} v_j(t), \quad 1 \leqslant j \leqslant n \quad , \qquad (8.1.1)$$

where $B_{\Delta x}$ may either be an explicit or an implicit operator. Let us compare the propagation of a single Fourier mode $e^{i\xi x}$ with wavenumber ξ by this equation, to that by the linear equation $u_t = u_x$. The exact solution,

$$u(x,t) = e^{i\xi(x+t)} ,$$

has phase speed 1 and constant amplitude.

We define $\hat{B}(\xi)$ to be the *spectral function* obtained by the Toeplitz operator $B_{\Delta x}$ acting on the Fourier modes $\{ e^{i\xi x_j} \}$. Then the function

$$w_{\xi,j}(t) = v_j(t) e^{i\xi x_j}, \qquad (8.1.2)$$

where $w_\xi(t)$, representing the propagation of the Fourier mode with wavenumber ξ, is a solution of (8.1.1) only if $v(t)$ is a solution of the equation

$$\frac{d}{dt} v(t) = \hat{B}(\xi) v(t) . \qquad (8.1.3)$$

Now, this equation has solution,

$$v(t) = v(0) e^{\hat{B}(\xi)t} ,$$

hence by (8.1.2) the solution to (8.1.1) is

$$w_{\xi,j}(t) = v_j(0) e^{\hat{B}(\xi)t} e^{i\xi x_j} . \qquad (8.1.4)$$

Consequently the numerical solution has *amplitude*

$$| w_\xi(t) | = | v(0) | e^{\mathrm{Re}\, \hat{B}(\xi)t} \qquad (8.1.5)$$

and *phase speed*

$$c(\xi) = \mathrm{Im}\frac{\hat{B}(\xi)}{\xi} \ . \qquad (8.1.6)$$

Therefore solutions of the semi-discretised equation will suffer *attenuation* of amplitude in time if $\mathrm{Re}\,\hat{B} < 0$. We say that the solution is *dissipative* of *order r* if,

$$\mathrm{Re}\,\hat{B}(\xi) = -\kappa\xi^{r} + O(\xi^{r+1}) \ , \qquad (8.1.7)$$

where $\kappa > 0$ is a real constant. This compares with the definition of dissipativity as given in Chapter 1 for fully discrete systems. Further, the solution is *conservative* if $\mathrm{Re}\,\hat{B} \equiv 0$. Notice that $\mathrm{Re}\,\hat{B} > 0$ is ruled out by stability considerations and since dissipativity ensures boundedness of solutions, it is sufficient for stability of the Cauchy problem.

Moreover, the phase speed of the numerical solution is dependent on the wavenumber and so initial conditions will not maintain their shape under propagation as they would with the exact solution. Instead, different Fourier modes propagate at different speeds and *dispersion* of polychromatic signals necessarily occurs.

However, the evolution of solutions of the SD equations is not completely determined by phase speed. We saw in Section 1.7 that linear partial differential equations

supporting solutions $e^{i(\omega t + \xi x)}$ have frequency ω which satisfies a *dispersion relation*,

$$\omega = \omega(\xi) \quad .$$

Also, for such solutions, energy associated with wavenumber ξ propagates asymptotically with the group speed

$$C(\xi) = \frac{d}{d\xi}\omega(\xi) \quad . \qquad (8.1.8)$$

Therefore exact solutions of the linear equation being considered, have $c(\xi) \equiv C(\xi) \equiv 1$ since $\omega = \xi$.

Contrary to this, the following analysis demonstrates that group speed and phase speed are not equivalent for the SD equation. To derive an expression for the group speed of this equation we investigate the propagation of a superposition of Fourier modes with nearly identical wavenumbers. An initial condition of this kind gives rise to an envelope of waves called a *wave packet*. As the wave packet consists of waves with low frequency variation we may define a local phase η and express each Fourier mode as

$$u_\eta(x,0) = a_\eta e^{i(\eta+\xi)x} \quad , \qquad (8.1.9)$$

so that the initial condition is

$$u(x,0) = \sum_\eta u_\eta(x,0) \quad . \qquad (8.1.10)$$

Then by (8.1.4) each mode propagates as

$$w_{\xi+\eta}(t) = a_\eta \, e^{i(\xi+\eta)x} \, e^{\hat{B}(\xi+\eta)t} \quad . \qquad (8.1.11)$$

Moreover, $|\eta| \ll |\xi|$ means that we may expand $\hat{B}(\xi+\eta)$ by Taylor expansion giving, to terms linear in η

$$w_{\xi+\eta}(t) = a_\eta \, e^{i\xi(x+c(\xi)t)} \, e^{i\eta(x+C(\xi)t)} e^{\mathrm{Re}\,\hat{B}(\xi)t} \quad , (8.1.12)$$

where

$$C(\xi) = \frac{d}{d\xi}(-i\hat{B}(\xi)) = \frac{d}{d\xi}(\mathrm{Im}\,\hat{B}(\xi)) - i\frac{d}{d\xi}(\mathrm{Re}\hat{B}(\xi)) \quad . (8.1.13)$$

Hence the group speed is independent of η and for conservative schemes (8.1.12) is the expression for a function with phase velocity $c(\xi)$ containing an envelope propagating without deformation at real velocity $C(\xi)$. For dissipative schemes we see that the evolution of the envelope in time is more complicated as the group speed is complex and its physical meaning is unclear. It might be expected that the real part of $C(\xi)$ takes the role of group speed and that the imaginary part measures the rate of dissipation of the envelope. Later, we investigate this for just two examples.

Since the general physical relevance is unclear we now assume that the model is conservative. Then any initial condition which may be decomposed as (8.1.10) will propagate in

a manner dictated by its individual modes and the initial superposition of Fourier modes propagates as

$$w_{\xi,j}(t) = \sum_{\eta} w_{\xi+\eta,j}(t)$$

$$= e^{i\xi(x_j+c(\xi))t} \sum_{\eta} a_{\eta} e^{i\eta(x+C(\xi)t)} \quad . \tag{8.1.14}$$

Fourier transforming this solution and applying Parseval's equality we see that the energy contained in the wave packet is given by

$$\| v \|_2^2 \equiv h \sum_{j=-\infty}^{+\infty} | v_j |^2$$

$$= \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} | w_{\xi}(t) |^2 \, d\xi \quad . \tag{8.1.15}$$

where h is the mesh size of the grid, $h = \Delta x$. As the envelope propagates without deformation, the energy contained in the wave packet propagates at the group speed $C(\xi)$ rather than the phase speed $c(\xi)$. Alternatively, if dissipation is introduced group speed no longer measures the flow of energy and the *energy velocity* needs to be considered [Br53].

Consequently, is has been shown that not only is the semidiscrete solution dispersive but it also transports energy at a different speed and may be attenuated. We would expect that, where dispersion dominates over dissipation,

the predictions of group velocity would still be approximately valid but that for strong dissipation the analysis presented is incomplete.

The notion of dissipativity discussed here is, in fact, equivalent to that of *monotonicity*, as described for example, [Ve83], [Sp83], [Da79] and [Ro82]. The numerical method is said to be unconditionally monotone if it is monotone for all grid sizes. Then Godunov's Theorem says that no multistep multistage method for solving the O.D.E. system has order $p > 1$ if it is unconditionally monotone [Ro82].

Now, the homogeneity of 8.1.1 means that *contractivity*, as discussed by Spijker, is equivalent to monotonicity and therefore his result, that unconditionally contractive methods have order bounded by one, is actually Godunov's Theorem in this framework. Moreover, other results for contractive methods are applicable. Spijker has shown that an implicit method, the trapezoidal rule, is contractive only in the circle of radius 1 with centre at $(-1,0)$. He has also given conditions for multistep multistage methods to be contractive in circles centred at $(-r,0)$ with radius r. From these conditions we have determined the largest circles for which some of our Runge Kutta methods are contractive (cf. Table 4 in Appendix C). Consequently, if contractivity, instead of stability, is imposed, then the maximal Courant numbers for integration are severely restricted. Verwer and Dekker have investigated the possibility of making Runge Kutta methods either conservative or contractive.

In particular, they demonstrate the technique for the classical fourth-order explicit Runge Kutta method [Ve83].

Recall that we did not choose to design methods based on these criteria. However, this brief discussion is relevant as it serves to highlight some of the properties of the numerical experiments described in the following sections.

8.2 The Finite-Difference Schemes

In this section we describe the experiments we have performed to compare the numerical solutions of hyperbolic partial differential equations by dissipative or conservative semi-discretisations integrated by standard or non-standard O.D.E. solvers. The SD's we use are two standard conservative schemes based on central differences as well as the two dissipative schemes given by 5.4.1. For time integration we employ the one- and two-step Runge-Kutta methods derived in Chapter 6 and two other standard second order methods. When integrating by the Runge-Kutta methods we use the method with stability region most suited to the location of the spectrum of the Toeplitz operator of the SD equation.

First we define the SD schemes, which we label by A, B, C and D, and give their order of accuracy p:

A) Case A p = 3:

$$2\frac{dv_j}{dt} + \frac{dv_{j+1}}{dt} = \frac{1}{\Delta x}\left(-\frac{1}{2}v_{j-1} - 2v_j + \frac{5}{2}v_{j+1}\right) .$$

B) Case B p = 3:

$$\frac{dv_{j-1}}{dt} - \frac{8\,dv_j}{dt} - \frac{5\,dv_{j+1}}{dt} = \frac{12}{\Delta x}(v_j - v_{j+1}) . \qquad (8.2.1)$$

C) Three-point central difference, p = 2:

$$\frac{dv_j}{dt} = \frac{1}{2\Delta x}(v_{j+1} - v_{j-1}) .$$

D) Five-point central difference, p = 4:

$$\frac{dv_j}{dt} = \frac{1}{2\Delta x}\left[\frac{4}{3}(v_{j+1} - v_{j-1}) - \frac{1}{6}(v_{j+2} - v_{j-2})\right] .$$

Clearly, from the previous section, the spectral functions of these SD's may be found by inserting $v_j(t) = v(t)e^{i\xi x_j}$, and cancelling common factors. Defining $\theta = \xi\Delta x$, $0 \leq \theta \leq 2\pi$, gives the following expressions,

A) $$\hat{B}_1(\xi) = \frac{1}{\Delta x}\left[\frac{-(\cos\theta - 1)^2 + i\sin\theta(8 + \cos\theta)}{5 + 4\cos\theta}\right] .$$

B) $$\hat{B}_2(\xi) = \frac{6}{\Delta x}\left[\frac{-(\cos\theta - 1)^2 + i\sin\theta(7 - \cos\theta)}{25 + 16\cos\theta - 5\cos^2\theta}\right] .$$

C) $\hat{B}_3(\xi) = \frac{i \sin\theta}{\Delta x}$. (8.2.2)

D) $\hat{B}_4(\xi) = \frac{i}{3\Delta x} \left[4\sin\theta - \frac{1}{2}\sin 2\theta \right]$.

The curves $\Delta x \hat{B}_1(\xi)$, $\Delta x \hat{B}_2(\xi)$ are the dashed curves given in Figure 2. Obviously in each case these SD's are strongly dissipative as $\mathrm{Re}\,\hat{B}_i(\xi)$, $i = 1,2$ extends a long way into $\mathbb{C}^-$. In comparison, the curves defined by $\hat{B}_3(\xi)$ and $\hat{B}_4(\xi)$ are the intervals $\frac{i}{\Delta x}[-1,1]$ and $\frac{i}{\Delta x}[-1.37, 1.37]$ on the imaginary axis respectively. Hence cases C and D are conservative SD.

As well as the Runge-Kutta methods the other O.D.E. solvers we use are an explicit method, the midpoint rule and an implicit method, the trapezoidal rule.

1) **Midpoint p = 2:**

$$v_j^{n+1} = v_j^{n-1} + 2\,\Delta t\, f(v_j^n) \quad . \qquad (8.2.3)$$

2) **Trapezoidal p = 2:**

$$v_j^{n+1} = v_j^n + \frac{\Delta t}{2}\left(f(v_j^{n+1}) + f(v_j^n)\right) \quad .$$

It is well known that the midpoint rule has interval of stability $[-i,i]$, whereas the trapezoidal rule is

unconditionally stable. Therefore we might expect to integrate all of A,B,C and D with the latter method but only C and D with the former. For the midpoint rule Courant numbers will be limited by 1.0 and .73 respectively. Note that this value of .73 is much less than the value $\mu = 2$ obtained by the Courant-Friedrichs-Lewy argument based on domains of dependence. If in addition we consider contractivity, then the trapezoidal rule is contractive only within the circle of radius two centred at (-2,0) and so A and B can only be integrated contractively for limited Courant numbers. However the midpoint rule cannot integrate any method in a contractive way.

The maximal values of μ with which we may integrate cases A and B by the Runge Kutta methods are given in Table 1 in Appendix C. We see immediately that efficiency is not at all comparable with that predicted by optimisation with respect to the eigenvalue curves of the finite matrices $B_{\Delta x}$. In Table 3 we give the maximal values of μ we might attain if integrating with the optimal schemes for the imaginary axis. Extension of stability along the imaginary axis almost certainly means that the stability regions will not extend far into $\mathbb{C}^-$. Thus, as expected, we can gain no improvement by integrating A and B by Runge Kutta schemes χ_1. Alternatively, for the two-step schemes, it may be better to use the Runge-Kutta schemes which have been optimised with respect to the wedge with angle 90°. However, at present these schemes have not been implemented. Hence in all cases we use the appropriate Runge Kutta as originally

intended. For comparison, the scaled Courant numbers are given in Table 2. Integration of A or B by the two-step, three-stage method is not worth considering, otherwise there is no general preference for one or two-step methods for A or B. However, for C and D there is a clear preference for the two-step schemes.

In the course of our experiments we have noticed that quite disastrous errors occur if we do integrate A and B out to the originally intended Courant numbers. Examples similar to this already exist in the literature ([Ve83], [Gr78]) and so we do not present our results here.

Before further discussion of the experiments it is interesting to consider what properties we might expect the SD's A to D to exhibit based on the analysis of the preceding section. Obviously any analysis applied without consideration of the time integration may be limited but should be valid in the limit as $\mu \to 0$. For low wavenumbers we may expand $\sin \xi h$ and $\cos \xi h$ by Taylor's theorem so that we can determine the order of dissipation of A and B. We find that, to lowest order,

$$\mathrm{Re}\,\hat{B}_1(\xi) \approx \frac{-h^3}{36}\xi^4 \quad , \quad \mathrm{Re}\,\hat{B}_2(\xi) \approx \frac{-h^3}{24}\xi^4 \tag{8.2.4}$$

and thus both schemes are dissipative of order four. Similarly the phase speeds for small wave numbers are given by

$$\begin{aligned} &c_1(\xi) \approx 1 + \frac{(\xi h)^4}{270} \quad , \quad c_2(\xi) \approx 1 + \frac{11}{720}(\xi h)^4 \\ &c_3(\xi) \approx 1 - \frac{(\xi h)^2}{6} + \frac{(\xi h)^4}{120} \text{ and } c_4(\xi) \approx 1 - \frac{(\xi h)^4}{30} \, . \end{aligned} \tag{8.2.5}$$

Therefore for A and B the leading error terms for phase and amplitude are both of order $p + 1$, whereas for C and D the phase errors are of order p. This agrees with the observation of Roe that, for p odd, the leading errors in phase and amplitude are of order $p + 1$, but for p even they are of order p and $p + 2$ respectively [Ro82]. For experiments performed under the same conditions, we would expect a noticeably larger phase error for the three-point central difference scheme as compared with the other methods.

Although we have already noted that the group velocity may not be particularly meaningful for cases A and B, we derive out of interest, expressions for the real part of the group velocity in all cases:

$$\mathrm{Re}\, C_1(\xi) = \frac{27 + 40\cos\xi h + 10\cos^2\xi h + 4\cos^3\xi h}{(5 + 4\cos\xi h)^2} \approx 1 - \frac{13}{162}(\xi h)^4 .$$

$$\mathrm{Re}\, C_2(\xi) = \frac{137 + 105\cos\xi h - 45\cos^2\xi h + 19\cos^3\xi h}{(25 + 16\cos\xi h - 5\cos^2\xi h)^2} \approx 1 + \frac{1}{16}(\xi h)^4 .$$

$$C_3(\xi) = \cos\xi h \approx 1 - \frac{(\xi h)^2}{2} + \frac{(\xi h)^4}{24} . \qquad (8.2.6)$$

$$C_4(\xi) = \frac{1}{3}[1 + 4\cos\xi h - 2\cos^2\xi h] \approx 1 - \frac{(\xi h)^4}{6} .$$

Clearly, a wave packet should move more quickly with the fourth-order method D rather than the second-order method C.

Further, it is possible to consider integration of C and D by the midpoint and trapezoidal rules as fully discrete schemes and derive their dispersion relations exactly. Therefore, on occasion we will refer to the results of Trefethen, who has analysed the fully discrete model of three of these schemes [Tr82]. With the two sets of figures for SD and FD schemes the influence of the time integration on group velocity may then be investigated.

In view of the complex nature of group velocity for the dissipative schemes the extent to which its imaginary part affects the speed of dissipation is also interesting. Therefore, we give expressions for this part of the group velocity which are as follows:

$$\mathrm{Im}\, C_1(\xi) = \frac{2 \sin \xi h (\cos \xi h - 1) (7 + 2 \cos \xi h)}{(5 + 4 \cos \xi h)^2} \approx - \frac{(\xi h)^3}{9} + O(\xi h)^5$$

$$\mathrm{Im}\, C_2(\xi) = \frac{12 \sin \xi h (\cos \xi h - 1) (9 + 26 \cos \xi h - 5 \cos^2 \xi h)}{(25 + 16 \cos \xi h - 5 \cos^2 \xi h)^2}$$

$$\approx - \frac{5}{36} (\xi h)^3 + O(\xi h)^5. \qquad (8.2.7)$$

Consequently, if this part of the group velocity does affect the dissipation we would expect its influence to be similar in each case.

