FINITE DIFFERENCE METHODS FOR THE ADVECTION EQUATION

PETER JOHN STEINLE

Thesis submitted for the degree of

Doctor of Philosophy

in

The University of Adelaide

(Faculty of Mathematical Sciences)

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by any other person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

3/12/93

Acknowledgment

I am indebted to many people for their support and encouragement during the period of study for this thesis. In particular, I would like to thank Drs. J. Noye and R. Morrow, for the many useful discussions over the years and their enthusiasm for the research contained within this work. The support from my parents, their suggestions and assistance during the writing of this thesis was greatly appreciated. Finally, but most importantly I would especially like to thank my wife, Vicki, for all the encouragement during the low periods, and for the manual preparation of this work. Without any of these people, I would have nothing to show for my research.

Table of Contents

List of Figures

List of Tables

Abstract

This thesis discusses the development of a class of highly accurate, positive definite, implicit finite difference schemes for modelling advection. These schemes are developed in two steps. Firstly, some high order implicit finite difference schemes are derived using the modified equivalent equation approach of Noye and Hayman (1986). These schemes are then made positive definite using the method of flux-corrected transport (FCT) as described by Boris and Book (1973) and Zalesak (1976).

The modified equivalent equation approach is shown to provide valuable information on the performance of a class of explicit finite difference schemes. Alternatively, by manipulation of the modified equivalent equation, it is shown how to obtain a finite difference scheme that advects well-resolved Fourier components with minimal error. This technique is used to develop some high order implicit finite difference schemes that advect all components (apart from those close to the Nyquist limit), to a very high degree of accuracy. These schemes are shown to be considerably more accurate than another class of implicit high order schemes and to be very efficient in comparison with explicit schemes.

From a discussion of techniques for obtaining positive definite solutions from finite difference schemes, it is apparent that flux-corrected transport is the most direct and efficient method for achieving this goal for implicit schemes. It was claimed in Boris and Book (1976), however, that FCT could not be used in con-

junction with implicit schemes due to the failure of the algorithm REVFCT. It is shown in this work how the REVFCT algorithm can be modified so as to be positive definite. The resulting algorithm is improved by using Zalesak's approach to FCT to obtain a robust and accurate algorithm for modelling advection.

The thesis concludes with further numerical tests involving variable velocity and a discussion on the application of this technique by Morrow (1991) to the highly non-linear problem of modelling gas-discharges.

Chapter 1

Introduction

This work is concerned with developing an accurate numerical scheme to model transport processes involving completely arbitrary velocity fields. Transport processes occur in many fields of study, for example gas discharges (Morrow, 1982), pollutant transport (Cunge et al., 1980), meteorology (Leslie et al., 1985) and oceanography (Owen, 1984).

The use of numerical models and computer simulations needs little justification due to their proven ability to give detailed information for any specified area of the solution domain and the ease of changing the parameters of the experiment. There is, however, always the problem of establishing whether a predicted phenomenon is due to the physics being described by the mathematical model or to the approximations in the numerical model. As such, the selection of the most appropriate numerical scheme for any specific application must be based on knowledge of the behaviour of the system being modelled, as well as the general behaviour of a variety of numerical schemes. Understanding the physical system being modelled will indicate certain desirable features of the numerical solution. An example of this is in pollutant transport where a negative concentration is meaningless. By examining a variety of schemes and keeping these desired features in mind, an appropriate scheme can be selected. This selection will be, to a certain extent, subjective and problem dependent.

This thesis aims to develop a scheme that will be robust and widely applicable.

To achieve this, it must be possible to rank the expected performance of schemes under a wide variety of conditions. Such a ranking will become apparent during the course of this work. The comparison of different schemes requires that the overall accuracy of a numerical method be quantified. There are, however, considerations other than the simple difference between the numerical model and the real world, such as the total time spent by people and machines in achieving the results and the nature of the errors in the numerical solution. The problem of how much computer time may be used in the modelling process is problem- and user- dependent. Some users (operational meteorologists, for example) impose absolute constraints on the run time taken for a numerical model, whereas some researchers have virtually unlimited computer time to obtain the results. All that can be done here is to relate the time spent in obtaining a numerical solution to the gain in numerical accuracy. Depending on the particular problem being modelled, different types of errors can be more serious than others. An example is the presence of spurious oscillations in numerical results, which in some problems, such as modelling gas discharges, leads to catastrophic errors.

The numerical models examined in this work will all be finite difference schemes, because they are the most widely used for modelling transport processes. Many other classes of numerical model exist but some are not suited to arbitrary solution domains (e.g. spectral techniques) or are better suited to solving steady state rather than time-evolution problems (e.g. finite elements). Furthermore, for reasons of space and time, it is not possible to give a detailed analysis of every numerical scheme developed. For these reasons, attention will be confined to finite difference schemes, which still allows for a great flexibility in the choice of schemes. It should be noted that, for problems in one spatial dimension, finite element schemes and finite difference schemes are equivalent.

Particular emphasis will be placed on the ability of a scheme to model gas discharges. The problem will not be discussed in detail until later, but certain aspects

of it are worth mentioning now. It is a problem in which the fine structure of the electric field is important, thereby making it an ideal case for numerical modelling. It also incorporates many features that present difficulties for numerical models, such as the presence of a boundary layer, of sharp gradients and a wide range of velocities. In addition, it is highly non-linear. Any scheme that accurately incorporates all these features will be widely applicable since some, or all, of these features occur in a wide variety of problems.

In many applications, and in solving for gas discharges in particular, it is very important for the numerical results to be *"smooth"*, that is, every local extremum in the numerical solution corresponds to a local extremum in the physical solution. Another term is therefore *"non-oscillatory"* since there can be no spurious oscillations in the numerical results. Schemes that produce such results are also referred to as *"positive definite"* since, if all the initial and boundary values are non-negative then so must be the numerical results. This is both necessary and sufficient for non-oscillatory behaviour in linear schemes that accurately model linear systems, because if oscillations are produced, then by subtracting appropriate values from the initial conditions, these oscillations can create negative values in the numerical solution.

The approach taken to develop a general, accurate and robust scheme for modelling transport processes is as follows. Taking the problem of modelling gas discharges as an example, certain desirable properties of the numerical results are obtained. In Chapter Two, a collection of standard explicit finite difference schemes is reviewed. Associated with this is an investigation into the causes of numerical dispersion and numerical diffusion which are two undesirable features in a numerical solution. Methods for comparing different schemes based on the modified equivalent equation (Noye and Hayman, 1986), and how effectively the schemes model single Fourier components of the initial condition, Noye (1984), and groups of Fourier components (Cathers and O'Connor, 1985) are demonstrated.

The distortion of the various Fourier components by a numerical scheme will be quantified by three functions known as the wave propagation parameters for that scheme. It will be shown that there is an equivalence in the order of accuracy of the modified equivalent equation and the accuracy with which different Fourier components of the solution are modelled. It is possible to explain the major flaws of each scheme by examining how successfully the different Fourier modes are transmitted. Using the modified equivalent equation, it is then possible to design schemes that improve the transmission of the Fourier modes, and are correspondingly more accurate in general. This process is shown to yield the more commonly used high order schemes and raises the possibility of improving other schemes. From this we obtain a benchmark with which to compare new schemes.

Chapter Three discusses implicit finite difference schemes. Starting from standard implicit schemes, some new high order schemes are obtained using the modified equivalent equation approach as discussed in Chapter Two. These methods are not only of a high order of accuracy in theory, but also in practice. It will also be shown that the approximations must be of high order in both space and time and that neglecting temporal approximations imposes a severe restriction on the overall accuracy of the scheme. A further complication in the use of implicit schemes is the problem of supplying appropriate boundary conditions. This will be briefly discussed, since although it is not a difficult problem to overcome, the application of incorrect boundary conditions may lead to a significant loss of accuracy. Results from some simple tests will verify that this process is successful in improving the accuracy of linear schemes. A collection of error measures will be used to quantify this improvement.

It will also be shown, that for the explicit schemes to obtain the accuracy of the new implicit schemes requires considerably more computational time, and so in this sense, the new schemes are very efficient. Furthermore, the improvement in accuracy due to increased resolution is much greater for the high order implicit schemes than

the other schemes discussed.

All the numerical schemes discussed in the next two chapters will be linear, in that if a linear system is being modelled the numerical scheme will also be linear. Godunov (1959) showed that if smooth results are required then first order upwinding is the least diffusive explicit two-level linear scheme. As shown in Chapter Two, however, this particular scheme is not satisfactory for modelling advective processes. As a result, various non-linear modifications have been developed. A review of these is presented in Chapter Four to provide a benchmark for positive definite schemes, and also to try to find an avenue by which some of the high order schemes developed in Chapter Three may be made positive definite.