All of the group velocity considerations discussed here are investigated in the following section by analysing the evolution of some carefully chosen initial conditions.

8.3 Experiments and Results

We recall that the solution of nonlinear hyperbolic partial differential equations can be considered as two separate problems. As the solution evolves in time, it passes from a transient stage where very violent changes are occurring, to a steady-state or asymptotic stage. In the transient phase, models are usually designed to take account of the direction in which information is moving. Some kind of upwinding approach may be used or a series of Riemann problems solved. The concentration is on a realistic solution recognising shocks and other discontinuities rather than great efficiency by fast time integration. Indeed, increased efficiency in time would ignore the rapid change occurring and therefore is not employed. The models we have considered are not suitable for solutions in this phase of the evolution. However, when steady state has been reached we are interested in integrating rather more quickly. We have already explained why the Runge Kutta schemes attain prominence in this context.

In the asymptotic state the solution of the nonlinear problem is likely to consist of regions in which the solution is smooth, interspersed with various discontinuities which are likely to dissipate. Therefore it is important that the numerical method employed to continue the solution should resolve these discontinuities correctly and evolve them in time without smoothing or expanding them. Inherent

in any numerical model are errors due to the inability of any solution to correctly model a discontinuity on a finite grid. These errors usually exhibit themselves as trains of oscillations in front or behind the discontinuity. Thus it is crucial that the model should damp out such spurious oscillations as quickly as possible to prevent interactions, and consequent development of non-physical shocks.

In our experiments we have not concentrated on evaluating the performance of our methods by solving any nonlinear problems. Rather, we have considered some initial conditions which might be present in the asymptotic stage of the solution and investigated their time evolution. All our experiments have been performed in the setting as suggested by Trefethen so that we may, on occasion, make comparison [Tr82]. The wavenumber is chosen so that there are eight grid points per wavelength, $\xi h = \frac{2\pi}{8}$ and we integrate with Courant number $\mu = 0.4$.

The first problem we consider is the evolution of step-function data which to some extent represents a shock-wave problem. Some of our results for the propagation of the data defined on the interval [0,3]

$$u(x,0) = \begin{cases} 0 & : \ |x-2| > \frac{1}{2} \\ 1 & : \ |x-2| \leq \frac{1}{2} \end{cases}$$

are presented in Figures 4 and 5 where the integration is taken up to time t = 1. Notice that the width of the step is more spread out for solutions based on the three-point central difference scheme. This reflects the lower order of this scheme.

TIME INTEGRATION

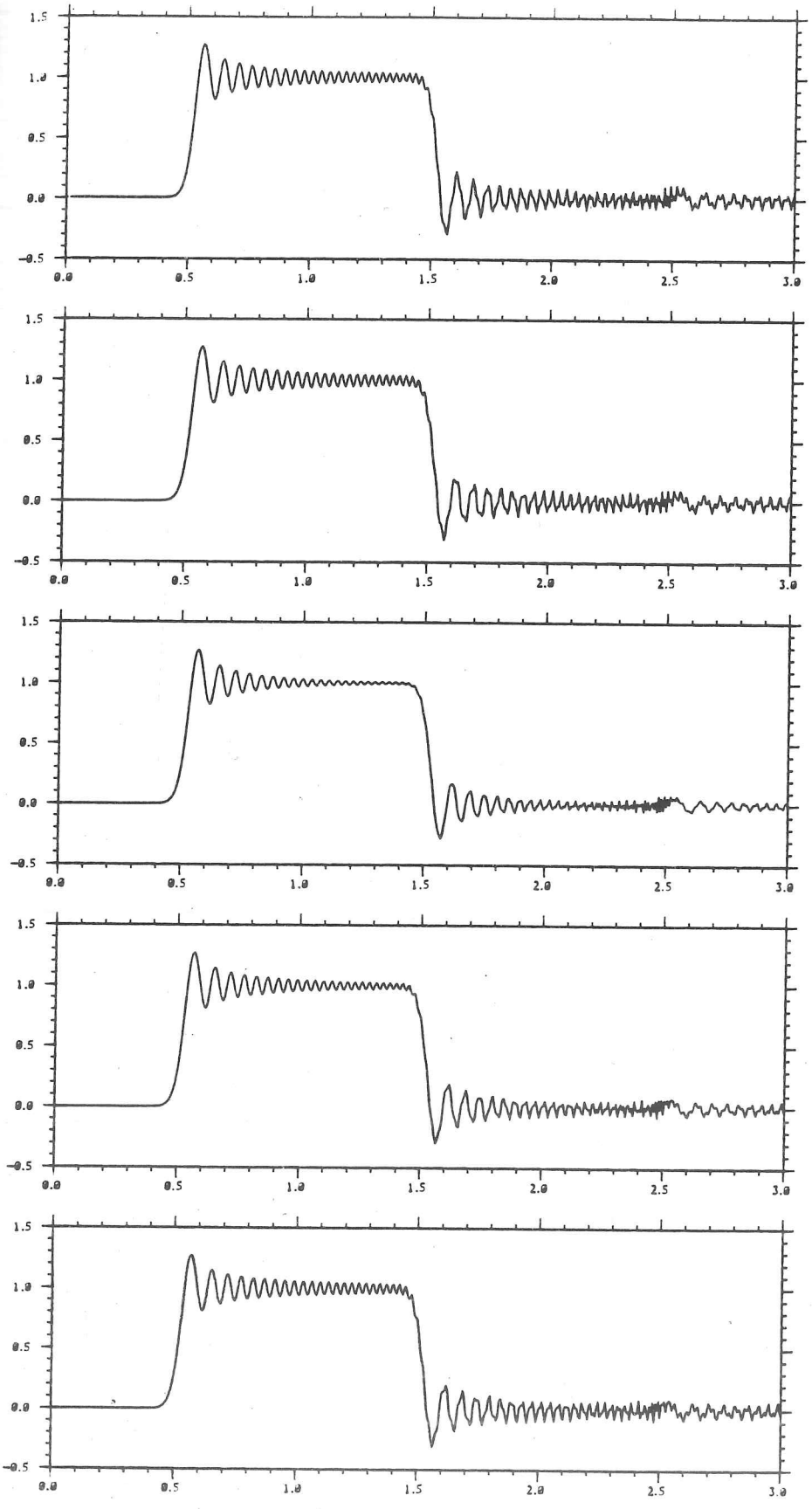
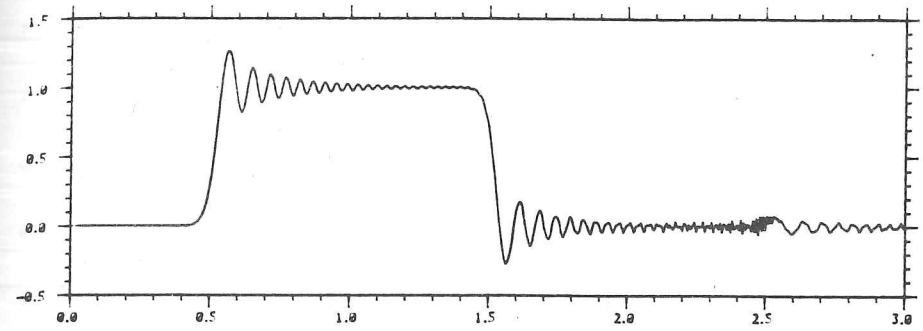
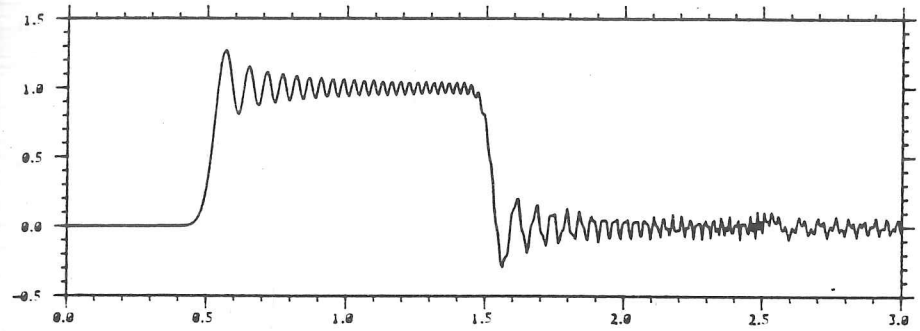


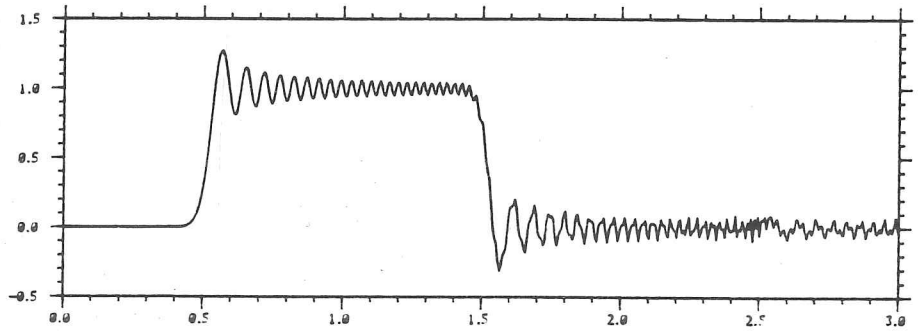
Figure 4a. Propagation of step function by three-point central difference semi-discretisation with $\mu = .4$, $h = 1/160$, $\xi h \simeq .79$.



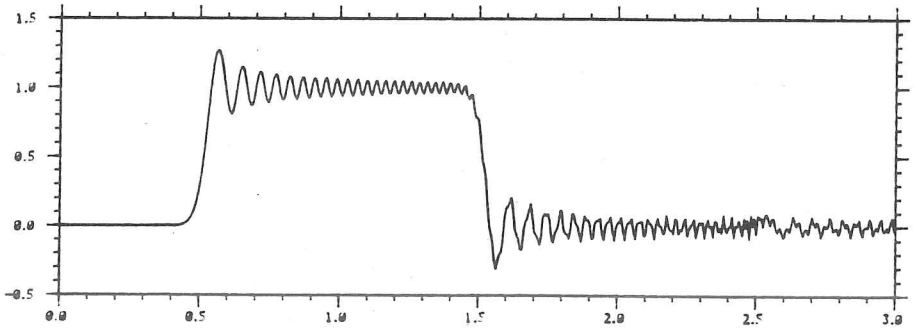
Runge Kutta
Two step
Two stage
Order three



Runge Kutta
Two step
Two stage
Order two



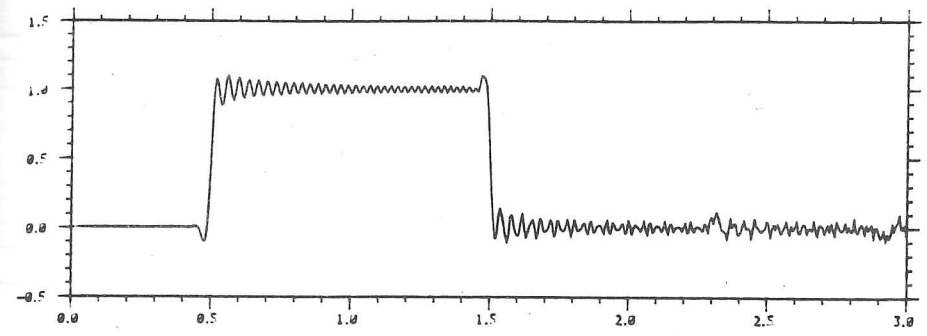
Runge Kutta
Two step
Three stage
Order three



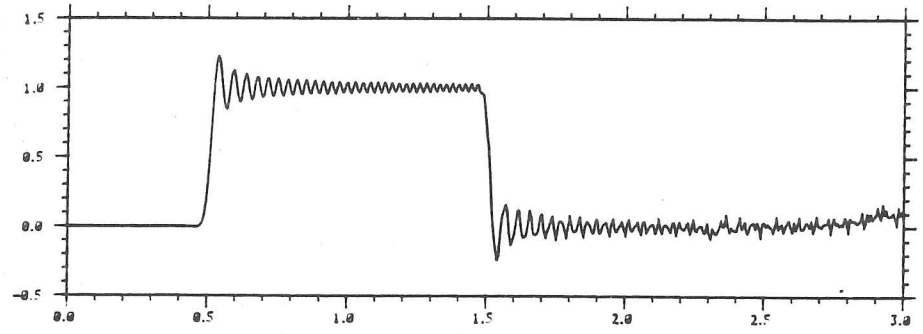
Runge Kutta
Two step
Three stage
Order two

Figure 4a. (continued)

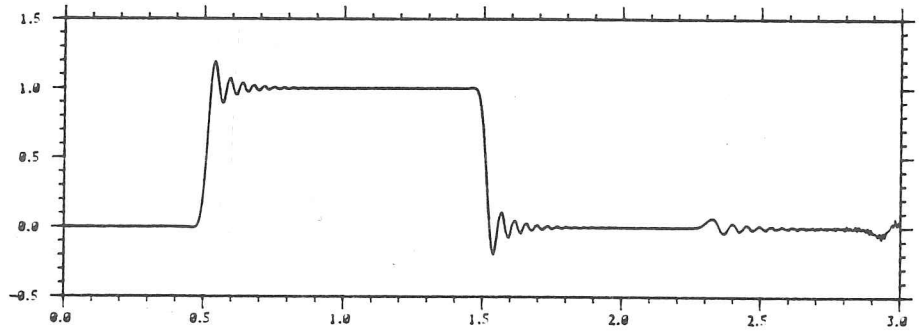
TIME INTEGRATION



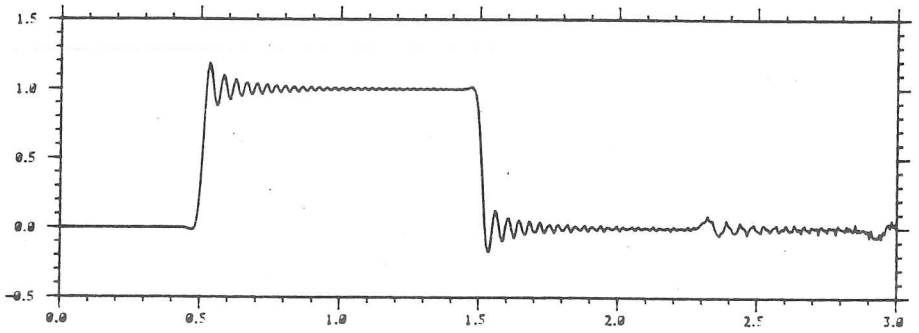
Midpoint



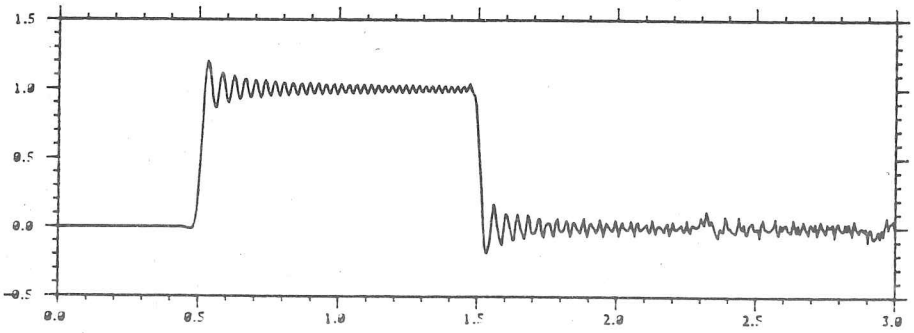
Trapezoidal



Runge Kutta
One step
Three stage
Order two



Runge Kutta
One step
Three stage
Order three



Runge Kutta
One step
Four stage
Order three

Figure 4b. Propagation of step function by five-point central difference semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq .79$.