Chapter Five details the development of high order implicit positive definite schemes. By combining the approach of Boris and Book (1973) and Zalesak (1979) with the new methods developed in Chapter Three, new schemes are obtained that preserve the accuracy of the high order schemes but remove the associated numerical oscillations. For some time it was considered that this approach could not succeed, but by careful consideration of all terms in the finite difference equation, a highly accurate implicit and smooth finite difference scheme can be obtained. Possible improvements to this new scheme are also discussed. There are essentially three components to the new scheme, the high order solution, the low order solution and the process by which the two are combined. Each of these will be dealt with in turn.

Chapter Six discusses additional numerical tests of the new method developed in Chapter Five for the case where the velocity field varies in space. This chapter also includes a discussion on the use of this new method in a model of gas discharges developed by Morrow (1985), demonstrating the success of this scheme in overcoming the difficulties mentioned earlier. A discussion on avenues for future research concludes this thesis.

Chapter 2

Standard Explicit Finite Difference Schemes

In this chapter, different approaches to obtaining accurate finite difference schemes are discussed. Comparing some standard explicit schemes that result from these approaches provides a benchmark for testing other schemes. The limitations of these different derivations in producing higher order schemes will be discussed. It will also be shown how the modified equivalent equation and wave propagation characteristics may be used to derive most of these schemes and so obtain a strategy for developing schemes of arbitrary order. An important part of modelling is the computational effort expended in obtaining results. The balance between accuracy and speed depends very heavily on the particular application. The length of the model time, the speed of the computer being used and the number of grid-points in the computational mesh must all be taken into consideration, and as computers become more and more powerful, the balance between speed and accuracy will continue to change. It will be demonstrated in Chapter Two that, for a given level of accuracy, it is more efficient to use a high order scheme than a low order scheme. As mentioned previously, other features of the numerical solution, such as smoothness, may also need to be considered, depending on the particular application. For the moment, however, attention will be focussed on the differences between the numerical schemes and the exact solution, leaving the problem of obtaining smooth results to Chapter Four.

To compare the accuracy of different numerical methods a model problem is required, such as solving the one-dimensional advection equation

$$\begin{aligned} \frac{\partial \hat{\rho}}{\partial t} + \frac{\partial (w\hat{\rho})}{\partial x} &= 0 \\ \hat{\rho}(x,0) &= f(x) \end{aligned} \tag{2.1}$$

for $\hat{\rho}(x,t)$ over the domain $x \in [0,1]$, $t \geq 0$ and where $w = w(x,t)$ is the advective velocity.

Numerical solutions to this equation exhibit the dispersive and/or diffusive errors that are typical of numerical simulations of all advective processes. The diffusive errors lead to undue smoothing of the numerical solution and the dispersive errors manifest themselves as oscillations producing many unphysical local extrema. These two different types of errors may even occur at the same time. The reason for their appearance will be seen later to be inherent to any numerical approximation.

A diffusive term is not included in the model equation since physical diffusion can serve to mask the numerical errors; by smoothing out the profile the oscillations are reduced and both the dispersive and diffusive errors are then further reduced by the physical diffusion. The spurious oscillations are damped out, and by smoothing the solution the numerical diffusion is not so obvious. Since a scheme should produce accurate results in as broad a range of conditions as possible, it is worth investigating the case of pure advection, modelled by Eq. (2.1).

The advection process will be modelled on a uniform space-time grid with grid-spacing Δx and time-step Δt. The grid coordinates will be given by $(x_j, t^n) = (j\Delta x, n\Delta t)$ for $j = 1, ..., J$ and $n = 1, ..., N$. The value of any function h at a point (x_j, t^n) on this mesh will be denoted by h_j^n. As in many numerical models, the velocity will be assumed to be given at half grid-points as in the Arakawa-C grid (or "Richardson lattice"). This gives a more accurate representation of the physical process as the grid values of w now represents the velocity of the material being advected between neighbouring grid-points. It is trivial, however, to change the difference equations for the case where the velocity is given at the grid-points.

To avoid errors due to the approximation of boundary conditions, cyclic boundary conditions will be used in the numerical tests discussed in this chapter. With cyclic boundary conditions, the comparisons between the different schemes are more representative of the relative performances in general. Since a widely applicable scheme is desired, it is reasonable to determine the general shortcomings of each scheme without reference to its application in specific cases.

To fully understand a numerical scheme it is essential to know how accurately the numerical solution approximates the true solution at the mesh points. From this, the cause of errors and their severity can be determined. This can be done in two ways. One method is to compare the results of numerical experiments, the other is to measure particular aspects of the numerical solution, such as how much dispersion and diffusion is introduced by the scheme and how well it models the true group velocity for a range of Fourier components. Each approach has its disadvantages. Numerical experiments only show how well the scheme performed with a particular initial condition (other initial conditions may produce different results), but they do serve to illustrate the theoretical findings. Alternatively, by looking at the distortion of different Fourier components by the numerical scheme, it is possible to see how well the scheme performs for general initial conditions, but only in the case of constant velocity, that is when solving

$$\frac{\partial \rho}{\partial t} + w\frac{\partial \rho}{\partial x} = 0 \tag{2.2}$$

where $w = u$, a positive constant.

Throughout this thesis, w will be used when the velocity may vary and u when the velocity is assumed constant.

It is also difficult to see the interplay between the different types of distortions. For example, some schemes do not transmit short Fourier components well, introducing a lag and damping, yet other schemes may only introduce a lag. It may not be obvious which is the better scheme, since the damping of small oscillations can stop the development of the spurious oscillations introduced by the phase lag. This

interaction between the various deformations is only shown by certain numerical tests. Used in conjunction, however, these two approaches give an indication of how a numerical scheme will perform in most problems.

To quantify the overall accuracy of a scheme it must be tested on a particular problem, such as the widely used test case of advecting a wavetrain of "Gaussian [1]" pulses at constant velocity. Later, other tests will also be used to illustrate certain features. A Gaussian pulse is useful as it possesses steep gradients, and a sharp peak. (Throughout this work, *"pulse"* shall be used to denote the body of the waveform, and the *"peak"* is defined to be the region within a few gridpoints of the maximum of the pulse.) There are, however, deficiencies with this test. For example, initial conditions may well not be infinitely differentiable in practice, and so there is some value in using an initial condition that is not so smooth. In later chapters, semi-elliptical pulses will be used to demonstrate specific failings of certain methods. Initially, the problem of the simple advection of infinitely differentiable pulses will be solved, before other complications are dealt with.

2.1 Wave Propagation Parameters

As mentioned earlier, it is useful to have analytic quantitative comparisons as well as numerical results to compare more readily how certain aspects of advection are modelled. In order to do this, formulae are needed that relate the numerical propagation of Fourier components to their analytic propagation. This can really only be done in a constant velocity field. The main features of the advection process which are of concern involve the amplitude, phase speed and group velocity of the numerical waves. To derive relationships between the numerical and the analytic propagation of waves, consider a Fourier component of wavelength λ being propagated by Eq. (2.1) and Eq. (2.2). By defining the non-dimensional wavelength and

[1]While it is acknowledged that functions of the form $\exp(-x^2/a)$ have only a slight connection with C.F. Gauss, it has become standard practice in the literature to refer to such functions as Gaussian.

wavenumber to be $N_\lambda = \lambda/\Delta x$ and $\beta = 2\pi/N_\lambda$, respectively, one obtains the exact solution to the difference equation, $\rho_j^n = G^n \exp\{i\beta j\}$, $i = \sqrt{-1}$ which may be compared with the exact solution of the differential equation Eq. (2.2), namely $\hat{\rho}_j^n = \exp\{2\pi i(x_j - wt^n)/\lambda\}$. The complex amplitude or the *"von Neumann amplification factor"* (O'Brien at al. 1950), G, depends on β as well as any constant appearing in the difference equation, such as the Courant number $c = w\Delta t/\Delta x$. An expression is obtained for G directly by substitution into the difference equation. Noye (1987) has shown that the numerical phase speed of a Fourier component is given by

$$u_N = -\frac{N_\lambda \Delta x}{2\pi \Delta t} \arg\{G(c,\beta)\} \quad , \tag{2.3}$$

so the ratio to the exact phase speed (u) is given by the relative phase speed,

$$\mu(c, N_\lambda) = \frac{u_N}{u} = -\frac{N_\lambda}{2\pi c} \arg\{G(c,\beta)\} \tag{2.4}$$

and the ratio to the numerical amplitude of the component to the true amplitude (unity) is the amplitude response

$$\zeta(c, N_\lambda) = |G(c,\beta)|^{N_\lambda/c} \quad . \tag{2.5}$$

An expression for the numerical group velocity is found by differentiating Eq. (2.3). Comparing this with the exact group velocity gives the result obtained by Cathers and O'Connor (1985) for the ratio of the numerical group velocity to the true group velocity, namely the relative group velocity

$$\begin{aligned} \gamma(c, N_\lambda) &= \left(\frac{\partial \beta u_N}{\partial \beta}\right) \Big/ \left(\frac{\partial \beta u}{\partial \beta}\right) \\ &= \frac{-\mathcal{R}e(G)\frac{\partial}{\partial\beta}[\mathcal{I}m(G)] + \mathcal{I}m(G)\frac{\partial}{\partial\beta}[\mathcal{R}e(G)]}{c|G|^2} \quad . \end{aligned} \tag{2.6}$$