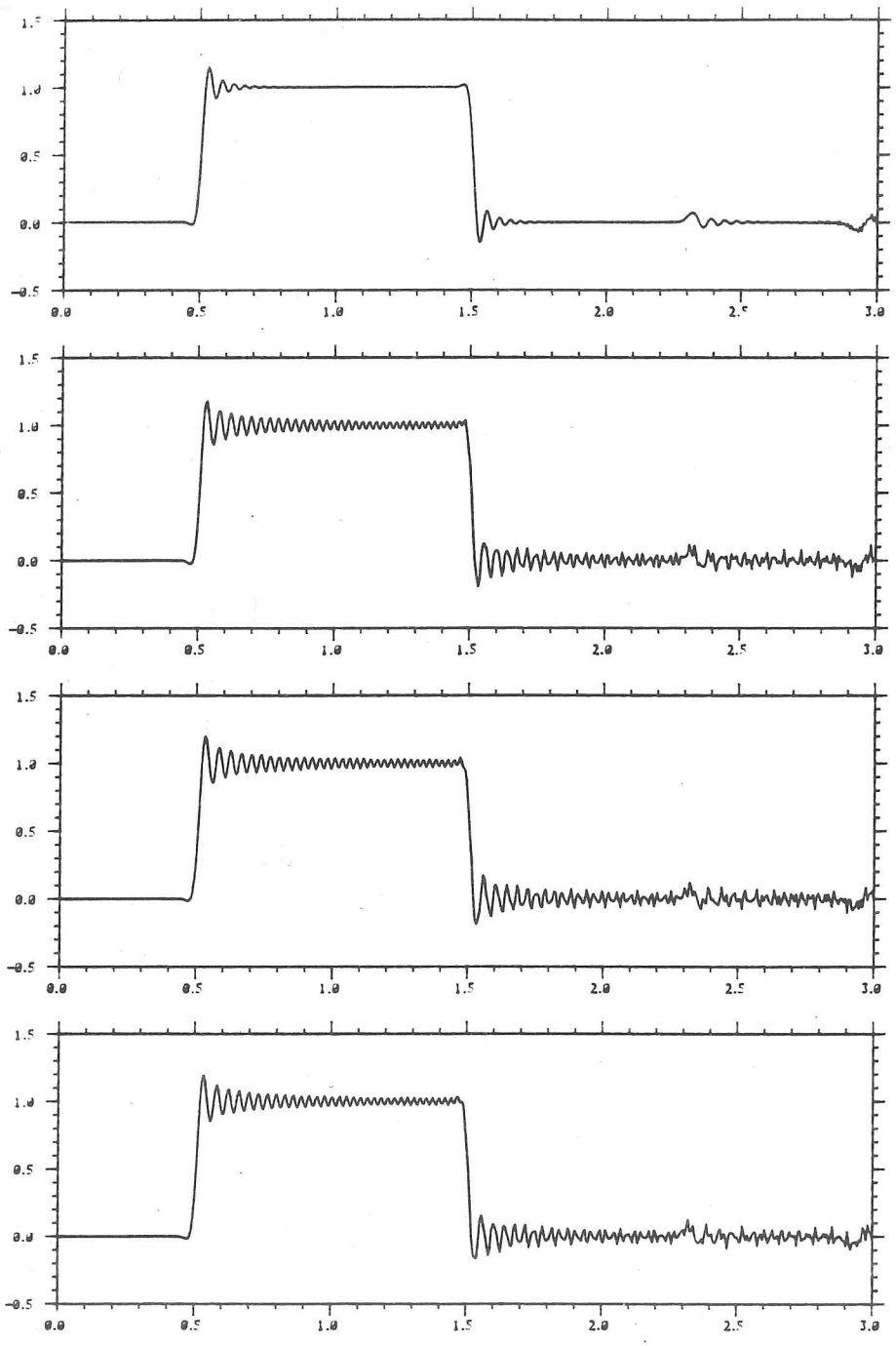


Figure 4b. (continued)

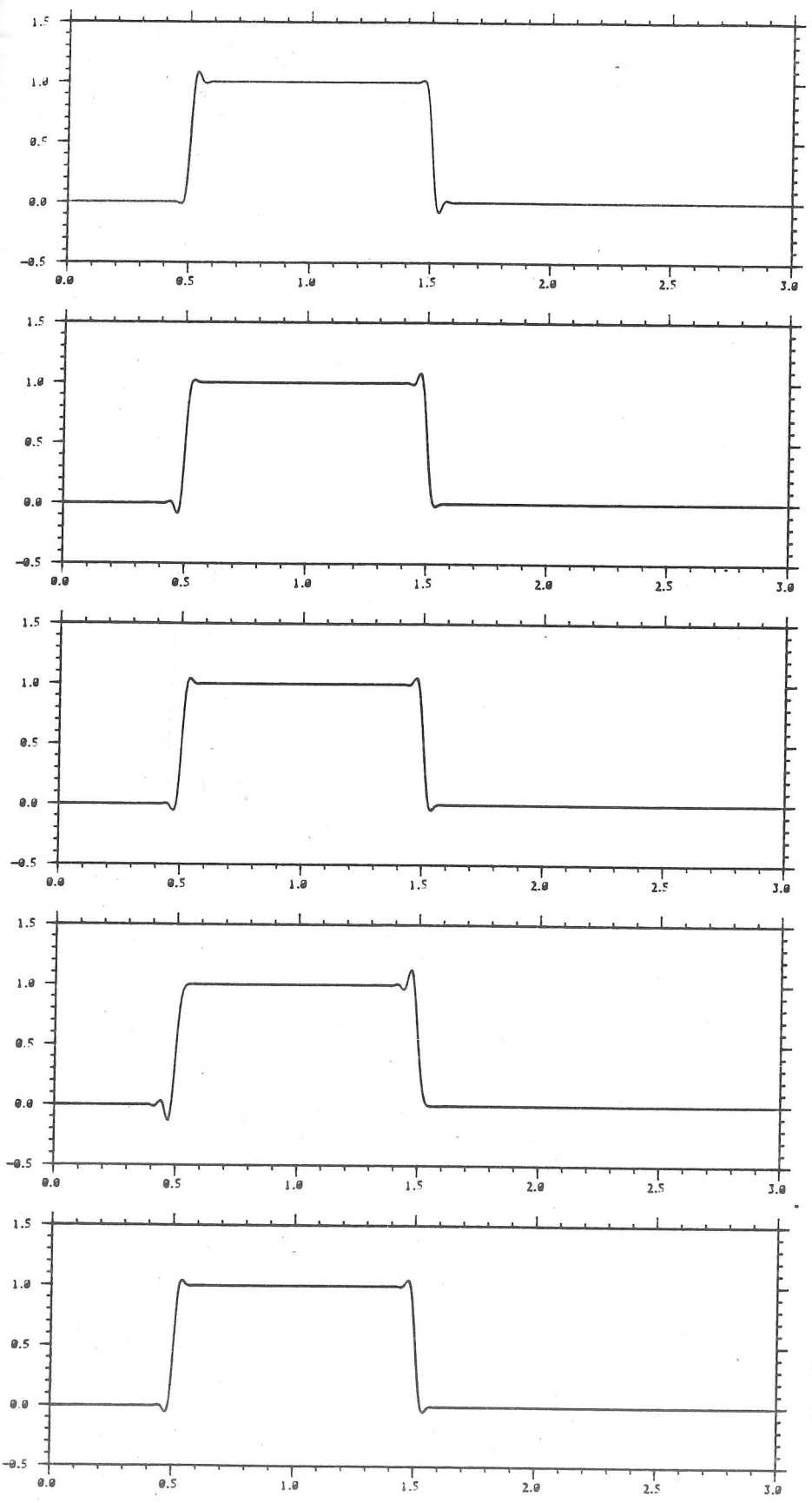


Figure 5a. Propagation of stepfunction by case A semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq .79$.

TIME INTEGRATION

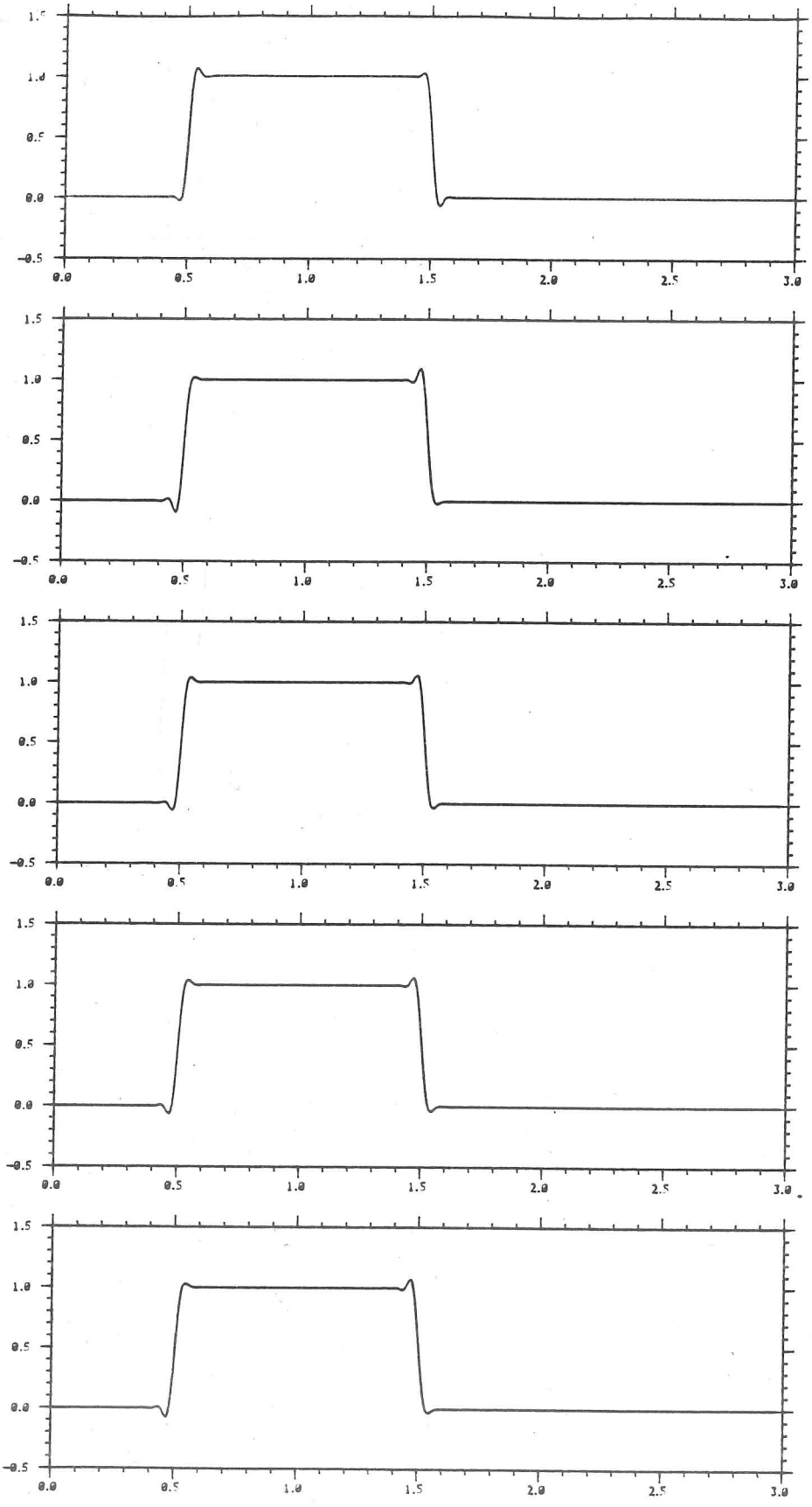


Figure 5b. Propagation of stepfunction by case B semi-discretisation, with $\mu = 4$, $h = 1/160$, $\xi h \simeq .79$ (except Runge Kutta two-step two-stage order three, $\mu = .35$).

For the conservative schemes there is a train of oscillations behind the discontinuity which is consistent with group velocity being less than one. Also, there is either overshoot or smearing ahead of the motion which cannot be totally attributable to group velocity but is instead a measure of the added dissipation of the O.D.E. solver. There seems to be some preference for both the one-step three-stage Runge Kutta, and for the two-step two-stage Runge Kutta of order three. The two-step method has a zero-stability parameter p_0 with modulus away from 1 so that one root of the zero-stability polynomial is well inside the unit circle. Hence, it is likely that schemes for which this is so possess more dissipation: certainly this scheme does have a stability region which extends well into $\mathbb{C}^-$. Also, the second-order one-step method involves very few operations for case χ_1 as our methods have been implemented in the most obvious ways rather than by one of the efficient algorithms of Chapter 7. Thus its superiority despite its lower-order accuracy, suggests that algorithms designed to reduce operations, as in Chapter 7, will be useful in practice for reducing spurious oscillations. Notice also the evidence of modes with higher wave numbers travelling slower and thus lying further back in the wave train.

The dissipative cases A and B with integration by χ_1 and χ_2 produce very good shock resolution with small overshoot ahead of the motion and an even smaller wiggle behind the wave. The situation is reversed for integration by the implicit trapezoidal rule. Thus the strongly

dissipative nature of cases A and B as predicted in the previous section is very much in evidence as oscillations have been very quickly damped out. We can also see the affect of integrating by explicit methods of different order. For the odd-order methods, the wiggles before and after the discontinuity are more balanced. The even-order Runge Kutta have much larger overshoot ahead of the motion. This indicates that the order-two methods will be slower to converge and therefore integrating with third order is preferable.

These stepfunction experiments certainly support the hypothesis that at least for small timesteps, the character of the semidiscretisation dominates the solution. To see how well the expressions for group velocity and amplitude can be used for prediction, we now consider the evolution of two more initial conditions.

Group speed is observed most simply by looking at a wave packet. The initial packet is a sine wave modulated by a Gaussian centred at $x = 2$,

$$u(x,0) = e^{-16(x-2)^2} \sin \xi x \quad , \quad \xi = \frac{\pi}{4h} .$$

Examples of the propagation of this monochromatic signal are given in Figures 6 and 7.

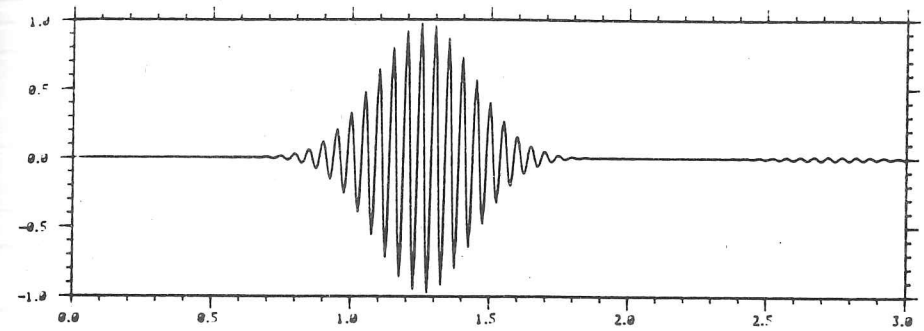
dissipative nature of cases A and B as predicted in the previous section is very much in evidence as oscillations have been very quickly damped out. We can also see the affect of integrating by explicit methods of different order. For the odd-order methods, the wiggles before and after the discontinuity are more balanced. The even-order Runge Kutta have much larger overshoot ahead of the motion. This indicates that the order-two methods will be slower to converge and therefore integrating with third order is preferable.

These stepfunction experiments certainly support the hypothesis that at least for small timesteps, the character of the semidiscretisation dominates the solution. To see how well the expressions for group velocity and amplitude can be used for prediction, we now consider the evolution of two more initial conditions.

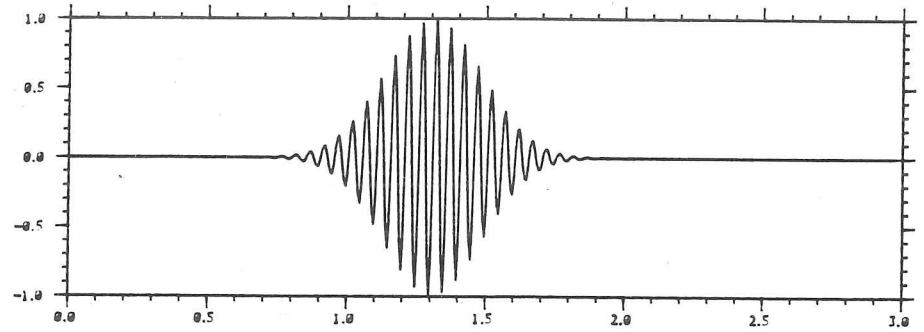
Group speed is observed most simply by looking at a wave packet. The initial packet is a sine wave modulated by a Gaussian centred at $x = 2$,

$$u(x,0) = e^{-16(x-2)^2} \sin \xi x \quad , \quad \xi = \frac{\pi}{4h} .$$

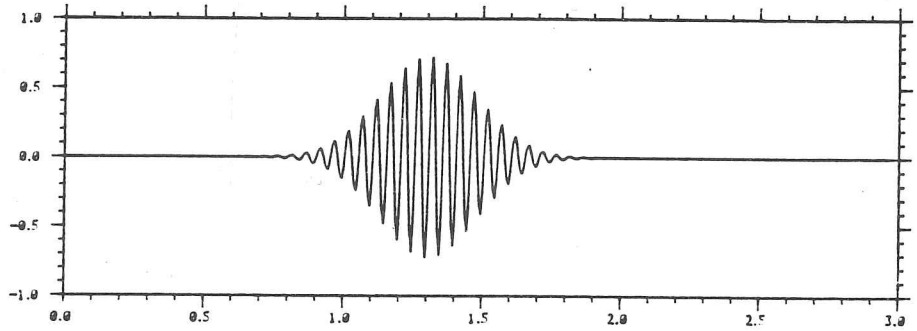
Examples of the propagation of this monochromatic signal are given in Figures 6 and 7.



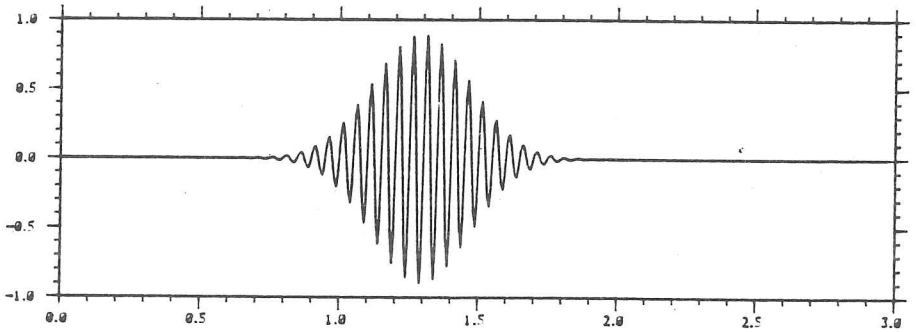
Midpoint



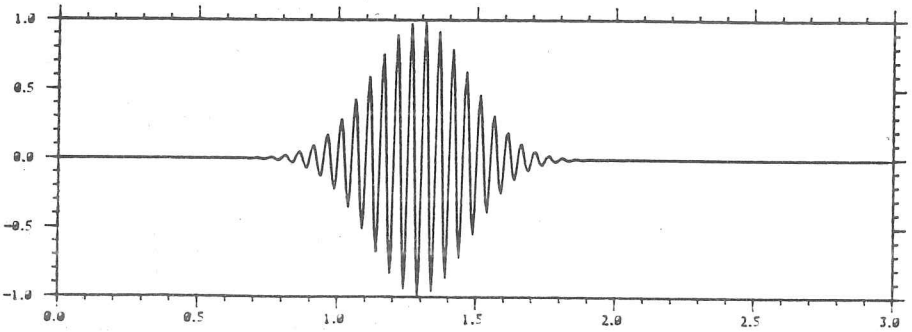
Trapezoidal



Runge Kutta
One step
Three stage
Order two



Runge Kutta
One step
Three stage
Order three



Runge Kutta
One step
Four stage
Order three

Figure 6a. Propagation of monochromatic wavepacket by three-point central difference semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq 1/160$.

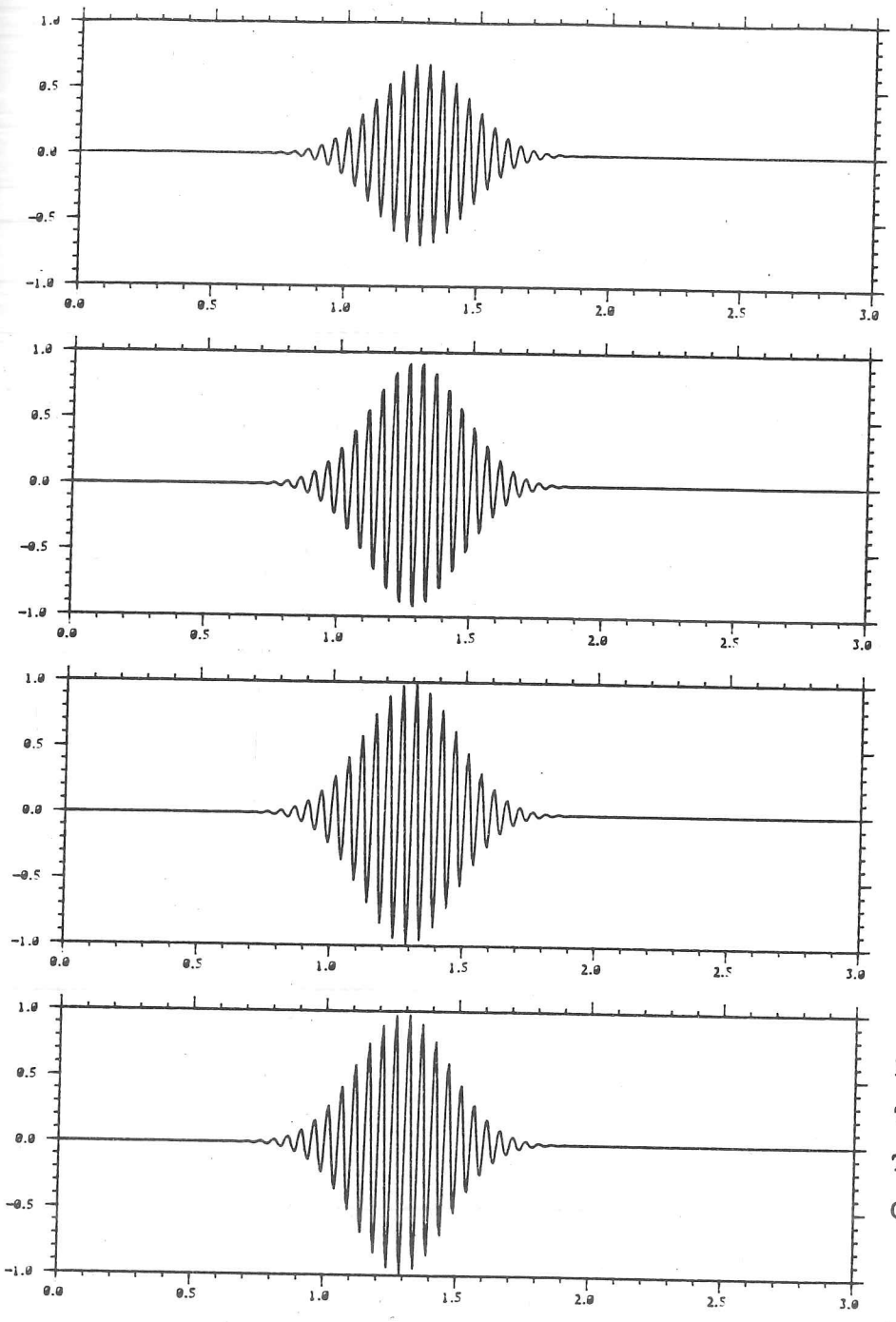


Figure 6a. (continued)

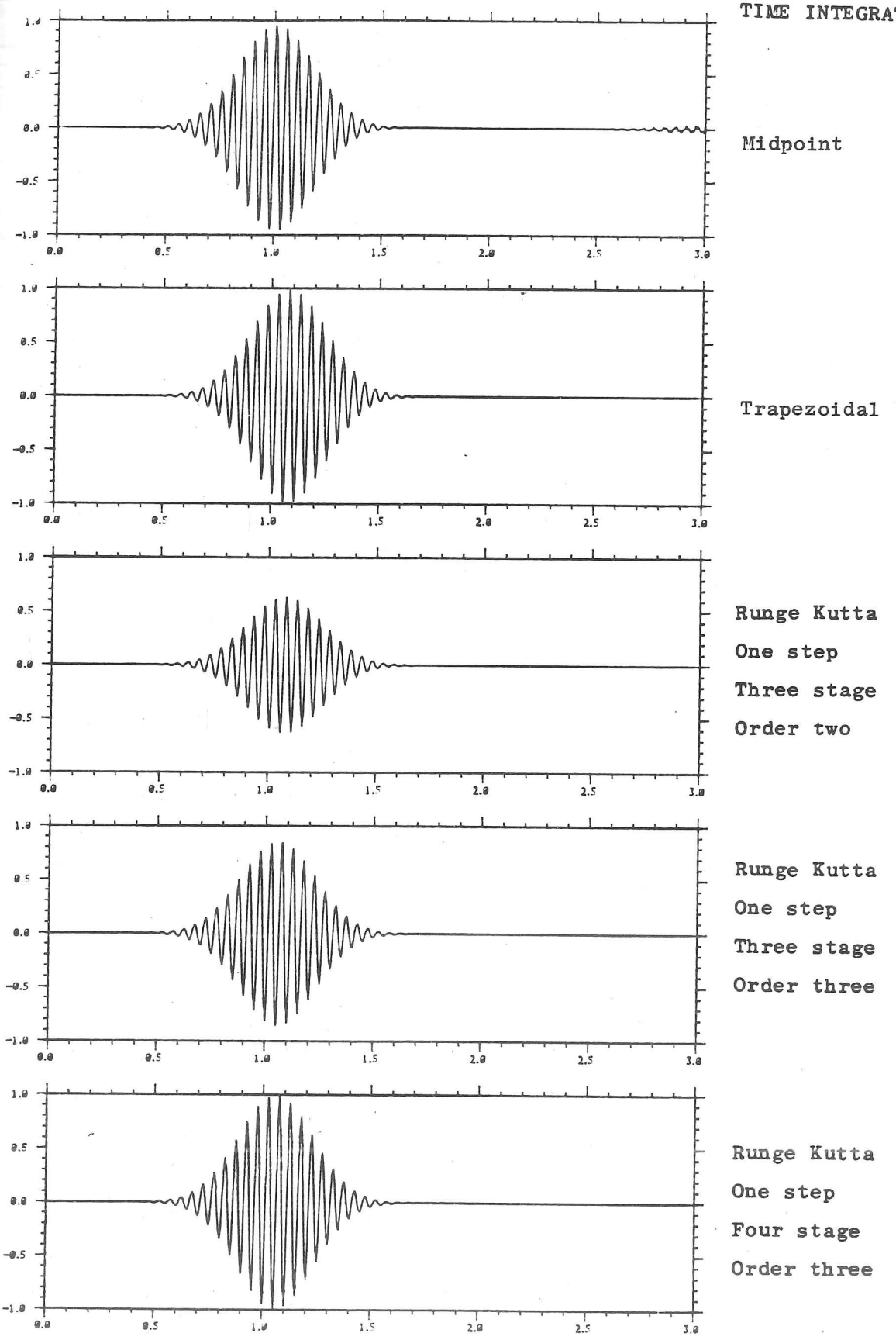


Figure 6b. Propagation of monochromatic wavepacket by five-point central difference semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq 1/160$.

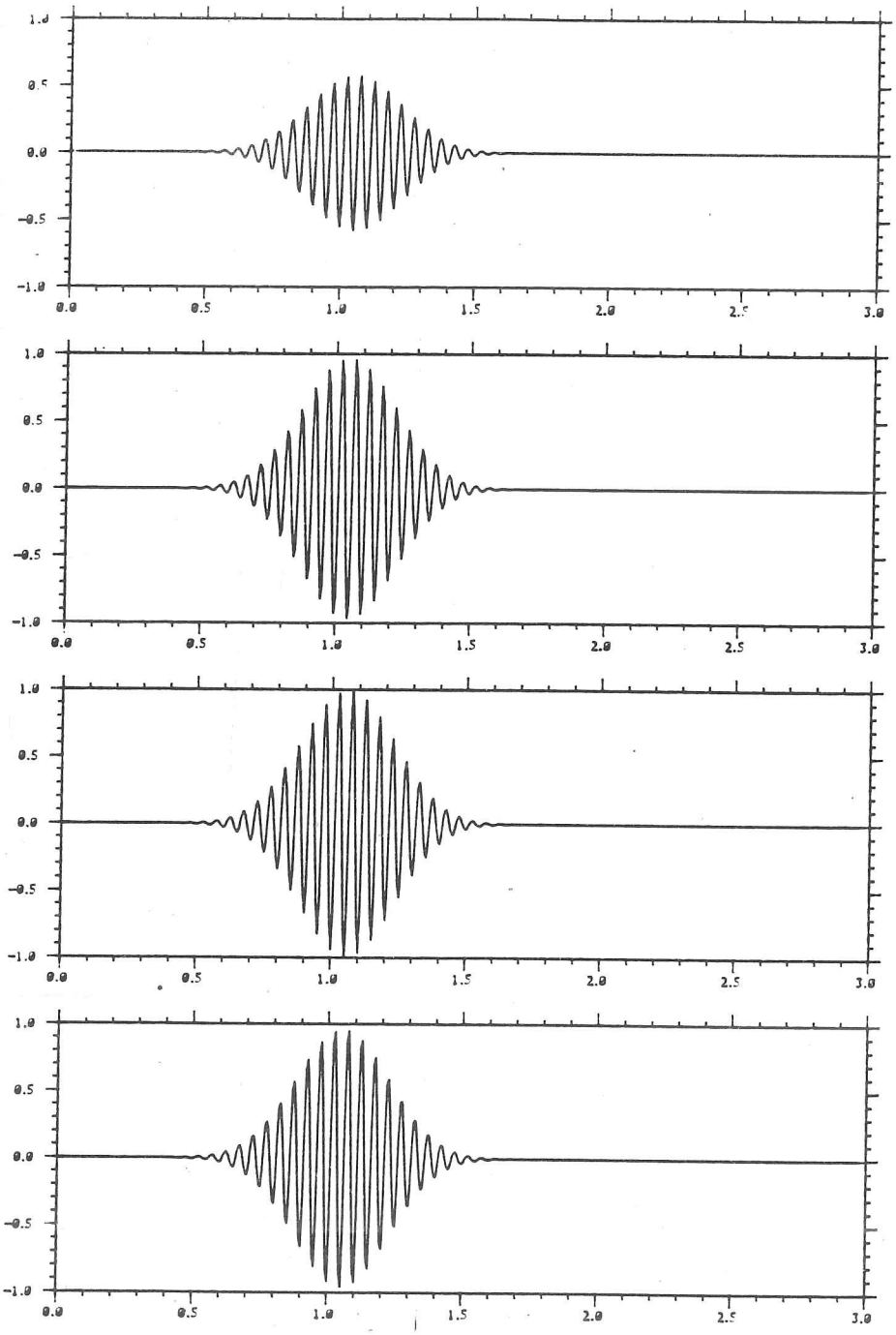


Figure 6b. (continued)

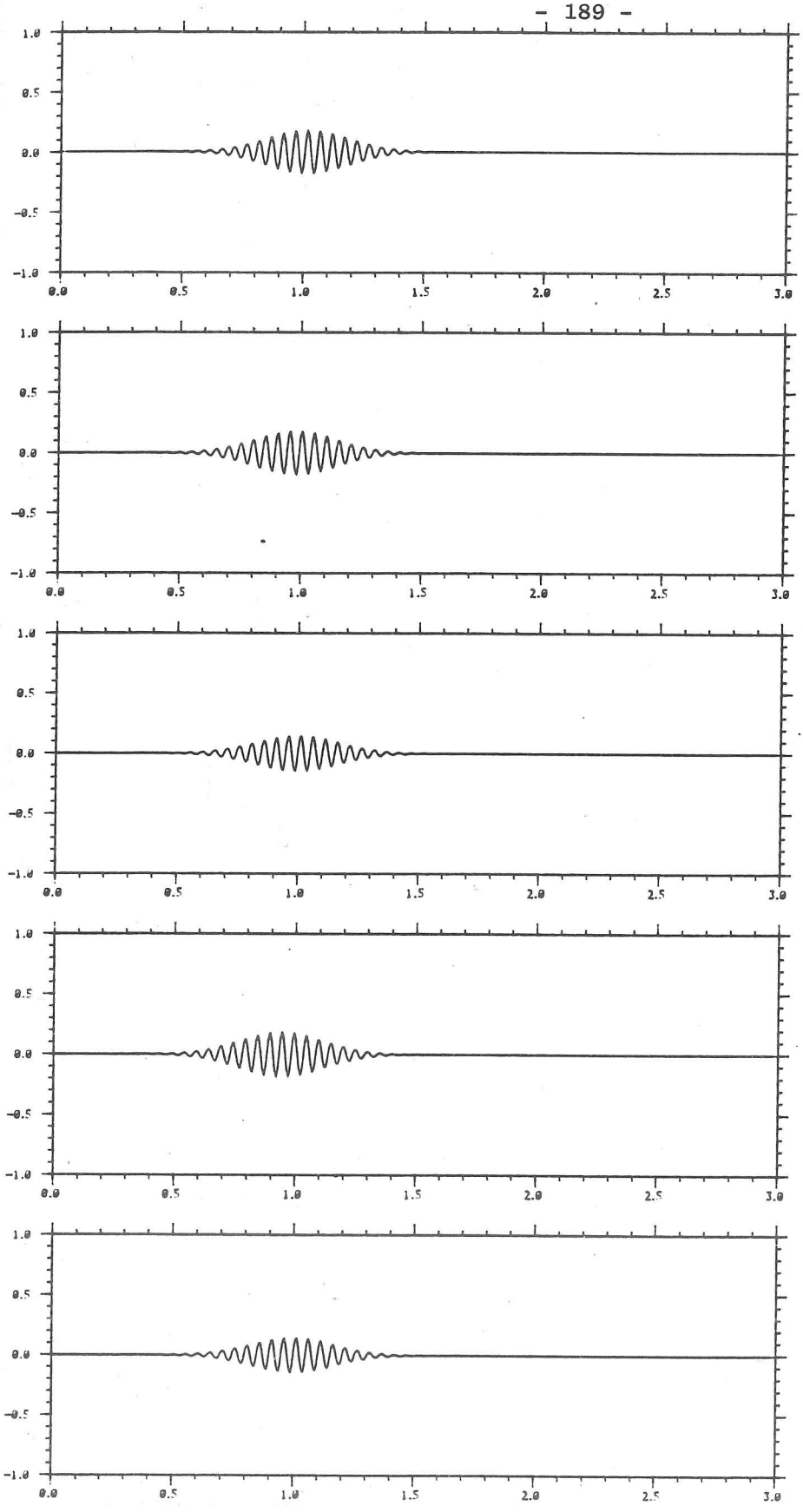


Figure 7a. Propagation of monochromatic wavepacket by case A semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq 1/160$.

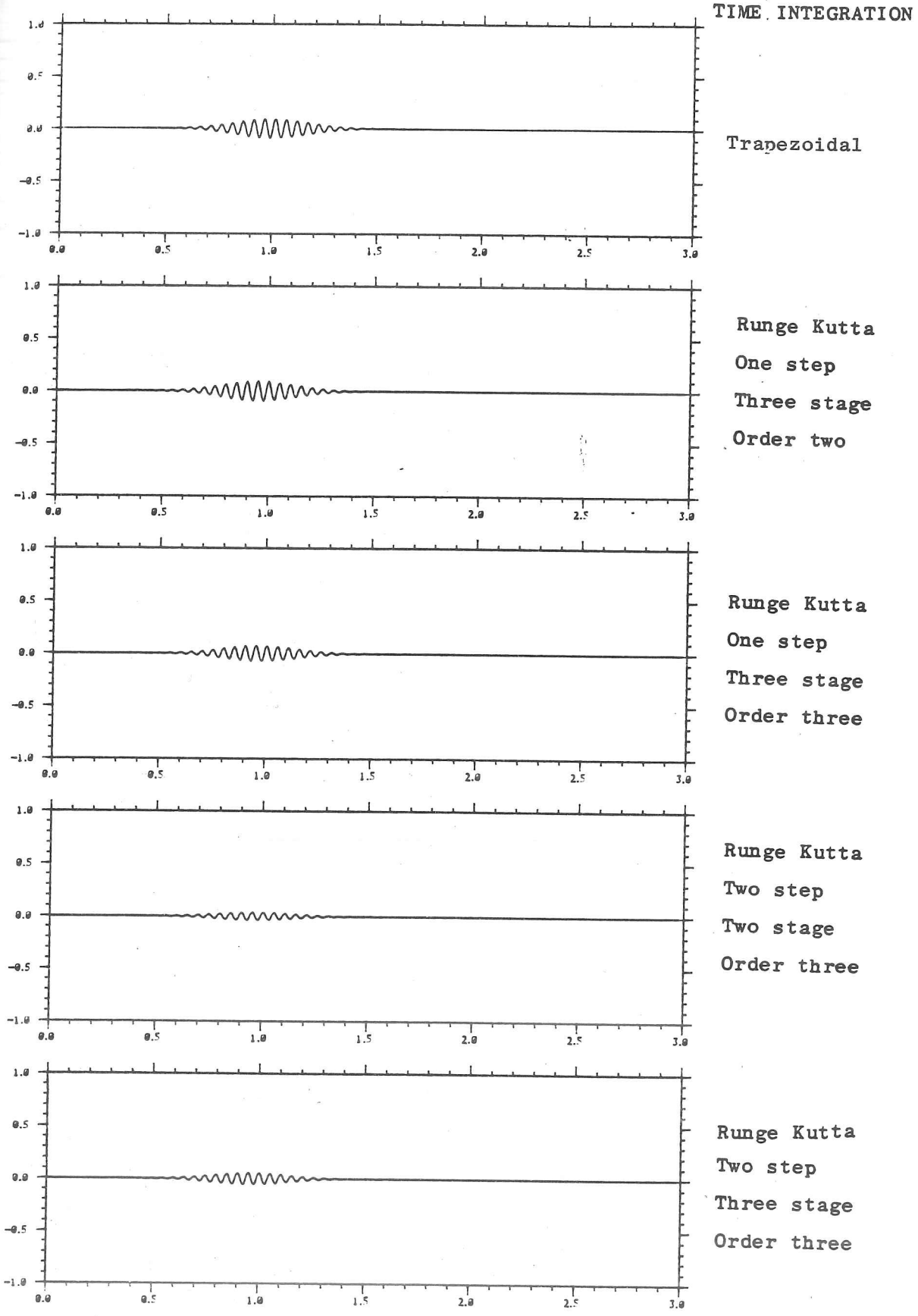


Figure 7b. Propagation of monochromatic wavepacket by case B semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq 1/160$ (except Runge Kutta two-step, two-stage order three, $\mu = .35$).

	mid-point	trape-zoidal	Runge Kutta			
			k = 2	p = 3	k = 1	m = 3
			m = 3	m = 2	p = 2	p = 3
Three point	1.27	1.30	1.29	1.29	1.29	1.30
central difference	.99	1.00	1.00	.70	.72	.90
Five point	1.01	1.06	1.06	1.06	1.08	1.05
central difference	.98	.99	.96	.58	.62	.80
Case A		1.01	1.00	.96	.98	1.00
		.18	.15	.18	.18	.15
Case B		.99	.95	.98	.95	.96
		.09	.06	.04	.10	.08

Table 8.3.1 Wave packet results.