In the particular case where $|G| = 1$ the above equation reduces to

$$\gamma(c, N_\lambda) = \frac{\frac{\partial}{\partial\beta}[\mathcal{R}e(G)]}{c\,\mathcal{I}m(G)} \quad . \tag{2.7}$$

2.1.1 Modified Equivalent Equation (MEPDE)

Another analytic comparison which shows how successfully the advection process is being modelled, is to determine how closely the finite difference equation resembles the advection equation, Eq. (2.2). For sufficiently differentiable functions this can be done by means of the *modified equivalent partial differential equation* (or MEPDE) as follows. Firstly, the the equivalent partial differential equation (p.d.e.) is obtained by replacing the discrete values ρ_j^n by the continuous function $\rho(x_j, t^n)$ and expanding each term of this equation about the point (x_j, t^n). The solution of this p.d.e. agrees with the difference equation at the gridpoints and will involve both spatial and temporal derivatives. From this, the modified equivalent equation may be obtained by repeatedly differentiating the equivalent p.d.e. and then cancelling various derivatives leaving an equation of the form

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \sum_{q=2}^{\infty} \frac{u(\Delta x)^{q-1}}{q!} \eta_q(c) \frac{\partial^q \rho}{\partial x^q} \tag{2.8}$$

where the functions $\eta_q(c)$ depend on the particular difference equation. For convenience, η_q will be used in place of $\eta_q(c)$, with the dependence on c being assumed. The term on the right hand side will be known as the *residual* of the modified equivalent equation.

This process was first described by Warming and Hyett (1974) and modified and expanded by Noye and Hayman (1986). In the following, a scheme is referred to as being of order k if η_q is identically zero for all $q \leq k$ and η_{k+1} is some non-zero function; that is, it approximates the original p.d.e. with an error $O\{(\Delta x)^k\}$. There is also the connection between the modified equivalent equation and the wave propagation parameters mentioned previously. It should be clear that this process assumes that the initial conditions (and hence the solution from then on) are everywhere infinitely differentiable. While this is not generally the case, it will be assumed for this chapter; the problems that arise when the initial conditions are not differentiable in some places will be discussed in Chapter Three.

It will be shown that if a scheme is of order k, where k is odd, then the difference

scheme has an amplitude response of the form $1 + O\{N_\lambda^{-k}\}$ as $N_\lambda \to \infty$, that is, the amplitude response of well resolved Fourier components is approximated with error $O\{(\Delta x)^k\}$, and that the numerical phase speed errors are at most $O\{(\Delta x)^{k+1}\}$. Similarly, it will be shown that for a scheme of even order, these components possess a numerical phase speed that is in error by terms of $O\{(\Delta x)^k\}$ and that the amplitude response is approximated with error are at most $O\{(\Delta x)^{k+1}\}$. So a connection exists between the leading error term of the modified equivalent equation and the dominant source of error. From this, it is clear that the odd terms in the modified equivalent equation contribute to the phase error and the even terms are associated with amplitude errors.

This approach also gives the minimum number of points needed to produce a scheme of order k, namely $k + 2$. This is because the Taylor series expansion of a difference scheme has as many degrees of freedom as the modified equivalent equation. To obtain a scheme of order k, the coefficients of ρ, $\partial^2\rho/\partial x^2$, ..., $\partial^k\rho/\partial x^k$ at (x_j, t^n) must all be set to zero and the coefficients of $\partial\rho/\partial t$, $\partial\rho/\partial x$ must match those in Eq. (2.2) , giving $k + 2$ degrees of freedom. Alternatively, it gives the maximum order possible for a scheme with a particular stencil.

As the modified equivalent equation was obtained by algebraic manipulation of the original Taylor series expansion, it equals the solution to the difference equation at the gridpoints. Now if the initial condition for the modified equivalent equation is a single Fourier component of wavelength λ, i.e. $\rho(x, 0) = \exp(i2\pi x/\lambda)$, then the solution is

$$\rho(x, t) = G^{t/\Delta t} \exp(i\beta x/\Delta x) \quad . \tag{2.9}$$

The solution to the difference equation is given by $\rho_j^n = G^n \exp(i\beta j)$. Substitution of Eq. (2.9) into the modified equivalent equation, along with the result $d(G^t)/dt = (\ln G)G^t$, gives

$$\ln G = -ic\beta + c\sum_{q=2}^{\infty} \eta_q \frac{(i\beta)^q}{q!} \quad . \tag{2.10}$$

Writing $Z = G^{2\pi/(\beta c)}$, then the amplitude response of the difference scheme is given

by $\zeta = |Z|$. It then follows that

$$\begin{aligned} lnZ &= \frac{2\pi}{\beta c} lnG \\ &= -2\pi i + \frac{2\pi}{\beta} \sum_{q=2}^{\infty} \eta_q \frac{(i\beta)^q}{q!} \end{aligned} \tag{2.11}$$

or

$$Z = \exp\left(2\pi i \sum_{q=2}^{\infty} \eta_q \frac{(i\beta)^{q-1}}{q!} \right) \quad . \tag{2.12}$$

Suppose that η_m is the first non-zero even coefficient in the series, then only the odd coefficients η_q are non-zero, and so $\eta_q (i\beta)^{q-1}$ is real for $q < m$, making

$$\theta = 2\pi \sum_{q<m} \eta_q (i\beta)^{q-1}/q! \tag{2.13}$$

real. Hence

$$Z = e^{i\theta} \exp\left[\frac{2\pi i \beta (i\beta)^{m-1}}{m!} \eta_m + 2\pi i \sum_{q>m} \frac{(i\beta)^{q-1}}{q!} \eta_q \right] \tag{2.14}$$

which, when expanded asymptotically, gives

$$\zeta = |Z| \sim 1 + 2\pi i^m \frac{\beta^{m-1}}{m!} \eta_m + O\{\beta^m\} \quad as \quad \beta \to 0 \quad . \tag{2.15}$$

If a scheme is order k then the modified equivalent equation is

$$\frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^k}{(k+1)!} \eta_{k+1} \frac{\partial^{k+1} \rho}{\partial x^{k+1}} + \sum_{q>k+1}^{\infty} \frac{u(\Delta x)^{q-1}}{q!} \eta_q \frac{\partial^q \rho}{\partial x^q} \tag{2.16}$$

and by definition, $m \geq k+1$. Hence, Eq. (2.15) implies that the amplitude response is at least order k. If a scheme is of odd order then the leading term is such that m equals $k+1$, and the leading error term is associated with amplitude errors.

In a similar fashion, it can also be shown that removal of terms from the modified equivalent equation also gives higher order approximations to the phase speed. The relative phase speed is given by $\mu = -\arg\{G\}/(c\beta)$ and Eq. (2.10) gives

$$G = \exp\left[-ic\beta + c \sum_{q=2}^{\infty} \frac{(i\beta)^q}{q!} \eta_q \right] \quad . \tag{2.17}$$

Denoting the first non-zero odd coefficient by η_m and redefining θ by

$$\theta = c \sum_{q<m} \eta_q (i\beta)^q / q! \tag{2.18}$$

which is real, gives

$$G = e^{\theta} \exp\left[-ic\beta \pm \frac{ic\beta^m}{m!}\eta_m + \sum_{q>m} \frac{(i\beta)^q}{q!}\eta_q\right] \quad , \tag{2.19}$$

the two solutions depending on whether $(m-1)/2$ is even or odd, respectively. Asymptotic expansion in β gives

$$\mu \sim 1 \pm \frac{\beta}{m!}^{m-1}\eta_m + O\{\beta^m\} \; as \; \beta \to 0 \tag{2.20}$$

showing that a difference scheme with modified equivalent equation of order k will also propagate Fourier components to within a factor of order k. Note that this analysis describes what happens as β tends to 0 which corresponds to λ (or N_λ) approaching ∞ and gives no information on how well the very short wave components are propagated.

Not only do wave propagation parameters, the MEPDE and numerical experiments give a comprehensive means of comparing different schemes, but the link between the first two gives the opportunity to design schemes with various properties. Noye (1987) has shown the link between a given computational stencil and the possible MEPDE's for that stencil and has used that to derive schemes which minimize (in some sense) either the numerical diffusion or dispersion present. The task here is to really minimize both, since the goal is to design a positive definite scheme. Using the MEPDE approach, however, restricts the choice to purely linear constant coefficient schemes. If only explicit schemes are considered, limits are imposed by the result of Godunov (1959) to a scheme with no less diffusion than first order upwinding, which will be shown to be unsatisfactory for the purposes of this argument. The equivalence between terms in the MEPDE and the wave propagation parameters will still be useful, since apart from the problem of smoothness, the higher the order of the residual terms in the wave propagation parameters, the more accurate in a broader class of problems is the numerical scheme. So, given any appropriate computational stencil, it is possible to construct a difference equation

that maximizes the order of the wave propagation parameters, i.e. maximize k in

$$\zeta = 1 + 2\pi i^k \frac{\beta^k}{(k+1)!}\eta_{k+1} + O\{\beta^{k+1}\} \tag{2.21}$$

for k odd, or

$$\mu = 1 \pm \frac{\beta^k}{(k+1)!}\eta_{k+1} + O\{\beta^{k+1}\} \quad . \tag{2.22}$$

for k even.

Since it is not possible to examine every stencil (as they are infinite in number), only two-level methods will be examined here. This still leaves considerable scope and has the advantage of including many schemes that are unconditionally stable (in the von Neumann sense)[2]. With two-level methods, it is relatively easy to ensure unconditional stability, whereas once data from more than two distinct time levels is included, it is not so simple. As shall be seen later, the unconditional stability of some schemes has considerable advantages.

2.2 Upwind Difference Schemes

In the remainder of this chapter, various well known finite difference schemes are analyzed in the light of the above measures of performance. Some of the schemes discussed here may be developed by many different approaches, however, only the approaches that best lend themselves to obtaining higher order schemes will be presented. This is not intended to be an exhaustive classification of all finite difference schemes but rather an attempt to disclose the failings of, and describe the attempts to improve, the commonly used methods. Some less commonly used schemes are included to illustrate how they fit into the general development of the high order schemes described in the next chapter.

The discussion begins with the basic techniques and points out their successes and failures. The precise nature of the errors of each scheme must be determined in order to compare the strengths and weaknesses of the different schemes.

[2]From hereon "stability" will be assumed to be in the von Neumann sense unless explicitly stated otherwise.

The comparisons will be made on the basis of MEPDE's, wave propagation parameters and numerical tests. The value of the first two was discussed earlier. The latter provides a guide to the amount of computational time required to obtain the solution, as well as being necessary to demonstrate the impact of high frequency waves in the initial conditions. There is also much value in an illustration of a numerical solution as this is often the best way to appreciate the effects of dispersion and diffusion which, to some extent, cancel each other.

There are many ways of deriving the class of finite difference schemes known as *upwind differencing*, one of which is based on following the *characteristics* for Eq. (2.2) given by $x = ut + C$ for any arbitrary constant C. Along this path, ρ is constant, giving $\rho(x, t + \Delta t) = \rho(x - u\Delta t, t)$. Thus the problem of approximating $\rho(x_j, t^{n+1})$ reduces to one of estimating $\rho(x_j - u\Delta t, t^n)$. This can be done by Lagrange interpolation between the values at time t^n, with different knots for the interpolation giving different schemes. This will be seen to be equivalent to successively removing terms from the modified equivalent equation.

To construct a k^{th} order upwind scheme requires a k^{th} order polynomial, involving $(k+1)$ values at time t^n, which is used to interpolate the value of $\rho(x_j - u\Delta t, t^n) = \rho_j^{n+1}$. This gives a stencil involving $k+2$ points, which is consistent with the modified equation approach. The selection of these k points leads to the different classes of upwind schemes. There is no restriction on how these points are chosen, except that if the interpolation is through the points $x_{j-l}, ..., x_j, ..., x_{j+m}$ then $m \leq l$ for stability. The simplest argument for this is based on physical reasons, namely that since information is advected from upstream to downstream points, it does not seem logical to incorporate more downstream information than upstream information. It should also be noted that when extended to the case where velocity $w = w(x, t)$, it is then assumed that the velocity is locally constant within the interval (x_{j-1}, x_j) so that the scheme is no longer strictly order k. This is also consistent with the modified equation approach as this corresponds to terms involving derivatives of w

entering the modified equivalent equation.

In all of the cases discussed in this thesis, the modified equation approach gives an identical difference equation to that obtained by upwind differencing, if the same stencil is used. This is as expected, since both Lagrange interpolation and the modified equation approach are based on removing successive terms of the Taylor series expansion of the difference equation. In fact, the author and associates have always found there to be a unique scheme of order k for any given stencil involving $k+2$ points, although no proof has been found that this is always true. Provided that there is a unique scheme of order k for a stencil of $k+2$ points (such as is the case for the stencils described here), then by the equivalence of the modified equivalent equation and the wave propagation parameters, it then follows that upwind schemes have minimal phase and amplitude error. In this case, the minimum is taken over all finite difference schemes defined on the same stencil as the particular upwind scheme.

2.2.1 First Order Upwinding

If linear interpolation between the points x_j and x_{j-1} is used to estimate $\rho(x_j - u\Delta t, t^n)$, then the interpolating polynomial is

$$P_1(x) = \rho_j^n + (x - x_j)(\rho_j^n - \rho_{j-1}^n)/\Delta x \quad . \tag{2.23}$$

Since $\rho_j^{n+1} = \rho(x_j - u\Delta t, t^n) \simeq P_1(x_j - u\Delta t)$, the resulting finite difference equation is

$$\rho_j^{n+1} = \rho_j^n - c(\rho_j^n - \rho_{j-1}^n) \tag{2.24}$$

where c, the Courant number, is constant. The modified equivalent equation for this scheme is then

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u\Delta x}{2}(c-1)\frac{\partial^2 \rho}{\partial x^2} + O\{(\Delta x)^2\} \quad . \tag{2.25}$$

Writing Eq. (2.24) in flux form, namely

$$\rho_j^{n+1} = \rho_j^n - f_{j+\frac{1}{2}} + f_{j-\frac{1}{2}} \tag{2.26}$$

where $f_{j+\frac{1}{2}} = c_{j+\frac{1}{2}}\rho_j^n$, gives a scheme suitable for use with variable velocity, namely

$$\rho_j^{n+1} = \rho_j^n - c_{j+\frac{1}{2}}\rho_j^n + c_{j-\frac{1}{2}}\rho_{j-1}^n \quad . \tag{2.27}$$

When the velocity, w, varies with time, $c_{j+\frac{1}{2}}$ should be calculated from a value of $w_{j+\frac{1}{2}}$ that is representative of the velocity over the interval $[t^n, t^{n+1}]$. The simplest form is an arithmetic mean, but other choices are available, depending on the particular problem at hand and user preference. The value of $c_{j+\frac{1}{2}}$ will be assumed to be given by such a representative value throughout this thesis.

For the remainder of this thesis, only the flux-form will be given, although the MEPDE and wave propagation parameters refer to the special case $w = u$, a constant. In all numerical results, it will always be the flux form that is used, even if the velocity is constant in the experiment. This is because the Central Processor Unit (CPU) time required by the scheme will naturally differ, depending on how the scheme is coded. Since it is the flux form that is used in practice, it is appropriate to quote times using this form rather than the constant velocity form. This has no effect on the accuracy of the scheme since, if the velocity is constant in a numerical experiment, the two formulations give identical results.

The difference equation Eq. (2.27) is still often used because of its simplicity and the fact that it guarantees smooth results. In many applications it is crucial that the numerical solution is smooth. A linear difference scheme is guaranteed to preserve smoothness in the data if it preserves monotonicity, which is equivalent to ρ_j^{n+1} being between ρ_{j-1}^n and ρ_j^n (Boris and Book, 1973). Stability requires $c \leq 1$ and thus this method must give a value of ρ_j^{n+1} which satisfies the smoothness condition. Unfortunately, this is the sole advantage of using this method. An example of the performance of this scheme is shown in Fig. 2.1(d), where the initial condition was a Gaussian pulse of the form $\exp\{-400(x-0.5)^2\}$. The pulse was advected at a speed $u = 0.8$ for 10 periods with a grid-spacing of 0.01, and a Courant number of 0.4. This test will form the basis for comparing most of the schemes discussed in this chapter. The excessive damping due to this scheme is readily apparent with

the numerical solution (the solid line) being so diffuse that only the last vestiges of the pulse remain. That such behaviour is typical of this scheme can be seen from Fig. 2.1(a) illustrating the amplitude response of a range of Fourier modes for a variety of Courant numbers. Modes of wavelength up to fifty grid-spacings are still significantly damped after only one period, and so it is not surprising that after ten periods all but the longest modes have been damped out. The phase speed of these components is better although it could still be improved. This is not surprising, as it can be seen from the modified equivalent equation that the amplitude response is of the form $1 + O\{N_\lambda^{-1}\}$ but the relative phase speed has form $1 + O\{N_\lambda^{-2}\}$, i.e. the amplitude response has first order errors but the errors in phase speed are second order.

The relative phase speed of this scheme is shown in Fig. 2.1(b) for completeness, along with the relative group velocity in Fig. 2.1(c). While the relative phase speed is mediocre in comparison with schemes discussed below, the considerable damping of the short Fourier components means that they are eliminated from the numerical solution and so it is immaterial how accurately they are propagated. For this reason the position of the numerical peak relative to the true peak is quite accurate, since only the components that have a relative phase speed near unity survive the numerical damping.