The first number for each case approximates the centre of the wave packet, and the second its maximum amplitude.

A summary of the positions and amplitudes of the signals after time $t = 1$ is given in Table 8.3.1. Obviously the choice of time integration does influence the propagation. As expected from the stepfunction results, the Runge Kutta methods previously discussed do add dissipation for cases C and D. For cases A and B, Equations (8.2.4) would predict that the amplitude is damped to .18 and .08 respectively. The experiments compare very favourably with these values. Moreover, Equations 8.2.7 predict a dissipation rate per unit time step of .07 and .09 respectively suggesting that group velocity is not relevant. Obviously there is again some additional dissipation due to the Runge-Kutta schemes.

The phase speeds and group speeds as given by (8.2.5) and (8.2.6), for $\xi h = \frac{\pi}{4}$, are approximately

$$C_1(\xi) = 1.00 \quad C_2(\xi) = 1.01 \quad C_3(\xi) = .90 \quad C_4(\xi) = .87$$
$$C_1(\xi) = .97 \quad C_2(\xi) = 1.02 \quad C_3(\xi) = .71 \quad C_4(\xi) = .94$$

It seems that the dissipative schemes have advanced faster than would be predicted by the group speed. This indicates that dissipation is dominating dispersion in which case the group velocity analysis is not sufficient. Additional analysis of energy velocity might completely describe the solution. However, there is reasonable agreement with the predicted values for the conservative cases C and D. In all cases adding dissipation by time integration has marginally affected the group speed. This is particularly noticeable for the five-point central difference formula. It is not completely clear why the influence of the time integration is not the same throughout. However, as the three point difference formula is far less affected by the time integration it is probable that the ability of group speed to be influenced is determined by the order of the variation of the speed from one. For case A this variation is only second order and thus variations of higher order due to higher order integration formulae will be less noticeable. Elsewhere the variation of the speed from one is fourth order and thus more liable to be affected by time integration.

It is interesting to remark that the expressions for group speed of the fully discrete models as given by

Trefethen yield the values

midpoint + C $C(\xi) = .74$

midpoint + D $C(\xi) = 1.05$

trapezoidal + C $C(\xi) = .67$

trapezoidal + D $C(\xi) = .91$

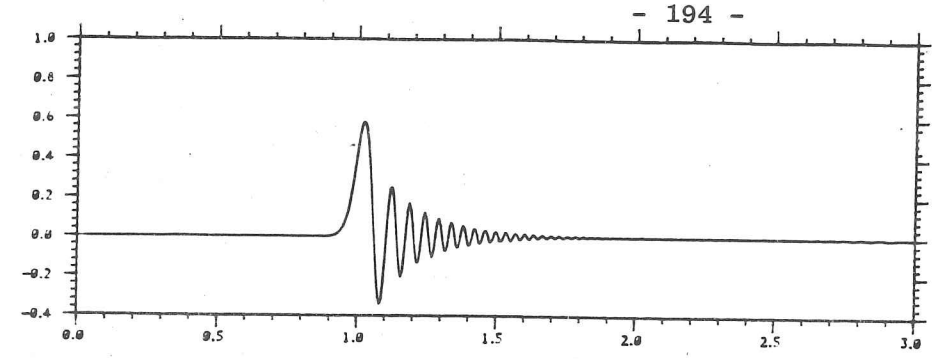
(cf. [Tr82]).

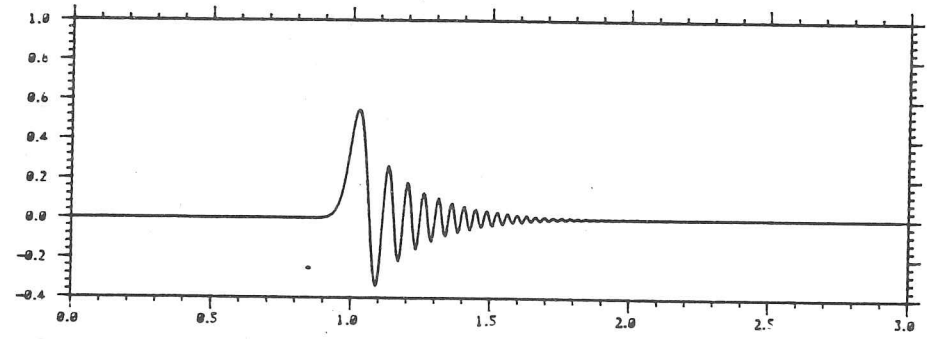
These values are not actually any more accurate than those predicted by the semidiscretisation alone. Therefore it seems quite realistic to use the expressions for group speed as predicted by the spatial semi-discretisation as a guide for all conservative methods.

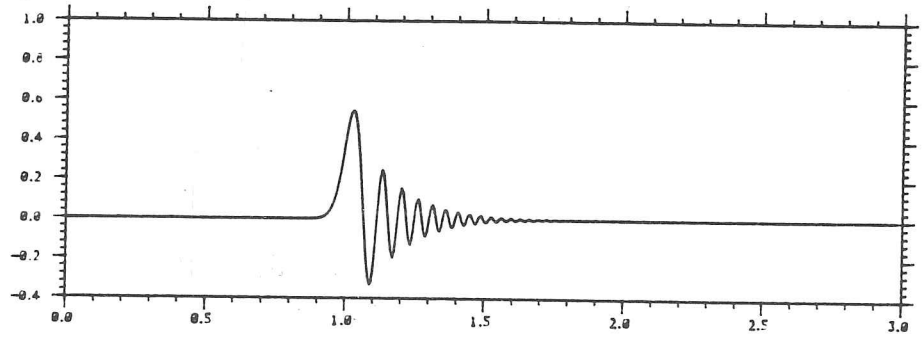
Clearly the wave packets are dispersed to a certain extent. However, to measure this dispersion it is far easier to look at the propagation of a polychromatic signal defined by

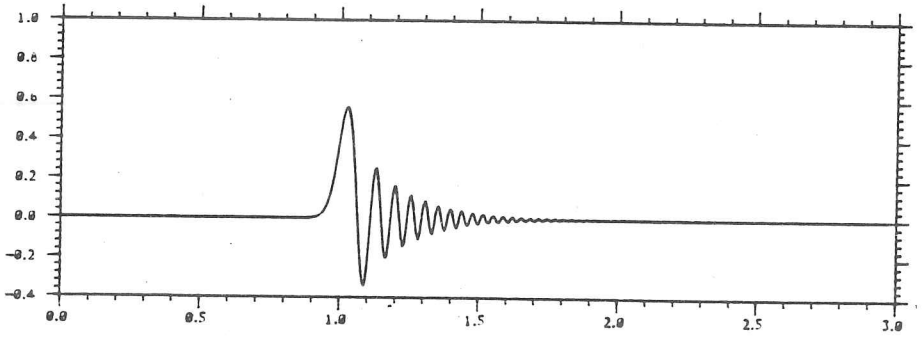
$$u(x,0) = e^{-3200(x-2)^2} .$$

This signal is much narrower than the envelope of the wave packet and it has a central wavenumber $\xi = 0$. The Fourier transform of a wave packet was a narrow spike, whereas a pulse has a transform which is broad. Therefore it can include wavenumbers with greatly varying group speeds and dispersion of these Fourier modes should be evident.









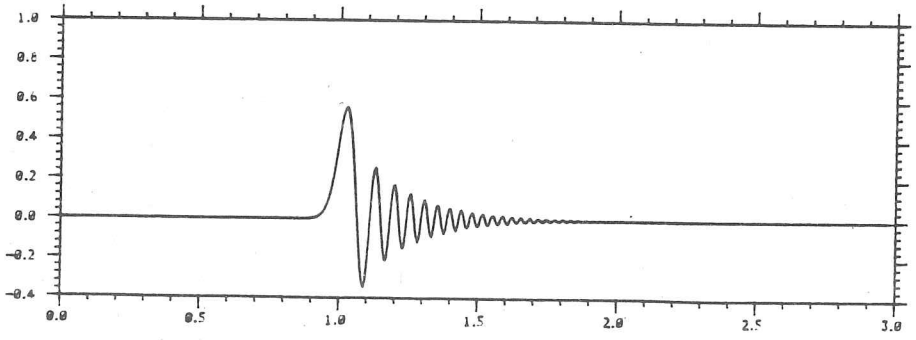


Figure 8a. Propagation of polychromatic pulse by three-point central difference semi-discretisation, with $\mu = 4$, $h = 1/160$, $\xi h \simeq .79$.

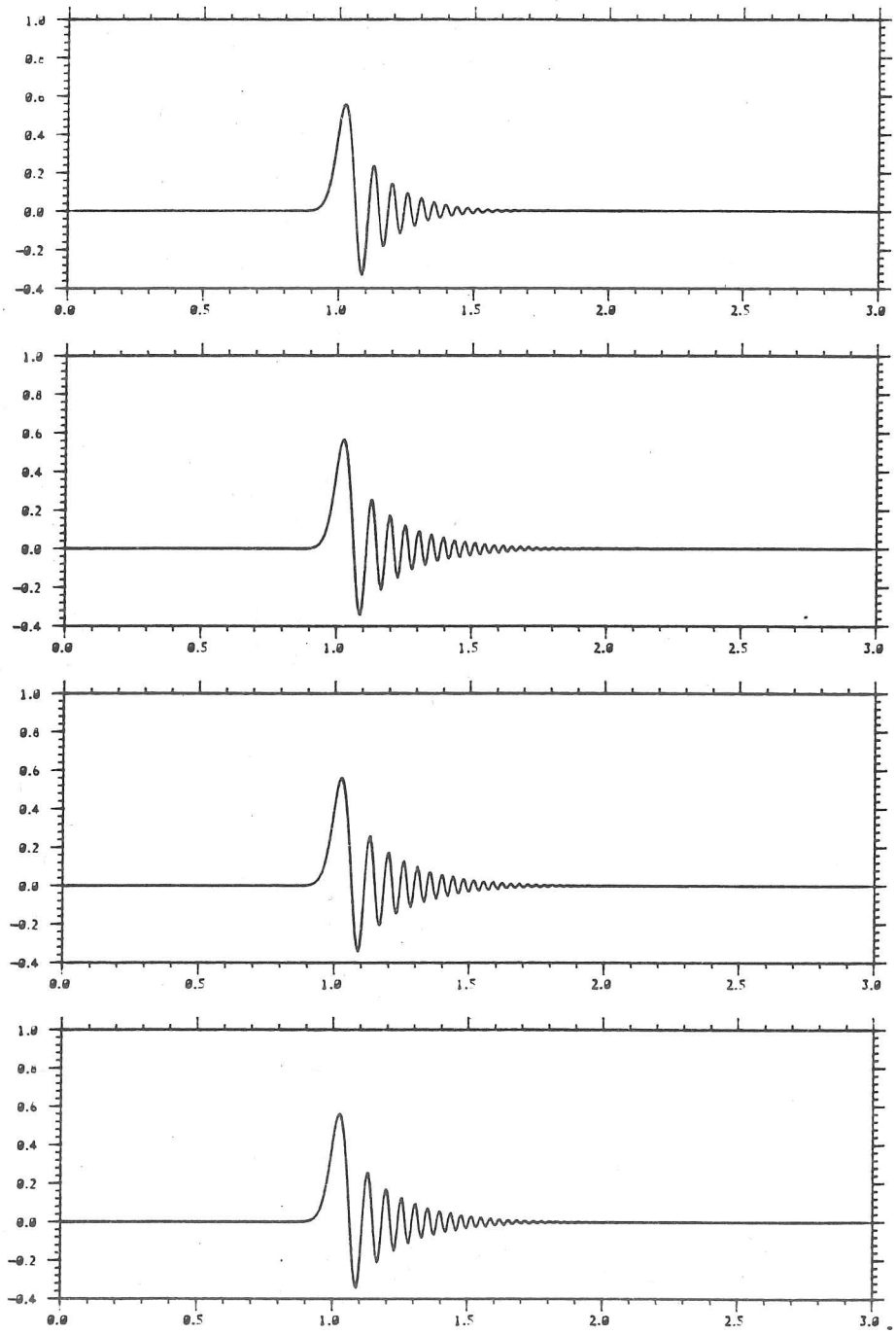
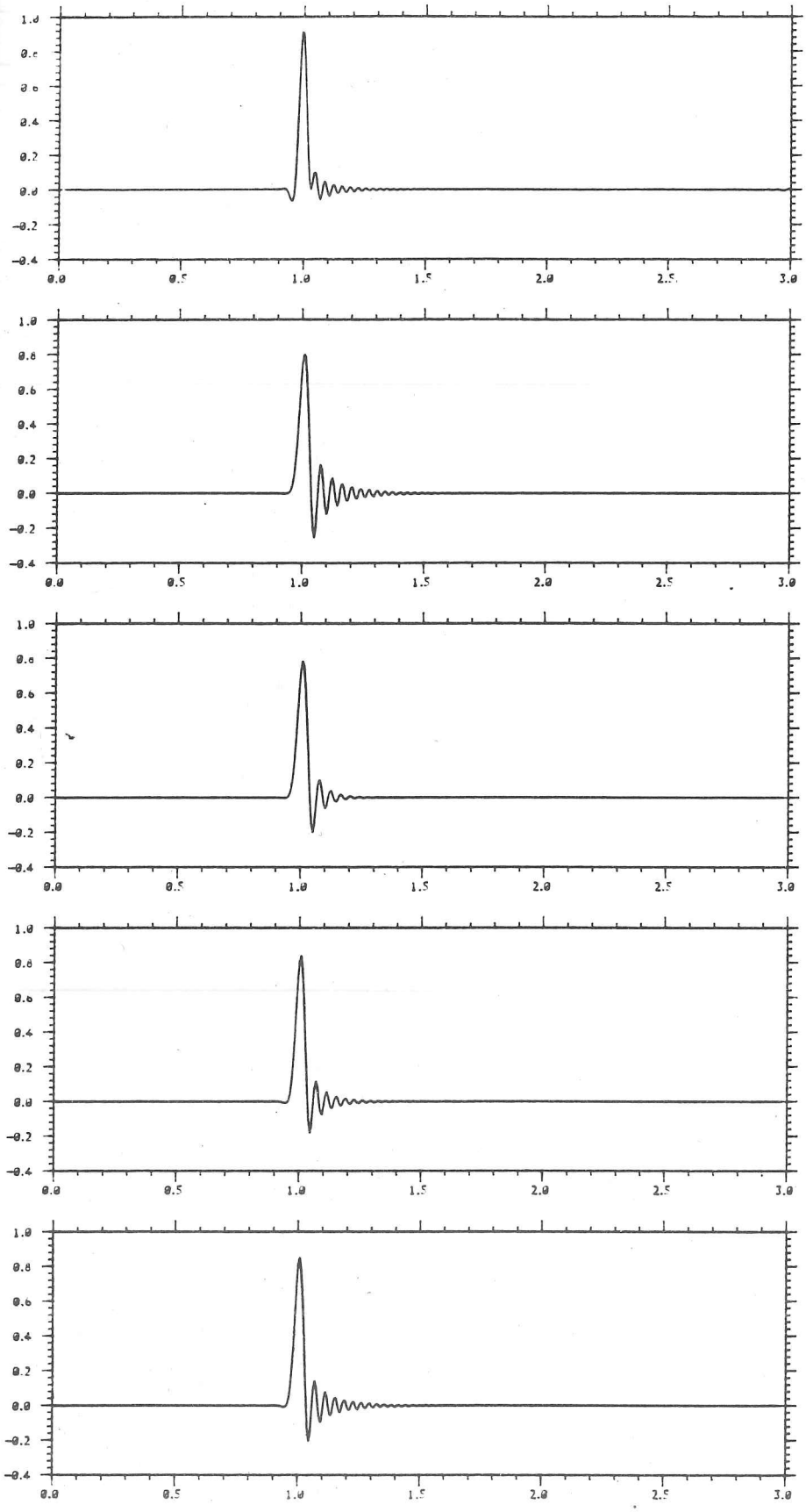


Figure 8a. (continued)

TIME INTEGRATION



Midpoint

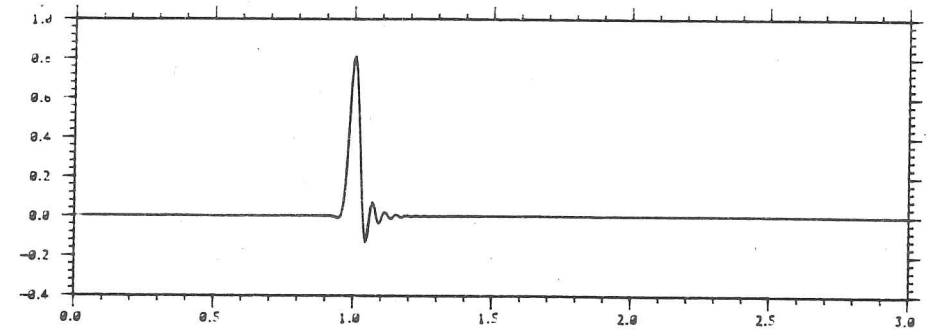
Trapezoidal

Runge Kutta
One step
Three stage
Order two

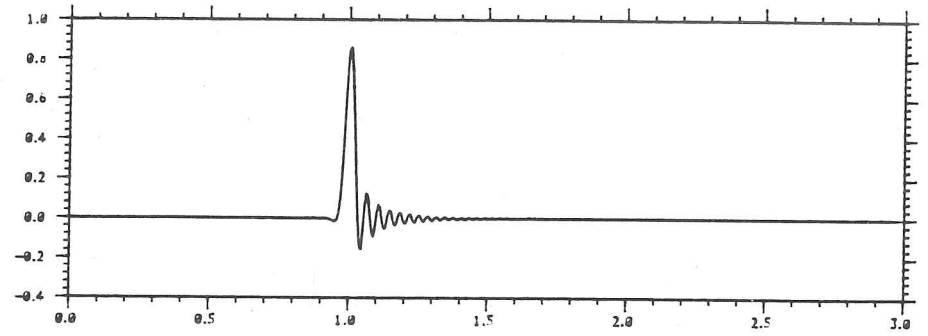
Runge Kutta
One step
Three stage
Order three

Runge Kutta
One step
Four stage
Order three

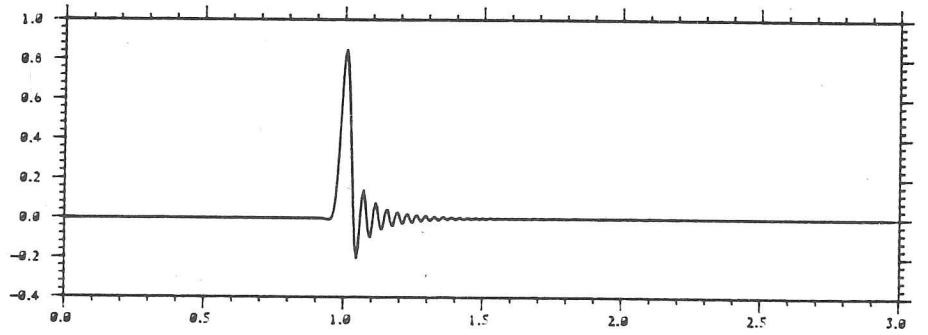
Figure 8b. Propagation of polychromatic pulse by five-point central difference semi-discretisation with $\mu = .4$, $h = 1/160$, $\xi h \simeq .79$.



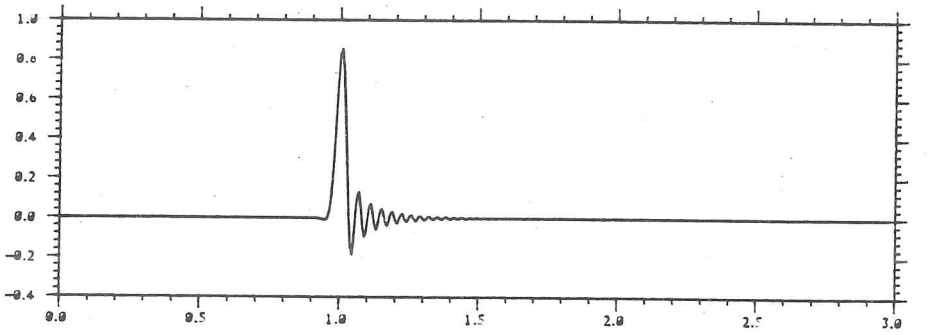
Runge Kutta
Two step
Two stage
Order three



Runge Kutta
Two step
Two stage
Order two



Runge Kutta
Two step
Three stage
Order three



Runge Kutta
Two step
Three stage
Order two

Figure 8b. (continued)

TIME INTEGRATION

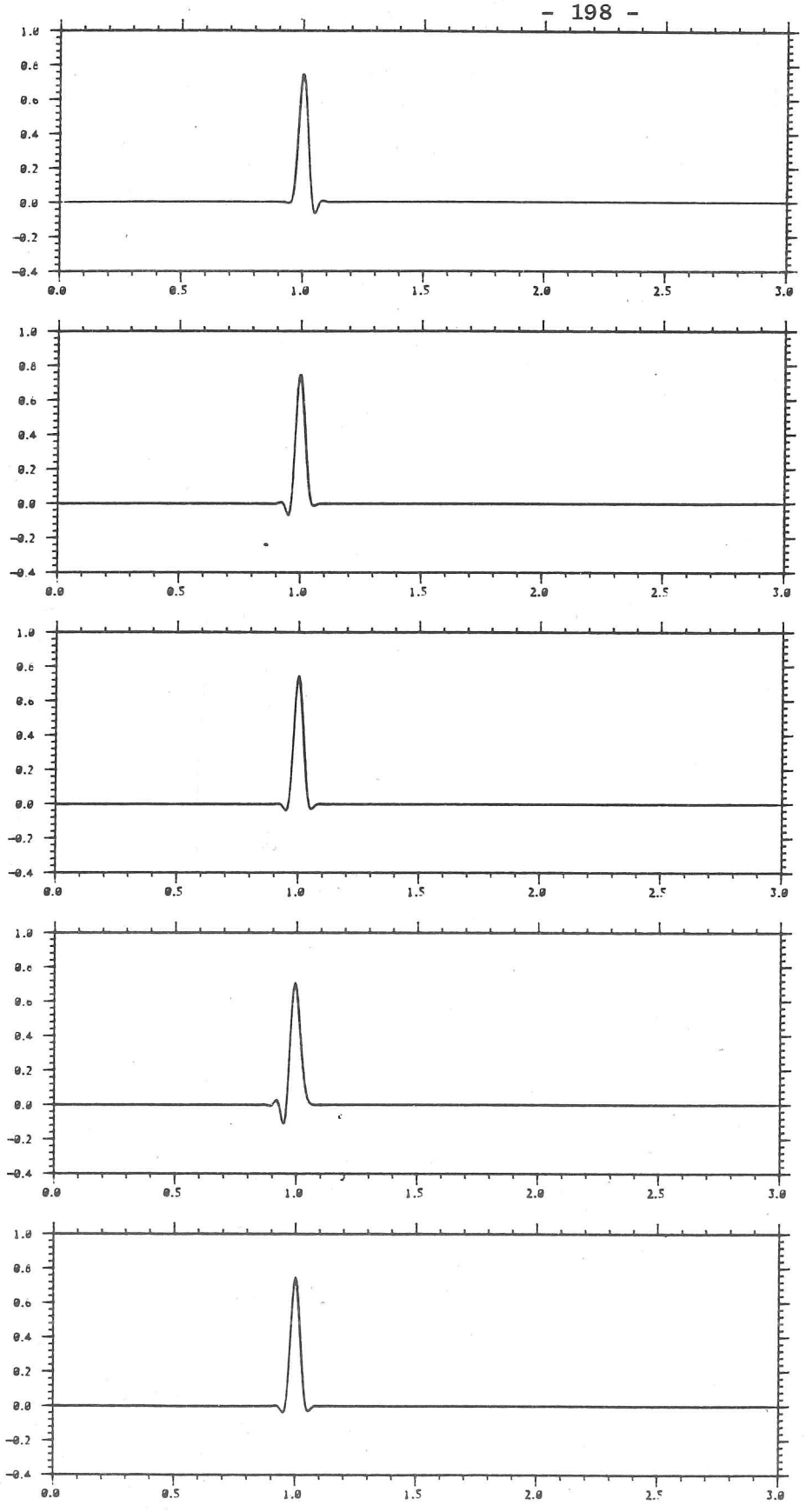


Figure 9a. Propagation of polychromatic pulse by case A semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq .79$.

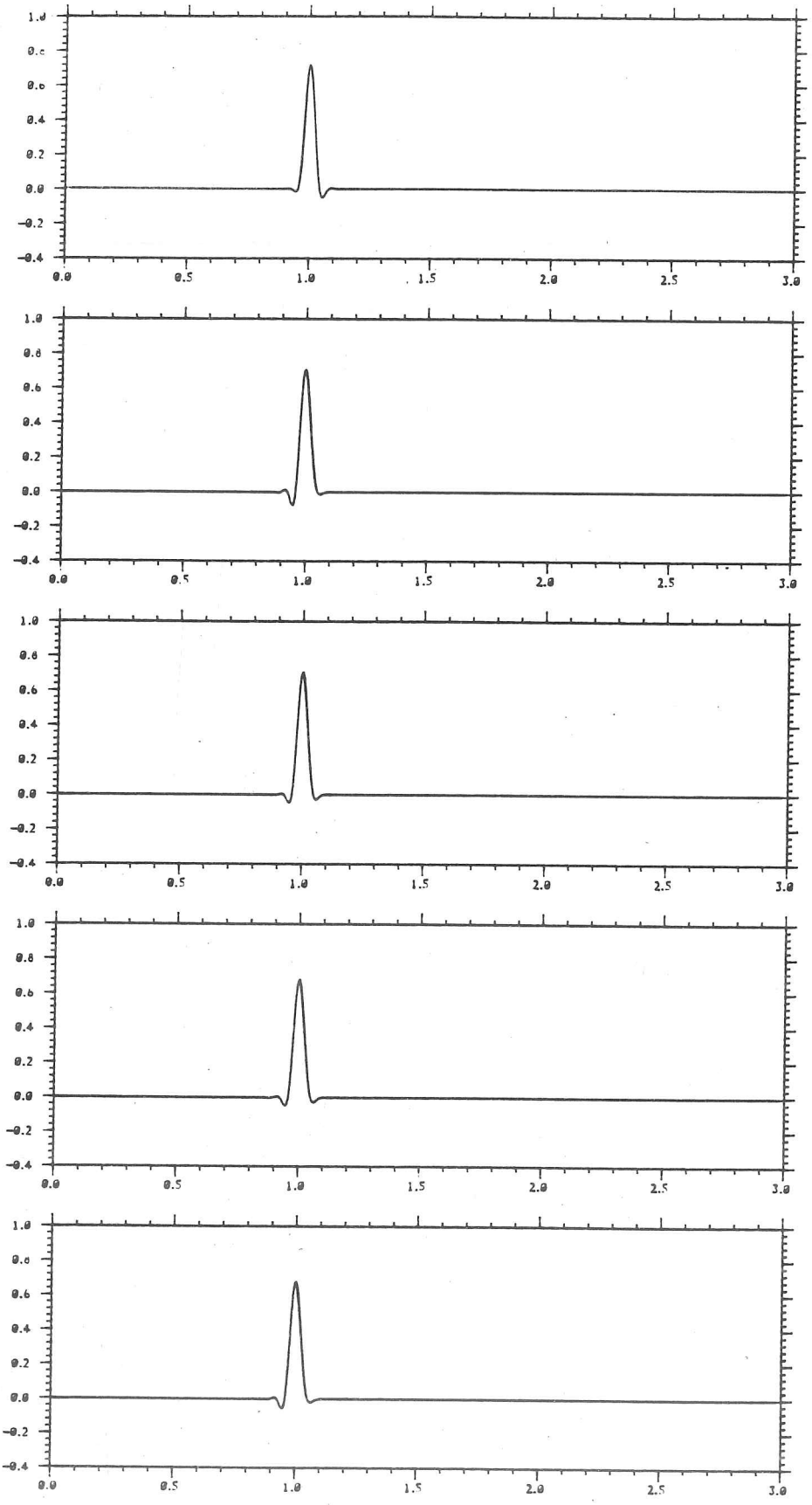


Figure 9b. Propagation of polychromatic pulse by case B semi-discretisation, with $\mu = .4$, $h = 1/160$, $\xi h \simeq 1/160$ (except Runge Kutta two-step two-stage order three, $\mu = .35$).

Examples of the propagation of the pulse are given in Figures 8 and 9. Clearly the second-order semidiscretisation C is much more dispersed. The pulse very quickly disperses into a train of oscillations. As stated by Trefethen, this is also explained by group speed. The waves with higher wave numbers travel more slowly for this method than for the fourth-order method, and so the pulse is more quickly dispersed. Once again we see that the Runge Kutta schemes which have been demonstrated to be dissipative for conservative schemes suffer less dispersion. The dissipative semi-discretisations allow very little dispersion. As for the stepfunctions, the wiggles around the pulse are more evenly balanced for odd-order rather than even-order methods of time integration.

Our experiments have therefore confirmed the view of Trefethen that evaluation of group velocity is very important in the development of robust numerical schemes which are conservative. However, we also remark that it seems to be sufficient to consider the group velocity of the semi-discretisation alone. We have not, as yet, attempted to extend our analysis to experiments with forcing terms; here the group velocity dependence on the time domain would be important. For dissipative schemes, we have shown that calculation of the predicted amplitude reduction is very useful for determining possible dissipative affects but that group velocity predictions are not helpful. Consequently, for any numerical model, by consideration of either group velocity or amplitude reasonable choices of grid sizes can be made so

as to minimise errors. In this way numerical models suited to particular kinds of problems may be chosen sensibly.

Finally we return to the original problem described at the beginning of this section. Very often the asymptotic stage of a nonlinear problem is solved with added artificial dissipation to damp unwanted oscillations near shocks. We have shown that by discretising in the spatial domain in a way other than by central differences, we may incorporate dissipation without adding extra terms to the original equation. Approximate measurements of the dissipation can be evaluated by a simple Fourier analysis and then grid size chosen to give the desired amount of dissipation. Such schemes have very good shock resolution properties and integrating in time by a method with comparable order helps ensure minimal overshoot.

Furthermore, our results demonstrate that equivalent shock resolution can be obtained using an explicit method of time integration as compared with an implicit method. Therefore the Runge Kutta methods are very promising for generalisations to higher dimensions, where solution of a large linear system by an iterative process would not be desirable. The main shortcoming of the Runge Kutta methods at present is the restricted size of their stability intervals. However, this may be overcome by applying the optimisation technique to design efficient schemes for particular spectral functions. There seems to be no particular preference for either case A or case B in terms of shock

resolution; they are both strongly dissipative. Case A seems to be more promising since its spectral curve extends less into $\mathbb{C}^-$. Moreover, less dissipative SD's may be considered and then considerably improved efficiency would be expected.

In situations where efficiency is more important than dissipativity conservative SD's are recommended. By designing a numerical model consisting of a conservative SD and a Runge-Kutta method with demonstrated dissipativity reasonable shock resolution is sometimes obtained. It is for these problems that some of the algorithms of Chapter 7 will be particularly useful. The experiments demonstrate that better shock resolution is achieved when fewer operations are required. In such cases internal instability, which is in any case negligible for so few stages, is further eliminated. This suggests that particularly good models would be obtained by using algorithms designed to be implemented with minimal numbers of operations which is, anyway, a desirable property for increased efficiency.

In conclusion our results do demonstrate that improved shock resolution and efficiency are achievable. Consequently, some of the numerical models described here are good candidates for inclusion in a numerical scheme for solving non-linear problems.

APPENDIX A: OPTIMISED COEFFICIENTS OF STABILITY POLYNOMIALS

In this Appendix we give the optimal coefficients for schemes with maximal interval of stability on the imaginary axis. These schemes are labelled by χ_1. We also give coefficients of near optimal stability polynomials suitable for integrating schemes efficiently in wedges with angles 49.1°, 60°, 45° and 90°. These schemes are labelled by χ_2 , χ_3 , χ_4 , χ_5 respectively.

In Table 1 we give the coefficients for one-step Runge Kutta where θ_3 and θ_4 are as defined in Appendix B. In Table 2 the coefficients for the two-step Runge Kutta are given.

		χ_1	χ_2	χ_3
m=3 p=2	θ_3	.25E+00	.1069585836271382E+00	.9346790288314027E-01
m=4 p=2	θ_3	.1667E+00	.9886493454938255E-01	.1125E+00
	θ_4	.4167E-01	.1163360523161749E-01	.105E-01
m=4 p=3	θ_4	.4167E-01	.24742981E-01	.2475991346207076E-01

Table 1. Coefficients of optimal one-step Runge Kutta

		X_2	X_3	X_1	X_4	X_5
m=2	p_o	.8451451442168584E-01	-.8E+00	-.95E+00	-.8E+00	.17865567E+00
p=2	p_1	.1749890016710914E+00	-.8312763116244109E+00	.025E+00	-.8271861203632445E+00	.32125698E+00
	p_2	.1691337598823489E+00	-.2293913702887822E+00	.006E+00	-.2070997408542243E+00	.526832147E-02
m=2 p=3	p_o	-.95E+00	-.5E+00	.2E+00	-.8E+00	.2E+00
	p_o	-.95E+00	-.95E+00	.9E+00	-.9498602398232146E+00	.1328892973220466E+00
m=3	p_1	-.9601462734927106E+00	-.9589723258234430E+00	-.039E+00	-.9610774733258250E+00	.1867862137573033E+00
p=2	p_2	-.2303556519384286E+00	-.2403755256952069E+00	-.009E+00	-.2183052948384030E+00	.1447453799931632E+00
	p_3	-.3285934381961209E-01	-.2315316182424289E-01	-.001E+00	-.1880745988278995E-01	.3568346078615244E-01
	s_3	.3529583902369824E-01	.2484002049172644E-01	.295E+00	.1981382053275292E-01	.1798768896385816E+00
m=3	p_o	.4656007508029430E+00	-.8E+00	1.00E+00	0.73E+00	.37862779E+00
p=3	p_1	.2608860734722656E+00	-.8231733436102460E+00	0.0	0.59E+00	.43085447E+00
	p_2	.2847099371820824E-01	-.4311016976241868E+00	0.0	0.14E+00	.24645432E+00
	p_3	.1324518317056125E-01	-.8062146734616475E-01	0.0	0.18E-01	.78023998E-01

Table 2. Coefficients of optimal two-step Runge Kutta.