For a process to be well-modelled by this scheme the grid-spacing should be less than about one-fiftieth of the smallest significant component. So although this scheme produces results very quickly (one multiplication and one addition per time-step per grid-point), to produce accurate results requires a very high number of grid-points and hence time-steps, since for stability the Courant number c is required to be no greater than one. It will be shown at the end of the following chapter that the increase in resolution required to obtain accurate results makes the scheme rather inefficient in terms of absolute computational resources required and as such, the scheme is of little practical use as it stands. The one feature in its favour

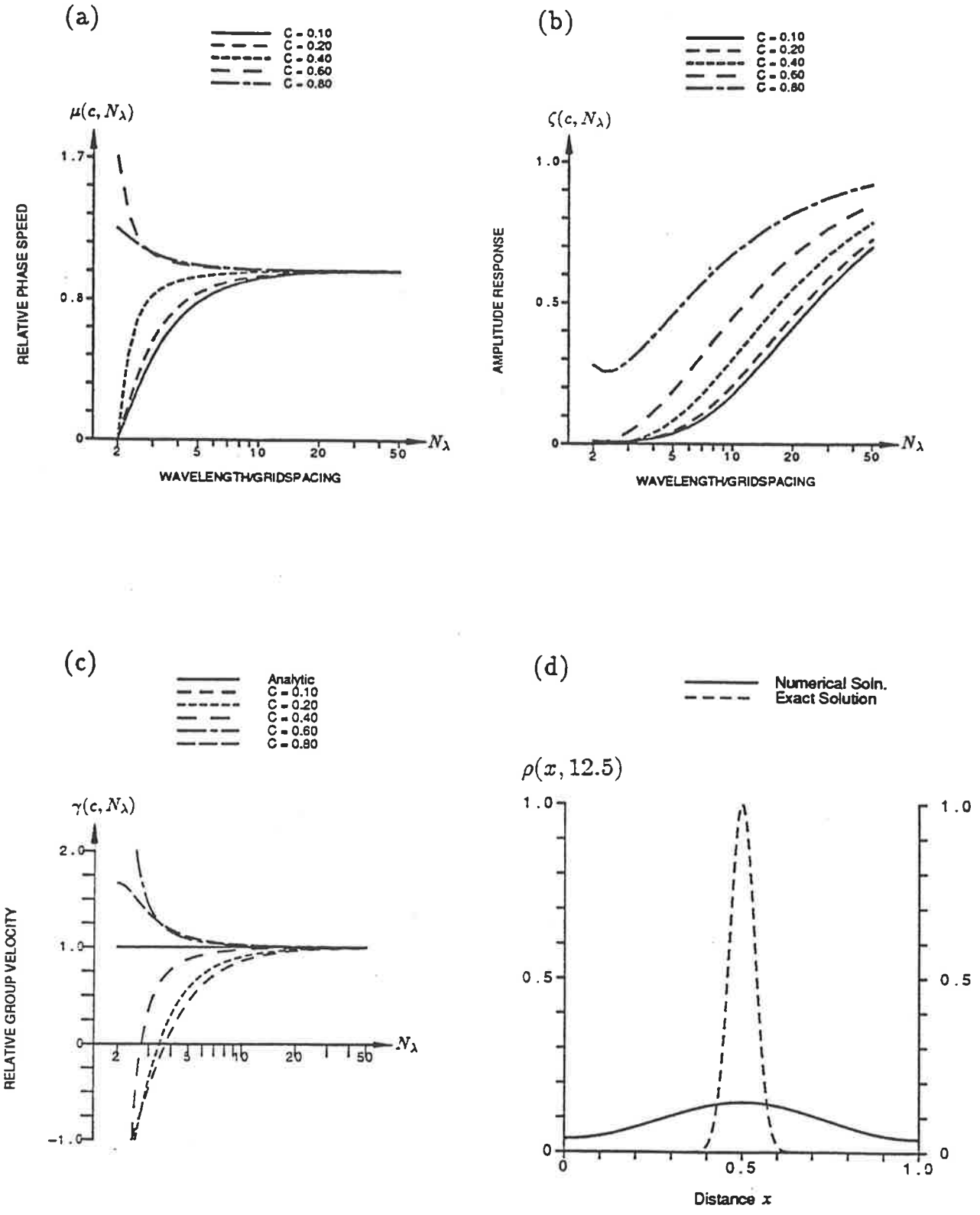

Figure 2.1: Illustration of the performance of First Order Upwinding, Eq. (2.27). The diagrams show (a) the relative phase speed, μ, (b) the amplitude response, ζ, (c) the relative group velocity, γ, and (d) the results of the the numerical test case with a Gaussian pulse as the initial condition and cyclic boundary conditions. This scheme is stable for $c \leq 1$.

is that the solution is guaranteed to be smooth. Godunov (1959) showed that this was the highest order explicit two-level linear scheme with this property. As this scheme is obviously unacceptable, and in the absence of any better smooth implicit solutions, considerable effort was spent on developing non-linear schemes, as discussed in Chapter Three. One of the characteristics of all these non-linear methods is that they generally require a high order scheme to blend with a positive definite (and therefore low order) scheme. As the low order scheme will always be very poor, its contribution to the solution should be minimized, which requires the high order scheme to be as accurate as possible. It will also be shown that there is a direct relationship between a scheme being high order and highly accurate. For this reason, and because higher order schemes turn out to be more efficient at obtaining accurate results, the rest of this chapter and the next will be devoted to developing higher order schemes.

2.2.2 Lax-Wendroff

This is another scheme which has many derivations. One approach is to consider it as a second order form of the scheme discussed above, which uses quadratic interpolation between the points (x_{j-1}, t^n), (x_j, t^n) and (x_{j+1}, t^n) to estimate $\rho(x_j - w\Delta t, t^n)$. This interpolating polynomial is now given by

$$P_2(x) = P_1(x) - \frac{(x - x_j)(x - x_{j-1})}{2(\Delta x)^2}(\rho^n_{j+1} - 2\rho^n_j + \rho^n_{j-1}) \tag{2.28}$$

giving the difference equation (written in flux form)

$$\begin{aligned}\rho^{n+1}_j = \rho^n_j \quad &- \quad (c_{j+\frac{1}{2}}\rho^n_j - c_{j-\frac{1}{2}}\rho^n_{j-1}) - \frac{c_{j+\frac{1}{2}}(1 - c_{j+\frac{1}{2}})}{2}(\rho^n_{j+1} - \rho^n_j) \\ &+ \quad \frac{c_{j-\frac{1}{2}}(1 - c_{j-\frac{1}{2}})}{2}(\rho^n_j - \rho^n_{j-1}) \quad .\end{aligned} \tag{2.29}$$

The modified equivalent equation is now

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^2}{3!}c(c^2 - 1)\frac{\partial^3 \rho}{\partial x^3} + O\{(\Delta x)^3\} \tag{2.30}$$

which means that the dominant errors are now phase errors and that the amplitude errors are third order. The increase in order of the amplitude errors over those for

first order upwinding is evident by comparing Fig. 2.1(b) and Fig. 2.2(b).

The scheme may also be derived by noting that the leading error in Eq. (2.25) is

$$\frac{(\Delta x)^2}{\Delta t}\frac{c}{2}(1-c)\frac{\partial^2 \rho}{\partial x^2} \tag{2.31}$$

and that Eq. (2.29) is merely the second order centred-space approximation to this term subtracted from Eq. (2.27). Such a scheme may thus be considered as first order upwinding without the dominant amplitude errors. As can be seen from Fig. 2.2(a-d), this provides little improvement over first order upwinding; while the amplitude errors have been partially corrected, the phase speed errors now dominate. Large oscillations in the numerical results are thereby produced.

2.2.3 Second order upstream biassed differencing

There was no particular reason for choosing the points x_{j-1}, x_j and x_{j+1} to be the knots for the interpolating polynomial. It would have been just as valid to use the points x_{j-2}, x_{j-1} and x_j. This choice of points, leads to the differencing known as *"upstream biassed"* which involves possibly some downstream points but always more upstream points than downstream points. The point x_j is considered to be neither upstream nor downstream, and so in the case of first order interpolation only one scheme is possible. With higher order schemes, differences arise in the location of the interpolation knots as mentioned above. Using the three points x_{j-2}, x_{j-1} and x_j gives an interpolating polynomial

$$P_{2u}(x) = P_1(x) - \frac{(x-x_j)(x-x_{j-1})}{2(\Delta x)^2}(\rho_j^n - 2\rho_{j-1}^n + \rho_{j-2}^n) \quad . \tag{2.32}$$

giving the difference equation (in flux form)

$$\begin{aligned}\rho_j^{n+1} = \; & \rho_j^n \; - \left(c_{j+\frac{1}{2}}\rho_j^n + \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}})}{2}\right)(\rho_j^n - \rho_{j-1}^n) \\ & + \left(c_{j-\frac{1}{2}}\rho_{j-1}^n \pm \frac{c_{j-\frac{1}{2}}(1 \pm c_{j-\frac{1}{2}})}{2}\right)(\rho_{j-1}^n - \rho_{j-2}^n) \quad .\end{aligned} \tag{2.33}$$

For the case of uniform velocity, $w = u$, the MEPDE is

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^2}{3!}(2-c)(1-c)\frac{\partial^3 \rho}{\partial x^3} + O\{(\Delta x)^3\} \quad . \tag{2.34}$$

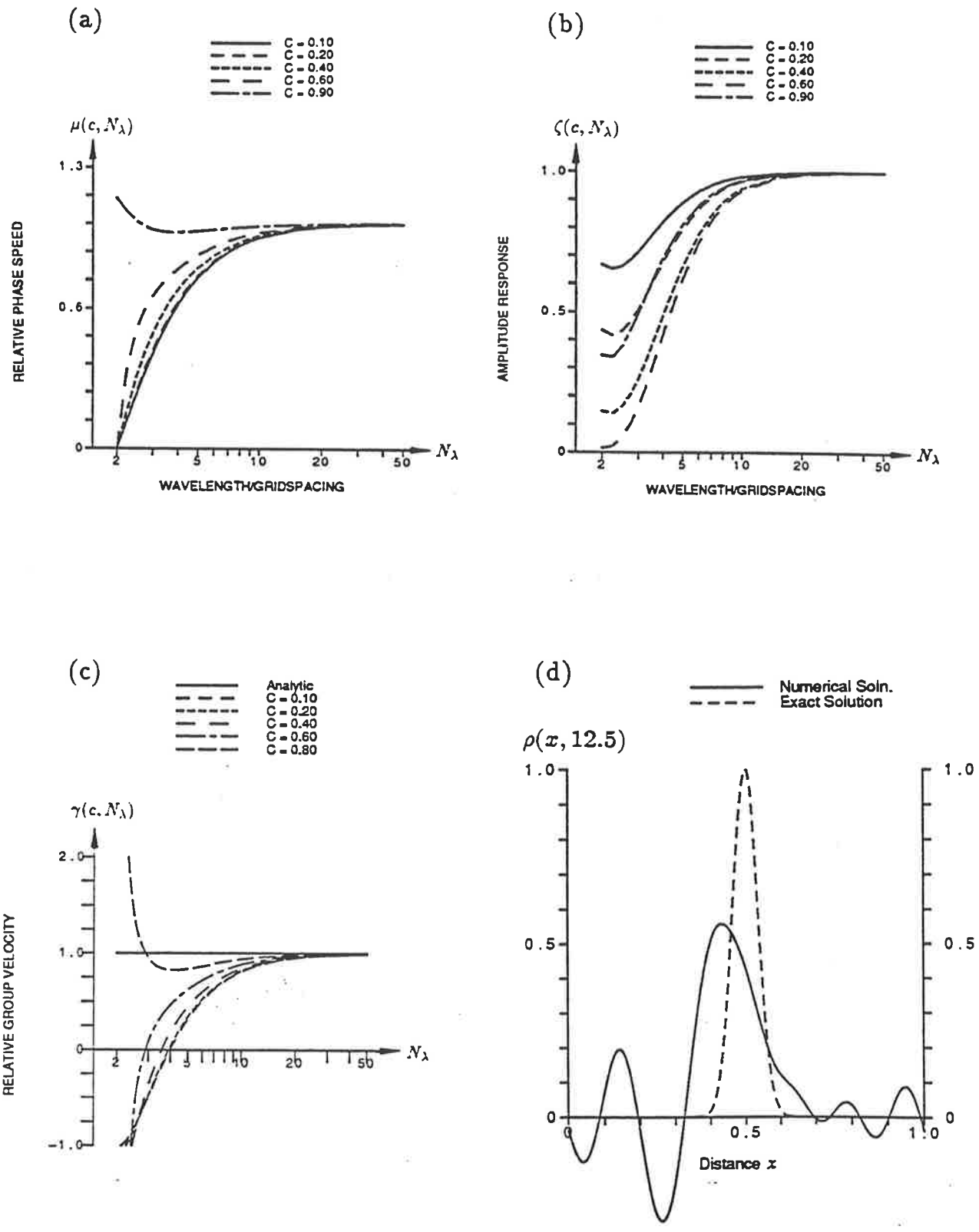

Figure 2.2: As in Fig. 2.1, but for the Lax-Wendroff scheme, Eq. (2.29). This scheme is stable for $c \leq 1$.

The wave propagation parameters and illustration of the results from the Gauss test are shown in Fig. 2.3(a-d). The amplitude response, shown in Fig. 2.3(b) is clearly comparable to that for the Lax-Wendroff scheme Fig. 2.2(b). The main difference between the two schemes is in the relative phase speeds and relative group velocities. For small values of Courant number, the second order fully upstream difference scheme tends to propagate the components and energy too rapidly whereas the Lax-Wendroff is somewhat too slow. This explains why the major negative lobe precedes the pulse in the test example using second order fully upstream differencing, whereas it trails the pulse in the results for the Lax-Wendroff scheme. As far as overall accuracy is concerned, there is little to separate the last two schemes.

By combining the four comparisons (i.e. amplitude response, relative phase speed, relative group velocity and the numerical example) it is possible to draw two conclusions. Firstly, that the poor performance of these three schemes is representative of their general behaviour and not merely an artifact of the particular test problem used, and secondly, that first or second order errors in phase speed or amplitude response lead to unacceptably large errors. It is difficult to say which of these three schemes is the best since none of the numerical solutions particularly resemble the analytic solution. To produce realistic numerical solutions in a wide variety of cases, an investigation of schemes that are at least third order is required.

2.3 High Order Upwind Difference Schemes

In this section, due to the poor performance of first and second order upwind schemes, some high order extensions to upwind differencing are discussed. These schemes are obtained by using higher order polynomials to interpolate the value of $\rho(x_j - w\Delta t, n\Delta t)$, with the order of the polynomial corresponding to the order of the resulting finite difference scheme.

The schemes presented in the previous section are extreme examples of methods dominated by numerical damping (in the case of first order upwinding) and by

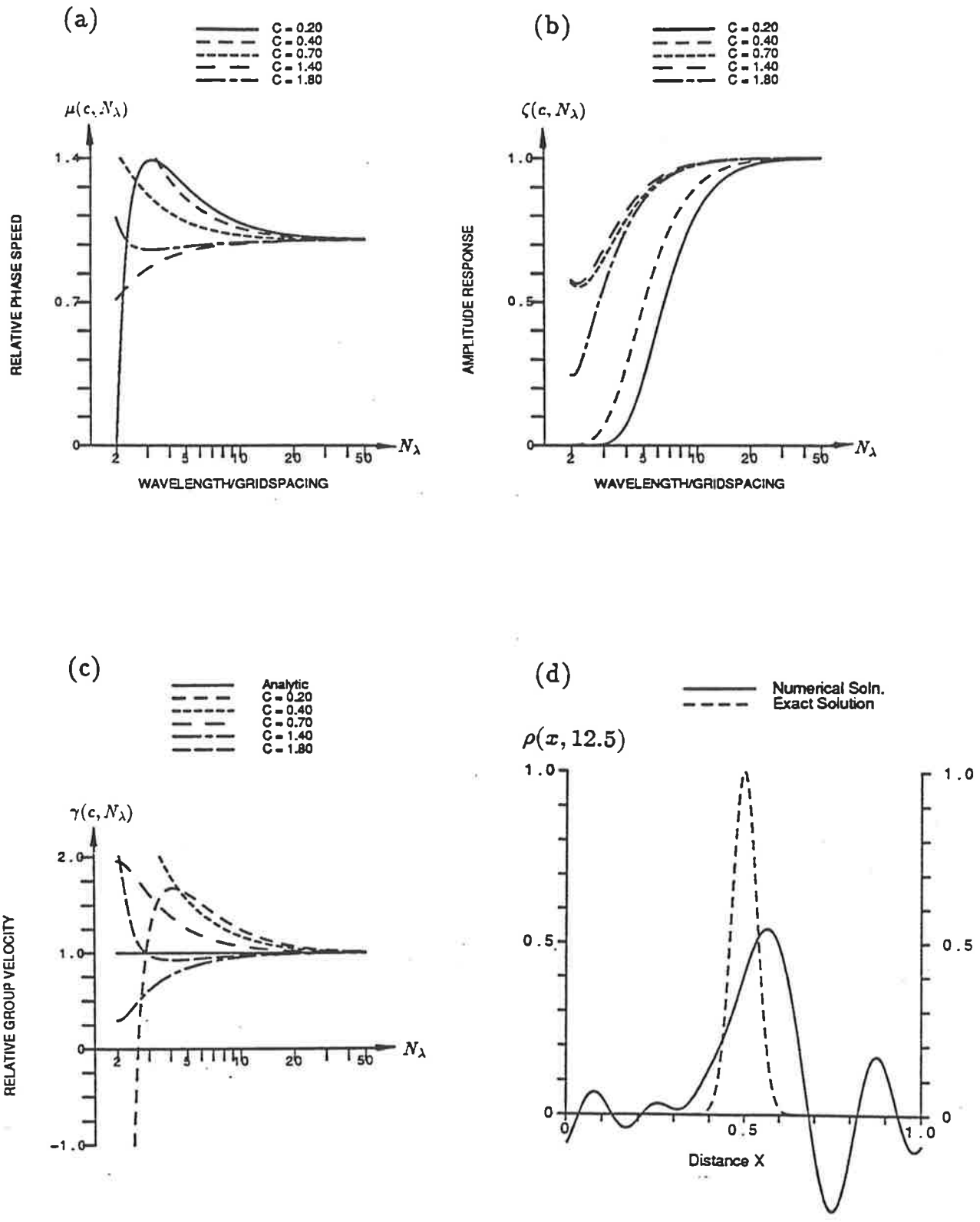

Figure 2.3: As in Fig. 2.1, but for second order upstream differencing scheme, Eq. (2.33). This scheme is stable for $c \leq 2$.