APPENDIX B: INTEGRATION PARAMETERS

First we give equations for integration parameters of the order two and three, two and three-stage two-step formulae. We give expressions for schemes with and without error control.

a) Two-stage two-step order two

Optimisation performed with respect to $\{p_o, p_1, p_2\}$

$$\beta = p_o \quad , \quad v_1 = p_1 - \frac{p_2}{\alpha} \quad , \quad v_2 = \frac{p_2}{\alpha}$$

$$w_1 = 1 + p_o - p_1 - \frac{1}{\alpha}\left[\frac{1}{2} - p_2 + p_1 - \frac{p_o}{2}\right] \quad \omega_2 = \frac{1}{\alpha}\left[\frac{1}{2} - p_2 + p_1 - \frac{p_o}{2}\right] .$$

Error control by two-stage two-step order-three scheme

$$p_o = p_o^* \quad v_1 = v_1^* \Rightarrow \quad \alpha = \frac{12\,p_2 - p_o + 5}{6(2\,p_1 - p_o + 1)}$$

$$v_2^* = \frac{(5 - p_o)(p_o - 1 - 2\,p_1)}{2(12\,p_2 - p_o + 5)} \quad , \quad w_2^* = -v_2^* \, , \, w_1^* = p_o + 1 - v_1^* .$$

b) Two-stage two-step order three

Optimisation performed with respect to $\{p_o\}$

$$\beta = p_o \, , \, v_1 = \frac{p_o - 1}{2} - \left(\frac{p_o - 5}{12\,\alpha}\right) \quad v_2 = \frac{p_o - 5}{12\,\alpha}$$

$$w_1 = \frac{3 + p_o}{2} - \left(\frac{5 - p_o}{12\,\alpha}\right) \quad w_2 = \frac{5 - p_o}{12\,\alpha} .$$

Error control by three-stage two-step order-four scheme

$$p_o = p_o^* \ , \ v_1 = v_1^* \Rightarrow \quad \alpha = \frac{4}{5 - p_o} \ , \ \delta = \alpha \quad .$$

The coefficients v_2^*, v_3^*, w_1^*, w_2^*, w_3^* may be evaluated from the equations in Table 6.3.3.

c) Three-stage two-step order-two

Optimisation performed with respect to $\{p_o, p_1, p_2, p_3, s_3\}$

$$v_1 = p_1 - \frac{1}{\delta\alpha^2}[\, p_3(\alpha - (\alpha + \delta)) + p_2\, \delta\alpha \,]$$

$$v_2 = \frac{1}{\delta\alpha^2}[\, p_1\, \delta\alpha - p_3(\gamma + \delta) \,]$$

$$v_3 = \frac{p_3}{\delta\alpha}$$

$$w_1 = 1 + p_o - p_1 - \frac{1}{\delta\alpha^2}[\, s_3(\alpha - (\alpha + \delta)) + (\tfrac{1}{2}(1 - p_o) + p_1 - p_2)\, \delta\alpha \,]$$

$$w_2 = \frac{1}{\delta\alpha^2}[\, \delta\alpha(\tfrac{1}{2}(1 - p_o) + p_1 - p_2) - s_3(\gamma + \delta) \,]$$

$$w_3 = \frac{s_3}{\delta\alpha} \quad .$$

In reasonable implementations $\gamma = 0$.

Error control by three-stage two-step order-three scheme

$$p_o = p_o^*, \quad v_i = v_i^* \quad 1 \leq i \leq 3 \quad , \quad \gamma = 0$$

$$\delta = \alpha - \frac{(\frac{1}{2}(1-p_o) + p_1)\alpha^2}{\frac{1}{6}(1+p_o) - \frac{p_1}{2} + p_2} \quad ,$$

$$\text{if } v_1 = 0 \text{ then } \alpha = \frac{p_2}{p_1} + \frac{3p_3(1 - p_o + 2p_1)}{4p_o - 2 - 9p_1 + 6p_2} \quad .$$

w_1^*, w_2^*, w_3^* as above with s_3 replaced by s_3^*

$$s_3^* = \frac{1}{6}(1 + p_o) - \frac{p_1}{2} + p_2 - p_3 \quad .$$

d) Three-stage two-step order-three

Optimisation performed with respect to $\{p_o, p_1, p_3, p_3\}$

$v_1, v_3, v_3, w_1, w_2, w_3,$ as given in c) where

$$s_3 = \frac{1}{6}(1 + p_o) - \frac{p_1}{2} + p_2 - p_3 \quad .$$

In addition α is a root of the equation

$$(\frac{1}{2}(1 - p_o) + p_1)\delta\alpha^2 + (\frac{1}{6}(1 + p_o) - \frac{p_1}{2} + p_2)((\gamma + \delta)^2 - \alpha(\gamma + 3\delta)$$

Error control by three-stage two-step order-four scheme

$$p_o = p_o^*, \quad v_i = v_i^* \quad 1 \leq i \leq 3 \quad , \quad \gamma = 0$$

$$\delta = \alpha - \frac{(\frac{1}{2}(1-p_o) + p_1)\alpha^2}{\frac{1}{6}(1+p_o) - \frac{p_1}{2} + p_2} \quad ,$$

$$\text{if } v_1 = 0 \text{ then } \alpha = \frac{p_2}{p_1} + \frac{3p_3(1 - p_o + 2p_1)}{4p_o - 2 - 9p_1 + 6p_2} \quad .$$

w_1^*, w_2^*, w_3^* as above with s_3 replaced by s_3^*

$$s_3^* = \frac{1}{6}(1 + p_o) - \frac{p_1}{2} + p_2 - p_3 \quad .$$

d) Three-stage two-step order-three

Optimisation performed with respect to $\{p_o, p_1, p_3, p_3\}$

$v_1, v_3, v_3, w_1, w_2, w_3,$ as given in c) where

$$s_3 = \frac{1}{6}(1 + p_o) - \frac{p_1}{2} + p_2 - p_3 \quad .$$

In addition α is a root of the equation

$$(\frac{1}{2}(1 - p_o) + p_1)\delta\alpha^2 + (\frac{1}{6}(1 + p_o) - \frac{p_1}{2} + p_2)((\gamma + \delta)^2 - \alpha(\gamma + 3\delta)$$

Error control by three-stage two-step order-four scheme

$$p_o = p_o^* \quad , \quad \gamma = 0$$

$$\Rightarrow \quad \delta = 3\alpha + \alpha^2 \left(\frac{p_o - 5}{2} \right)$$

$$\text{and} \quad \alpha = \frac{-4(1 + p_o + 3p_1 + 6p_2)}{(p_o - 5)(1 + p_o - 3p_1 + 6p_2) + 6(1 - p_o) + 12p_1} \quad .$$

Coefficients $v_1^*, v_2^*, v_3^*, w_1^*, w_2^*, w_3^*$ given in Table 6.3.3

Before giving the equivalent formulae for the one-step schemes we must clarify our notation. We define the one-step m-stage Runge-Kutta scheme by

$$y^{n+1} = y^n + h \sum_{i=1}^{m} \theta_i \ell_i \quad ,$$

$$\ell_i = f(y_i^n) \quad 1 \leqslant i \leqslant m \quad ,$$

$$y_1^n = y^n, \quad y_i^n = y^n + h \sum_{j=1}^{i-1} \lambda_{ij} \ell_j \quad .$$

Then consistency conditions up to order four of these schemes are, in the notation of (6.3.1), as follows:

$$C_1 = 1 - \sum_{i=1}^{m} \theta_i$$

$$C_{21} = \frac{1}{2} - \sum_{i=2}^{m} \sum_{j=1}^{i-1} \theta_i \lambda_{ij}$$

$$C_{31} = \frac{1}{6} - \frac{1}{2} \sum_{i=2}^{m} \theta_i \left(\sum_{j=1}^{i-1} \lambda_{ij} \right)^2$$

$$C_{32} = \frac{1}{6} - \sum_{i=3}^{m} \sum_{j=2}^{i-1} \sum_{k=1}^{j-1} \theta_i \lambda_{ij} \lambda_{jk}$$

$$C_{41} = \frac{1}{24} - \frac{1}{6} \sum_{i=2}^{m} \theta_i \left(\sum_{j=1}^{i-1} \lambda_{ij} \right)^3$$

$$C_{42} = \frac{1}{8} - \sum_{i=3}^{m} \sum_{j=2}^{i-1} \sum_{k=1}^{j-1} \theta_i \lambda_{ij} \lambda_{jk} \sum_{j=2}^{i-1} \lambda_{ij}$$

$$C_{43} = \frac{1}{24} - \frac{1}{2} \sum_{i=3}^{m} \sum_{j=2}^{i-1} \theta_i \lambda_{ij} \left(\sum_{k=1}^{j-1} \lambda_{jk} \right)^2$$

$$C_{44} = \frac{1}{24} - \sum_{i=4}^{m} \sum_{j=3}^{i-1} \sum_{k=2}^{j-1} \sum_{\ell=1}^{k-1} \theta_i \lambda_{ij} \lambda_{jk} \lambda_{k\ell} \quad .$$

Also, the coefficients of the stability polynomial defined by

$$y^{n+1} = R(z)\, y^n$$

can be shown in the usual way to be given by

$$r_1 = \sum_{i=1}^{m} \theta_i$$

$$r_j = \sum_{k_1=j}^{m} \sum_{k_2=j-1}^{k_1-1} \cdots \sum_{k_j=1}^{k_{j-1}-1} \theta_{k_1} \lambda_{k_1 k_2} \cdots \lambda_{k_{j-1} k_j} \quad .$$

Thus formulae for calculating the coefficients of these schemes with and without error control can now be given.

e) Three-stage one-step order-two scheme

Optimisation performed with respect to r_3

$$\theta_3 = \frac{r_3}{c_2 c_1} \quad \theta_2 = \frac{1}{2c_1} - \frac{r_3}{c_1^2} \quad \theta_1 = 1 - \frac{r_3}{c_2 c_1} - \frac{1}{2c_1} + \frac{r_3}{c_1^2}$$

Error control

$$\theta_3^* = \frac{1}{6c_2 c_1} \quad \theta_2^* = \frac{1}{c_1}\left(\frac{1}{2} - \frac{1}{6c_1}\right) \quad \theta_1^* = 1 - (\theta_2^* + \theta_3^*)$$

Choose $c_2 = c_1 = \frac{1}{2}$, then $\theta_1 = \theta_1^* = 0$

$$\theta_3 = 4r_3 \quad \theta_2 = 1 - 4r_3 \quad \theta_3^* = \frac{2}{3} \quad \theta_2^* = \frac{1}{3} \ .$$

f) Three-stage one-step order-three scheme

$$\theta_3 = \frac{1}{6c_2 c_1} \quad \theta_2 = \frac{1}{c_1}\left(\frac{1}{2} - \frac{1}{6c_1}\right)\theta_1 = 1 - (\theta_2 + \theta_3) \ .$$

Error control by four-stage one-step order-four

$$\theta_4^* = \frac{1}{24c_1 c_2 c_3} \quad \theta_3^* = \frac{1}{6c_2 c_1} - \frac{1}{24c_1^2 c_2}$$

$$\theta_2^* = \frac{1}{2c_1} - \frac{1}{6c_1^2} - \frac{1}{24c_1^2} + \frac{1}{24c_1^3} \quad \theta_1^* = 1 - \theta_2^* - \theta_3^* - \theta_4^* \ .$$

Reduce operations by putting $\theta_1 = \theta_1^*$, $\theta_2 = \theta_2^*$, $\theta_1 = 0$ then $c_1 = c_2 = c_3 = \frac{1}{2}$.

$$\theta_3 = \frac{2}{3} \qquad \theta_2 = \frac{1}{3} \qquad \theta_1 = 0$$

$$\theta_4^* = \frac{1}{3} \qquad \theta_3^* = \frac{1}{3} \qquad \theta_2^* = \frac{1}{3} \qquad \theta_1 = 0 \quad .$$

9) Four-stage one-step order-two scheme

Optimisation with respect to $\{r_2, r_3\}$

$$\theta_4 = \frac{r_3}{c_1 c_2 c_3} \qquad \theta_3 = \frac{r_2}{c_2 c_1} - \frac{r_3}{c_1^2 c_2} \qquad \theta_2 = \frac{1}{2c_1} - \frac{r_2}{c_1^2} - \frac{r_3}{c_1^2 c_2} + \frac{r_3}{c_1^3}$$

$$\theta_1 = 1 - (\theta_2 + \theta_3 + \theta_4) \quad .$$

Error control four-stage one-step order-three scheme

Choose $\theta_1 = \theta_2 = \theta_3 = 0$ then $c_1 = \frac{r_3}{r_2}$, $c_2 = 2r_2$, $c_3 = \frac{1}{2}$ and $\theta_4 = 1$.

Also choose $\theta_1^* = 0$ then

$$\theta_4^* = \frac{1}{6}\left[\frac{12r_3^2 - 2r_2^2 + r_3 + 6r_2 r_3}{2r_2 r_3 - 2(r_3^2 + r_2^3)}\right]$$

$$\theta_3^* = -\frac{1}{6}\left[\frac{r_3 + 6r_3 r_2 + 3r_2^2}{2r_2 r_3 - 2(r_3^2 + r_2^3)}\right]$$

$$\theta_2^* = \frac{12\, r_3^2 + 5\, r_2^2}{6\,[2r_2 r_3 - 2(r_3^2 + r_2^3)]} \quad .$$

h) Four-stage one-step order-three

Optimisation with respect to $\{r_3\}$

$$\theta_4 = \frac{r_3}{c_1c_2c_3} \quad \theta_3 = \frac{1}{6c_2c_1} - \frac{r_3}{c_1^2c_2} \quad \theta_2 = \frac{1}{2c_1} - \frac{1}{6c_1^2} - \frac{r_3}{c_1^2c_2} + \frac{r_3}{c_1^3}$$

$$\theta_1 = 1 - (\theta_2 + \theta_3 + \theta_4) \quad .$$

Error control four stage one step order-four

$$\theta_4^* = \frac{1}{24c_1c_2c_3} \quad \theta_3^* = \frac{1}{6c_2c_1} - \frac{1}{24c_1^2c_2} \quad \theta_2^* = \frac{1}{2c_1} - \frac{1}{6c_1^2} - \frac{1}{24c_1^2c_2} + \frac{1}{24c_1^3}$$

$$\theta_1^* = 1 - (\theta_2^* + \theta_3^* + \theta_4^*) \quad .$$

Put $c_1 = c_2 = c_3 = \frac{1}{2}$ then

$$\theta_4^* = \theta_3^* = \theta_2^* = \frac{1}{3} \quad \theta_1^* = 0$$

$$\theta_4 = 8r_3 \quad \theta_3 = \frac{2}{3} - 8r_3 \quad \theta_2 = \frac{1}{3} \quad \theta_1 = 0 \quad .$$

APPENDIX C : MAXIMAL COURANT NUMBERS

In Table 1 we give the following maximal numbers:

1) maximal Courant numbers for integration of semi-discretisations A and B by schemes χ_2 and χ_3 ;
2) maximal radii of wedges with angles $\pi / 4$ and $\pi / 2$
3) maximal intervals of stability on the imaginary axis.

In Table 2 we give the same numbers as in Table 1, but scaled by the number of stages so that a comparison of efficiency may be made.

In Table 3 we give the maximal Courant numbers for integration of cases A and B by scheme χ_1 and the actual observed values for integration by χ_2 and χ_3 respectively.

In Table 4 we either give the radius of the circle of contractivity or a bound on it. The circle of contractivity of radius r is defined to be circle with centre at $(-r,0)$ and with radius r in which a given scheme is contractive with respect to an arbitrary norm defined on R^n .

No. of steps	No. of stages	Order p	Maximal values for				
			A	B	χ_1	χ_4	χ_5
2	2	2	.4282	.2020	1.987	2.747	1.467
		3	.5041	.3546	1.00	2.060	1.00
	3	2	.0650	.0765	2.935	5.9167	2.289
		3	.5673	.6131	2.647	3,296	2.212
1	3	2	.8620	.6377	2		
		3	.6258	.4188	$\sqrt{3}$		
	4	2	.8344	.8333	$2\sqrt{2}$		
		3	.9651	.6755	$2\sqrt{2}$		

Table 1. Maximal parameters.

No. of steps	No. of stages	Order p	Scaled maximal values for				
			A	B	χ_1	χ_4	χ_5
2	2	2	.2141	.1010	.9935	1.374	.735
		3	.2521	.1773	.50	1.030	.5
	3	2	.0217	.0252	.9783	1.972	.763
		3	.1891	.2044	.9428	1.0987	.737
1	3	2	.2873	.2126	.6667		
		3	.2086	.1396	.5774		
	4	2	.2086	.2083	.7071		
		3	.2413	.1689	.7071		

Table 2. Scaled maximal parameters

No.of steps	No.of stages	Order p	Predicted CFL by χ_1 for A	Predicted CFL by χ_1 for B	Observed for A	Observed for B
2	2	2	.12	.01	.4	< .3
		3	.35	.25	.5	.3
	3	2	.03	.02	< .4	< .4
		3	0	0	.5	.6
1	3	2	.49	.33	.8	.6
		3	.63	.42	.6	.4
	4	2	.70	.46	.8	.8
		3	.70	.46	.9	.6

Table 3. Observed maximal parameters.

No.of steps	No.of stages	Order p	Radius of the circle of contractivity χ_2	χ_3	χ_1
2	2	2	<.517	0	0
		3	0	0	0
	3	2	0	0	0
		3	0	0	0
1	3	2	1.55	1.78	2/3
		3	1	1	1
	4	2	< 2.212	0	0
		3	1.68	1.68	1

Table 4. Radii of circles of contractivity.

Index of Notation

IFD	implicit full discretisation	1.2
ISD	implicit semi-discretisation	1.2
I, I^-, I^+	strips in $\mathbb{C}$, $\mathbb{C}^-$ and $\mathbb{C}^+$	3.1
I_m^p	optimal imaginary stability polynomial	5.6
k	number of timesteps	5.3
k_i	stage of Runge Kutta scheme	6.2
L	linear differential operator	1.2
ℓ_i	stage of Runge Kutta scheme	6.2
M	degree of numerator of $A(z,\mu)$, $M = \tilde{r} + \tilde{s}$	4.1
m	degree of $G(z)$, $m = r + s$; Chapters 1-4	3.2
m	number of stages in Runge Kutta scheme; Chapters 5-8	5.3
N	degree of denominator of $A(z,\mu)$, $N = \tilde{R} + \tilde{S}$	4.1
n	degree of $F(z)$, $n = R + S$	3.2
O.D.E.	ordinary differential equation	6.7
$P(z)$	stability polynomial of two-step Runge Kutta	6.4
p	order of accuracy	2.2
$R(z)$	stability polynomial	1.8
R, r	numbers defining SD	1.5
$\tilde{R}, \tilde{r}$	numbers defining FD	1.5
S	Stability region	5.3
$S(z)$	stability polynomial of two-step Runge Kutta	6.4
S, s	numbers defining SD	1.5
$\tilde{S}, \tilde{s}$	numbers defining FD	1.5
$T_n(z)$	Chebyshev polynomial $T_n(k) = \cos n\theta$, $\theta = \arccos z$	5.6
t	time variable	
$U_n(z)$	Chebyshev polynomial of second kind $U_n(z) = \frac{\sin(n+1)\theta}{\sin\theta}$, $\theta = \arccos z$.	5.6
$u(x,t)$	exact solution of partial differential equation	1.1

$v_j(t)$	discrete approximation to $u(x_j, t)$	1.2
v_i, v_i^*	real coefficients of two-step Runge Kutta formula	6.2
$W^{(i)}$	generic unit of storage	7.1
$W^{(n)}$	array representing y^n	7.1
WHN	Wanner-Hairer-Nørsett	3.1
w_i, w_i^*	real coefficients of two-step Runge Kutta formula	6.2
x	space variable	
y(t)	solution of ordinary differential system of equations	5.3
y^n	discrete approximation to $y(t_n) = y(n\Delta t)$	5.3

****

α_{ij}	real coefficients of two-step Runge Kutta formula	6.2
β	zero stability parameter of two-step Runge Kutta formula	6.2
β_{imag}	imaginary stability boundary	5.6
β_{real}	real stability boundary	5.5
Δt	mesh size in time	1.2
Δx	vector of mesh sizes in spatial domain	1.2
$\kappa(B)$	condition number of matrix B	1.4
λ_{ij}	real coefficients of one-step Runge Kutta formula	App.B
λ	complex eigenvalue	1.4
μ	Courant number	1.3
ξ	real wavenumber	1.7
π	pi	
$\rho(B)$	spectral radius of B	1.8
$\sigma(B)$	spectrum of B	1.4
$\sigma(z)$	order star function	3.1
$\Phi(\alpha, z)$	characteristic function of multistep multistage formula	1.3

$\phi(x)$	initial condition of differential equation ----------	1.1
ω	real frequency ----------	1.7
χ_i	label for Runge Kutta formula ----------	App.A
∂	Jordan curve in complex plane ----------	3.1

REFERENCES

[Ba75] G.A. Baker, Essentials of Padé Approximants, New York, Academic Press (1975).

[Br75] P. Brenner, V. Thomée and L.B. Wahlbin, Besov Spaces and Applications to Difference Methods for Initial Value Problems, Springer-Verlag, New York (1975).

[Br53] L. Brillouin, Wave Propagation in Periodic Structures, Dover, New York (1953).

[Bu62] J.C. Butcher, Coefficients for the study of Runge-Kutta integration processes, J. Aust. Math. Soc. 3, 185-201 (1963).

[Da56] G. Dahlquist, Numerical integration of ordinary differential equations. Math. Scand. 4, 33-53 (1956).

[Da63] G. Dahlquist, A special stability problem for linear multistep methods, BIT 3, 27-43 (1963).

[Da79] G. Dahlquist and R. Jeltsch, Generalised discs of contractivity for explicit and implicit Runge Kutta methods. Report TRITA-NA-7906. Dept. Comp. Sci. Roy. Inst. of Techn., Stockholm (1979).

[De77] K. Dekker, Stability of linear multistep methods on

the imaginary axis,
BIT 21, 66-79 (1981).

[En80] B. Engquist and S. Osher, Stable and entropy satisfying approximations for transonic flow calculations,
Math. Comp. 34, 45-75 (1980).

[En81] B. Engquist and S. Osher, One sided difference approximations for nonlinear conservation laws,
Math. Comp. 36, 321-352 (1981).

[Go64] S.K. Godunov and V.S. Ryabenki, Introduction to the Theory of Difference Schemes,
Interscience Publishers, New York, also Fizmatigiz, Moscow (1962).

[Go77] D. Gottlieb and S.A. Orszag, Numerical Analysis of Spectral Methods: Theory and Applications,
SIAM, Philadelphia (1977).

[Go78] M. Goldberg and E. Tadmor, Scheme independent stability criteria for difference approximations of hyperbolic boundary value problems I,
Maths. Comp 32, 1097-1107 (1978).

[Go81] M. Goldberg and E. Tadmor, Scheme independent stability criteria for difference approximations of hyperbolic boundary value problems II,

Maths. Comp. 36, 603-626 (1981).

[Gr58] V. Grenander and G. Szegö, Toeplitz Forms and their Applications,
University of California Press, Berkeley and Los Angeles (1958).

[Gr69] R.T. Gregory and D.L. Karney, A Collection of Matrices for Testing Computational Algorithms,
Wiley Interscience, New York (1969).

[Gr78] D.F. Griffiths, I. Christie and A.R. Mitchell, Analysis of error growth for explicit difference schemes in conduction convection problems,
Rep. NA/29, University of Dundee (1978).

[Gu72] B. Gustafsson, H.O. Kreiss and A. Sundström, Stability theory of difference approximations for initial boundary value problems II,
Math. Comp. 26, 649-686 (1972).

[Ha76] A. Harten, T.M. Hyman and P.D. Lax, On finite difference approximations and entropy conditions for shocks,
Comm. Pure and Appl. Maths. 29, 297-322 (1976).

[He62] P. Henrici, Discrete Variable Methods in Ordinary Differential Equations

Wiley, New York (1962).

[Ho77] P.J. van der Houwen, Construction of Integration Formulae for Initial Value Problems,
North Holland, Amsterdam (1977).

[Ho80] P.J. van der Houwen and B.P. Sommeijer, On the internal stability of explicit, m stage Runge Kutta methods for large m values,
ZAMM 60, 479-485 (1980).

[Ho82] P.J. van der Houwen and B.P. Sommeijer, A special class of Runge Kutta methods with extended real stability interval,
I.M.A. J. Num. Anal. 2, 183-209 (1982).

[Ho83] P.J. van der Houwen and B.P. Sommeijer, Predictor corrector methods with improved absolute stability regions,
I.M.A. J. Num. Anal. 3, 417-437 (1983).

[Is79] A. Iserles, A note on Padé approximations and generalized hypergeometric functions,
BIT 19, 543-545 (1979).

[Is82] A Iserles, Order stars and a saturation theorem for first order hyperbolics,
I.M.A. J. Num. Anal. 2, 49-61 (1982).

[Is83a] A. Iserles and G. Strang, The optimal accuracy of difference schemes,
Trans. Amer. Math. Soc. 2, 779-803 (1983).

[Is83b] A. Iserles, Order stars, approximations and finite differences, I: the general theory of order stars,
SIAM J. Num. Anal. (to appear, 1985).

[Is83c] A. Iserles, Order stars, approximations and finite differences, II: theorems in approximation theory,
SIAM J. Num. Anal. (to appear, 1985).

[Is83d] A. Iserles, Order stars, approximations and finite differences, III: finite differences for $u_t = \omega u_{xx}$,
SIAM. J. Num. Anal. (to appear, 1985).

[Is84a] A. Iserles and R.A. Williamson, Stability and accuracy of semidiscretized finite difference methods,
I.M.A. J. Num. Anal. 4, 289-307 (1984).

[Is84b] A. Iserles, Numerical Analysis of Differential Equations,
Springer Verlag, New York (to appear 1984).

[Is84c] A Iserles, Multistep discretisations of linear hyperbolic equations,
University of Cambridge, D.A.M.T.P. NA5 (1984).

[Is84d] A. Iserles, private communication.

[Ja83] A Jameson, Transonic aerofoil calculations using the Euler equations.
In: Numerical Methods for Aeronautical Fluid Dynamics, ed. P.L. Roe, Academic Press (1982).

[Je78] R. Jeltsch, Stability on the imaginary axis and A-stability of linear multistep methods,
BIT 18, 170-174 (1978).

[Je81] R. Jeltsch and O. Nevanlinna, Stability of explicit time discretisations for solving initial value problems,
Num. Math. 37, 61-91 (1981).

[Je82] R. Jeltsch and O. Nevanlinna, Stability and accuracy of time discretisations for initial value problems,
Num. Math. 40, 245-296 (1982).

[Je83] R. Jeltsch and O. Nevanlinna, Stability of semi-discretisations of hyperbolic problems,
SIAM J. Num. Anal 20, 1210-1218 (1983).

[Je84] R. Jeltsch and K.G. Strack, Accuracy bounds for semidiscretisations of hyperbolic problems,
Math. Comp. (to appear 1984).

[Ki84a] I.P.E. Kinnmark and W.G. Gray, One step integration

methods with maximum stability regions,
Math. Comput. Simulation 26, 87-92 (1984).

[Ki84b] I.P.E. Kinnmark and W.G. Gray, One step integration methods of third-fourth order accuracy with large hyperbolic stability limits, Math. Comput. Simulation 26, 181-188 (1984).

[Kr59] H.O. Kreiss, Uber Matrizen die Beschrankte Halbgruppen Erzeugen,
Math. Scand. 7, 71-80 (1959).

[Kr64] H.O. Kreiss, On difference approximations of the dissipative type for hyperbolic differential equations,
Comm. Pure Appl. Math. 17, 335-353 (1964).

[Kr66] H.O. Kreiss, Difference approximations for the initial boundary value problem for hyperbolic differential equations,
In: Numerical Solution of Nonlinear Differential Equations, D. Greenspan (ed.), Proc. Adv. Symp. Math. Res. Ctr., University of Wisconsin, Wiley (1966).

[Kr68] H.O. Kreiss, Stability theory for difference approximations of mixed initial boundary value problems, I.
Math. Comp. 22, 703-714 (1968).

[Kr70] H.O. Kreiss, Initial boundary value problems for

hyperbolic systems,
Comm. Pure Appl. Math. 23, 277-298 (1970).

[La75] G. Laptev, Conditions for the uniform well posedness of the Cauchy problem for systems of equations,
Soviet Math. Dokl. 16 (1975).

[La66] J.D. Lawson, An order five Runge Kutta process with extended region of stability,
SIAM J. Num. Anal. 3, 593-598 (1966)

[La56] P.D. Lax and R.D. Richtmyer, Survey of the stability of linear finite difference equations,
Comm. Pure Appl. Math. 9, 267-293 (1956).

[La60a] P.D. Lax, The scope of the energy method,
Bull. Amer. Math. Soc. 66, 32-35 (1960).

[La60b] P.D. Lax and B. Wendroff, Systems of conservation laws,
Comm. Pure Appl. Math. 13, 217-237 (1960).

[La61] P.D. Lax On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients,
Comm. Pure Appl. Math. 14, 497-520 (1961).

[La62] P.D. Lax and B. Wendroff, On the stability of

difference schemes,
Comm. Pure Appl. Math. 15, 363-371 (1962).

[La73] P.D. Lax, Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves,
SIAM, Philadelphia (1973).

[Le79] B. van Leer, Towards the ultimate conservative difference scheme, V , a second order sequel to Godunov's method,
J. Comp. Phys. 32, 101-136 (1979).

[Ma77] T.A. Manteuffel, The Tchebychev iteration for non-symmetric linear systems,
Num. Math. 28, 307-326 (1977).

[Mi71] J.H. Miller, On the location of zeros of certain classes of polynomials with applications to numerical analysis,
J. Inst. Maths. Applics. 8, 397-406 (1971).

[Mi80] A.R. Mitchell and D.F. Griffiths, The Finite Difference Method in Partial Differential Equations,
Wiley, Chichester (1980).

[Os69] S. Osher, Systems of difference equations with general homogeneous boundary conditions,
Trans. Amer. Math. Soc. 137, 177-201 (1969).

[Pi83] J. Pike and P.L. Roe, Accelerated convergence of Jameson's finite volume Euler scheme using van der Houwen integrators,
Computers and fluids (to appear).

[Ri67] R. Richtmyer and K. Morton, Difference Methods for Initial Value Problems, Wiley Interscience (1967).

[Ri72] W. Riha, Optimal stability polynomials,
Computing 9, 37-43 (1972).

[Ro81] P.L. Roe, Approximate Riemann solvers, parameter vectors and difference schemes,
J. Comp. Phys. 43, 357-372 (1981).

[Ro82] P.L. Roe, Numerical modelling of shockwaves and other discontinuities,
Tech. Me. Aero 1951, Royal Aircraft Establishment, HMSO London (1978).

[So80] B.P. Sommeijer and J.G. Verwer, A performance evaluation of a class of Runge Kutta Chebyshev methods for solving semi-discrete parabolic differential equations,
Report NW91/80 Centrum voor Wiskinde en Informatica, Amsterdam (1980).

[So84] P. Sonneveld and B. van Leer, A minimax problem

along the imaginary axis,
Nieuw Archief voor Wiskunde (to appear 1984).

[Sp83] M.N. Spijker, Contractivity in the numerical solution of initial value problems,
Numer. Math. 42, 271-290 (1983).

[Sp84] M.N. Spijker, On the relation between stability and contractivity,
Rep. No. 84-03, Inst. Appl. Math. Comp. Sci., University of Leiden, Leiden (1984).

[St62] G. Strang, Trigonometric polynomials and difference methods of maximum accuracy.
J. Math. Phys. 41, 147-154 (1962).

[St64a] G. Strang, Accurate partial difference methods II: Nonlinear problems,
Num. Math. 6, 37-46 (1964).

[St64b] G. Strang, Wiener-Hopf difference equations,
J. Math. Mech. 13, 85-96 (1964).

[St66] G. Strang, Implicit difference methods for initial boundary value problems,
J. Math. Anal. Applic., 16, 188-198 (1966).

[St82] G. Strang and A. Iserles, Barriers to stability,

SIAM J. Num. Anal. 20, 1251-1257 (1983).

[St80] J.C. Strikwerda, Initial boundary value problems for the method of lines,
J. Comp. Phys. 34, 94-107 (1980).

[Tr82a] L. Trefethen, Group velocity for finite difference schemes,
SIAM Review 24, 113-136 (1982).

[Tr82b] L. Trefethen, Wave propagation and stability for finite difference schemes,
Ph.D. Dissertation, Dept. of Comp. Sci., STAN-CS-82-905, Stanford Univ, (1982).

[Tr83] L. Trefethen, Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundström, J. Comp. Phys. 49, 199-217 (1983).

[Ve76a] J.G. Verwer, Multipoint multistep Runge Kutta methods I: On a class of two step methods for parabolic equations,
Rep. N.W. 30/76, Centrum voor Wiskunde en Informatica, Amsterdam (1976).

[Ve76b] J.G. Verwer, Multipoint multistep Runge Kutta methods II: The construction of a class of stabilised three step methods for parabolic equations,

Rep. N.W. 31/76, Centrum voor Wiskunde en Informatica, Amsterdam (1976).

[Ve77] J.G. Verwer, A class of stabilised three step Runge Kutta methods for the numerical integration of parabolic problems,
J. Comp. Math. 3, 155-166 (1977).

[Ve79] J.G. Verwer, On a class of explicit three step Runge Kutta methods with extended real stability intervals,
Rep. N.W. 77/79, Centruum voor Wiskunde en Informatica, Amsterdam (1979).

[Ve80] J.G. Verwer, An implementation of a class of stabilised explicit methods for the time integration of parabolic equations,
ACM Trans. on Math. Software 6, 188-205 (1980).

[Ve82] J.G. Verwer, A note on a Runge Kutta Chebyshev method,
Z. Angew. Math. Mech. 62, 561-563 (1982).

[Ve83] J.G. Verwer and K. Dekker, Step by step stability in the numerical solution of partial differential equations,
Rep. N.W. 161/83, Centrum voor Wiskunde en Informatica, Amsterdam (1983).

[Vi82] R. Vichnevetsky and J. Bowles, Fourier Analysis of

Numerical Approximations of Hyperbolic Equations,
SIAM, Philadelphia (1982).

[Wa67] J.M. Watt, The asymptotic discretisation error of a class of methods for solving ordinary differential equations,
Proc. Camb. Phil. Soc. 61, 461-472 (1967).

[Wa78] G. Wanner, E. Hairer and S.P. Nørsett, Order stars and stability theorems,
BIT 18, 475-489 (1978).

[Wi65] J.H. Wilkinson, The Algebraic Eigenvalue Problem,
Oxford University Press, London (1965).

[Wi84] R.A. Williamson, Padé approximations in the numerical solution of hyperbolic differential equations,
In: Padé Approximation and its Applications, Bad Honeff 1983, H. Werner and H.J. Bunger (eds.), Springer Verlag, New York (1984).