numerical dispersion (as in the case of the second order schemes). Similar errors will also be seen to occur with the odd order schemes (dominated by dispersive errors), although with increasing order these errors decrease. Such behaviour is a direct consequence of their derivation, since for a scheme of order k, the interpolation removes all terms up to order k from the Taylor series expansion of the difference equation, in turn eliminating all terms up to those of k^{th} order from the modified equivalent equation. The leading term in the modified equivalent equation residual is thus $\eta_k(c)(u\Delta x)^k/k!\,\partial^k\rho/\partial x^k$. The equivalence between this term and the leading term in the error of the wave propagation parameters, indicates that the largest errors in a scheme of order k (k odd) will be amplitude errors and if k is even, the major errors will be due to dispersion, both types of errors decaying at a rate of N_λ^{-k}.

It should also be noted that upwind schemes will still contain errors due to both diffusion and dispersion (and experience indicates this is particularly true for very short wavelengths). The overall performance, however, will be dominated by diffusive errors if the scheme is of odd order or dispersive errors if it is of even order. Despite the very short Fourier components still being propagated poorly by these higher order schemes, improving the propagation of the well-resolved components will improve the overall accuracy. This can be seen by examining the illustrations of the numerical tests in the previous section Fig. 2.1(d), Fig. 2.2(d) and Fig. 2.3(d). Consider for the moment, first order upwinding; the major errors in this scheme are due to the damping of the longer Fourier wavelengths. This must be the case, since the amplitude of the short components must be considerably less than those of the longer components, and so although the short components are completely damped out in this case, the loss of amplitude corresponding to the damping of the short wavelength components is insignificant when compared to the loss of amplitude due to the damping of the longer components. Similarly, an examination of the amplitude of the oscillations in Figs.2.2(d) and 2.3(d) shows that the oscillations

were mainly due to the poor phase speed of the intermediate components. This can also be seen by noting that the oscillations are about $20\Delta x$ in wavelength, although care must be taken here as the oscillations are due to components of many wavelengths. Nevertheless, it should be clear that improving the accuracy of the well-resolved components will increase the overall accuracy of the schemes, and hence the higher order schemes will be substantially more accurate.

Another view was given by Leonard (1984) who showed that upwinding schemes of odd order possess "*negative feedback sensitivity*" and so do not suffer as badly from oscillatory errors as even order schemes. For an explicit finite difference scheme of the form

$$\rho_j^{n+1} = \rho_j^n + \sum_i \xi_i^{(j)} \rho_{j+i}^n \tag{2.35}$$

to possess negative feedback sensitivity, the coefficient $\xi_0^{(j)}$ must be less than zero. This leads to a damping of oscillations with a wavelength comparable to the width of the computational stencil. So, if third or fifth order interpolation is chosen, then although the results are dominated by amplitude errors, these errors are third or fifth order in Δx, respectively, and so they should give a considerable improvement on the results seen up to now.

2.3.1 Odd Order Upwinding

As mentioned previously, a polynomial of any order may be used to interpolate the value $\rho(x_j - w\Delta t, t^n)$. These polynomials may be constructed by the straightforward use of divided differences (Kreyszig, 1983). Given any particular stencil, the coefficients of the interpolating polynomial will be unique as these only depend on the interpolation knots, which will coincide with the points in the computational mesh. Depending on which points are chosen, different schemes are obtained. If the interpolation uses the points x_{j-2}, x_{j-1}, x_j and x_{j+1}, the resulting cubic is

$$\begin{aligned} P_3(x) = \rho_j^n \quad &+ \quad \frac{x}{\Delta x}(\rho_j^n - \rho_{j-1}^n) + \frac{x(x+\Delta x)}{2(\Delta x)^2}(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \\ &- \quad \frac{x(x^2-(\Delta x)^2)}{6(\Delta x)^3}(\rho_{j+1}^n - 3\rho_j^n + 3\rho_{j-1}^n - \rho_{j-2}^n) \end{aligned} \tag{2.36}$$

and substituting $x = -w\Delta t$ gives the finite difference equation (in flux form)

$$\begin{aligned}
\rho_j^{n+1} = \rho_j^n \;&-\; c_{j+\frac{1}{2}}\rho_j^n + c_{j-\frac{1}{2}}\rho_{j-1}^n \\
&-\; \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}})}{2}(\rho_{j+1}^n - \rho_j^n) \\
&+\; \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}})}{2}(\rho_j^n - \rho_{j-1}^n) \\
&+\; \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)}{6}(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \\
&-\; \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)}{6}(\rho_j^n - 2\rho_{j-1}^n + \rho_{j-2}^n) \quad .
\end{aligned} \tag{2.37}$$

If instead the points $x_{j-3}, \ldots, x_{j+2}$ are used, the resulting quintic is

$$\begin{aligned}
P_5(x) \;&=\; P_3(x) \\
&-\; \frac{x(x^2-(\Delta x)^2)(x+2\Delta x)}{24(\Delta x)^4}(\rho_{j+1}^n - 4\rho_j^n + 6\rho_{j-1}^n - 4\rho_{j-2}^n + \rho_{j-3}^n) \\
&-\; \frac{x(x^2-(\Delta x)^2)(x^2-4(\Delta x)^2)}{120(\Delta x)^5} \\
&\qquad \times \left(\rho_{j+2}^n - 5\rho_{j+1}^n + 10\rho_j^n - 10\rho_{j-1}^n + 5\rho_{j-2}^n - \rho_{j-3}^n\right)
\end{aligned} \tag{2.38}$$

which gives

$$\begin{aligned}
\rho_j^{n+1} = \rho_j^n \;&-\; c_{j+\frac{1}{2}}\rho_j^n + c_{j-\frac{1}{2}}\rho_{j-1}^n \\
&-\; \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}})}{2}(\rho_{j+1}^n - \rho_j^n) \\
&+\; \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}})}{2}(\rho_j^n - \rho_{j-1}^n) \\
&+\; \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)}{6}(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \\
&-\; \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)}{6}(\rho_j^n - 2\rho_{j-1}^n + \rho_{j-2}^n) \\
&+\; \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)(2-c_{j+\frac{1}{2}})}{24}(\rho_{j+2}^n - 3\rho_{j+1}^n + 3\rho_j^n - \rho_{j-1}^n) \\
&-\; \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)(2-c_{j-\frac{1}{2}})}{24}(\rho_{j+1}^n - 3\rho_j^n + 3\rho_{j-1}^n - \rho_{j-2}^n) \\
&-\; \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)(4-c_{j+\frac{1}{2}}^2)}{120} \\
&\qquad \times \left(\rho_{j+2}^n - 4\rho_{j+1}^n + 6\rho_j^n - 4\rho_{j-1}^n + \rho_{j-2}^n\right) \\
&+\; \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)(4-c_{j-\frac{1}{2}}^2)}{120}
\end{aligned} \tag{2.39}$$

$$\times \left(\rho_{j+1}^n - 4\rho_j^n + 6\rho_{j-1}^n - 4\rho_{j-2}^n + \rho_{j-3}^n\right) \quad .$$

This process of using successively higher and higher order polynomials to interpolate the value $\rho(x_j - w\Delta t,\ n\Delta t)$ can be extended indefinitely. As can be seen above, fifth order upwinding requires use of points at x_{j-3} and x_{j+2} and to go to higher order upwinding requires using points still further away from the point of interest. This introduces complications at the boundaries, and for this reason the width of the computational stencil will not be extended beyond six points. This was given as a reason by Leonard (1984) for not going above third order, let alone fifth order. It is still worth examining fifth order schemes, however, just to see what gains are obtained and at what expense.

The modified equivalent equation for third order upwinding (cubic interpolation) is

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^3}{4!}\left(-2 + c + 2c^2 - c^3\right)\frac{\partial^4 \rho}{\partial x^4} + O\{(\Delta x)^4\} \qquad (2.40)$$

and for fifth order upwinding is

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \qquad (2.41)$$
$$\frac{u(\Delta x)^4}{6!}\left(12 - 4c - 15c^2 + 5c^3 + 3c^4 - c^5\right)\frac{\partial^6 \rho}{\partial x^6} + O\{(\Delta x)^6\} \quad .$$

As expected, these schemes have very good phase speed, the errors being $O\{N_\lambda^{-4}\}$ and $O\{N_\lambda^{-6}\}$, respectively, the amplitude response being one order worse. This is shown again in the numerical results, Fig. 2.4(d) and Fig. 2.5(d) where the distortion of the Gaussian pulse is quite small. The oscillations still persist but the pulse has reasonable shape, especially when compared with the results of previous methods. It should also be noted that the behaviour of these schemes exhibited in Fig. 2.4(d)and Fig. 2.5(d) are typical for Courant numbers less than about 0.4, since as shown in Fig. 2.4(a-c) and Fig. 2.5(a-c), the wave propagation characteristics show little variation with c for Courant numbers of this size. As such, the results presented here are an appropriate representation of the performance of this scheme.

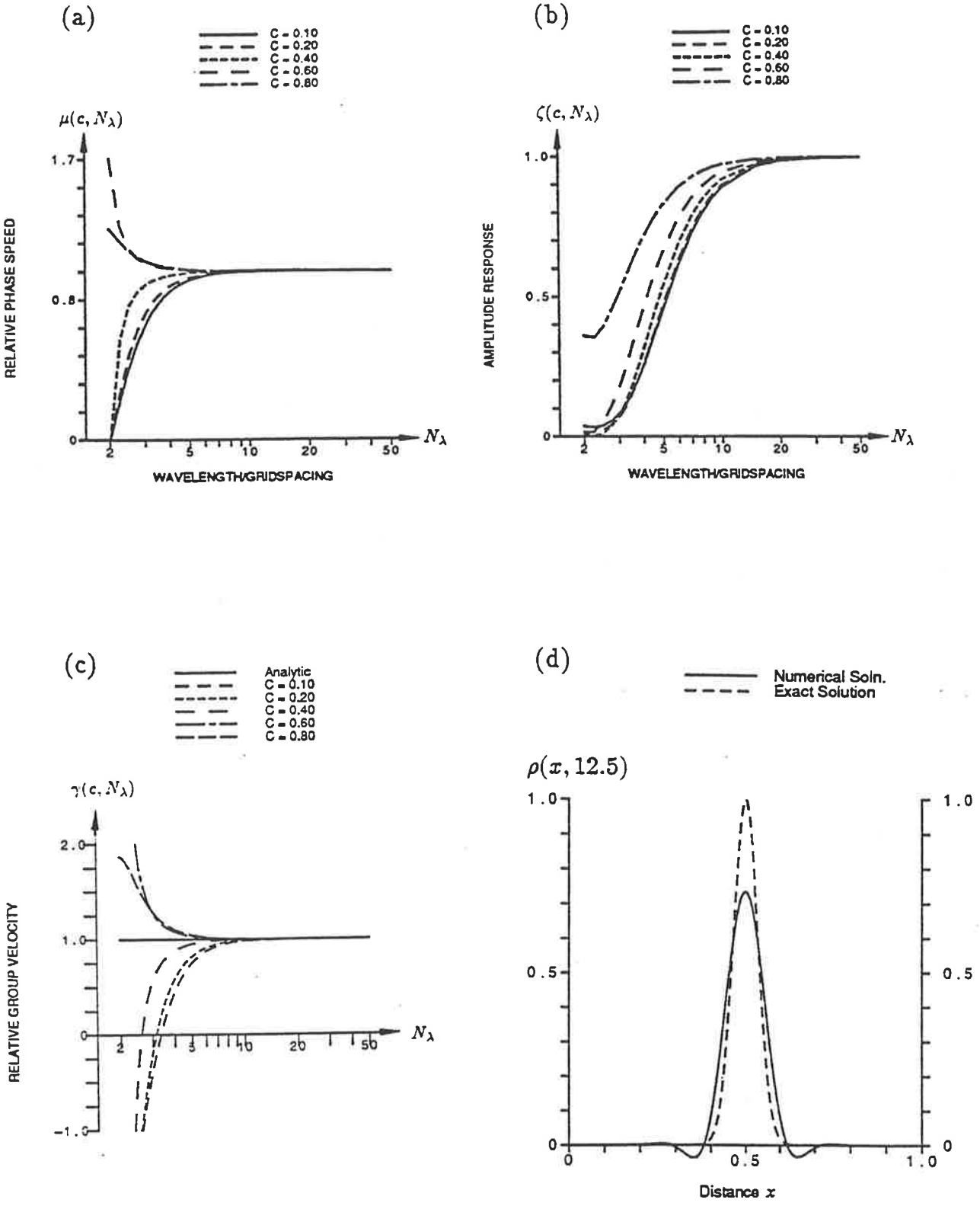

Figure 2.4: As in Fig. 2.1, but for third order upwinding, Eq. (2.37). This scheme is stable for $0 < c \leq 1$.

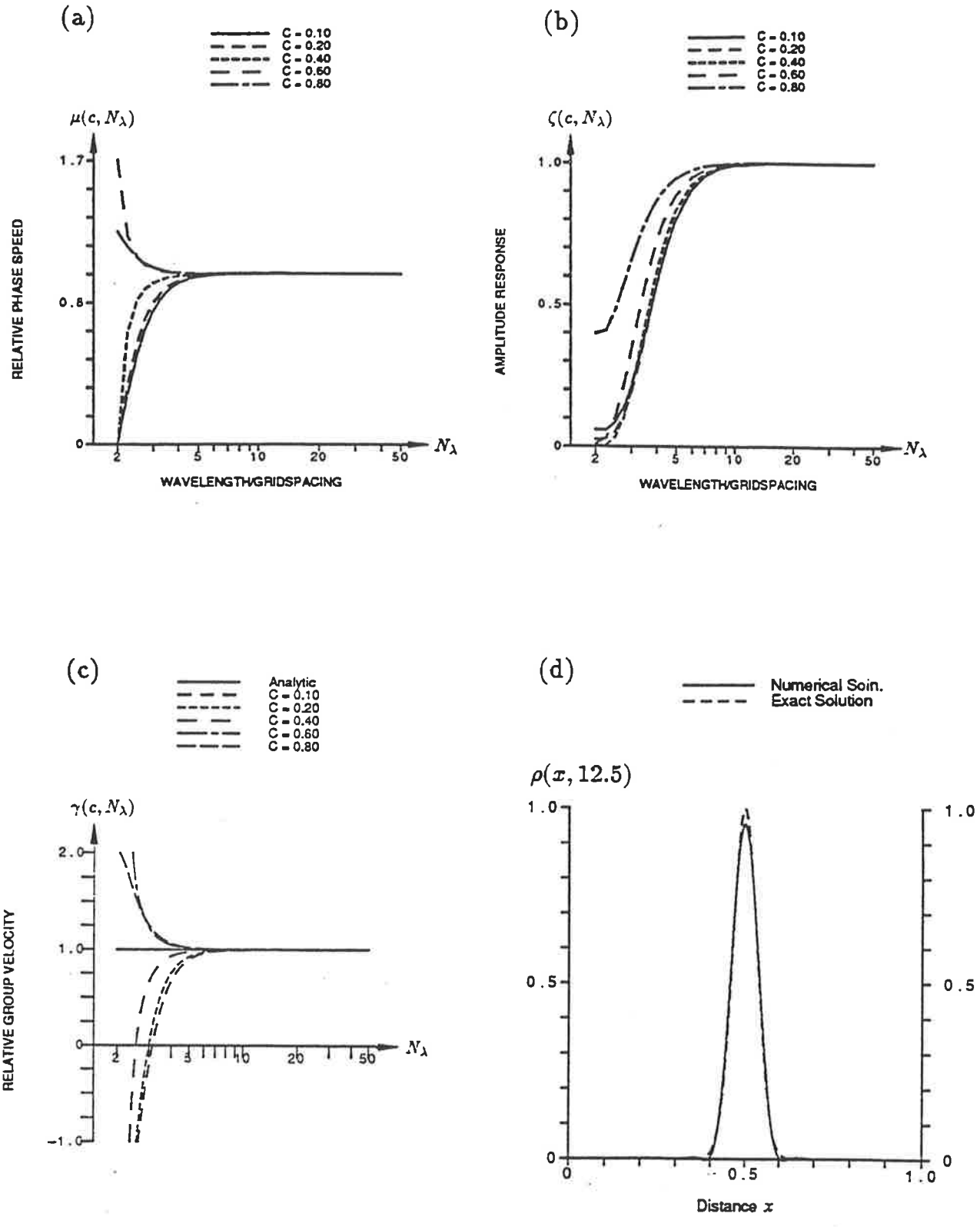

Figure 2.5: As in Fig. 2.1, but for fifth order upwinding, Eq. (2.39). This scheme is stable for $0 \leq c \leq 1$.

There are other choices of knots that will give third and fifth order schemes, obtained by forcing the interpolating polynomial to pass through different points. For example, a third order scheme can be obtained by using the points $x_{j-3}, ..., x_j$, giving the polynomial

$$\begin{aligned} P_{3u}(x) = \rho_j^n \quad &+ \quad \frac{x}{\Delta x}(\rho_j^n - \rho_{j-1}^n) + \frac{x(x+\Delta x)}{2(\Delta x)^2}(\rho_j^n - 2\rho_{j-1}^n + \rho_{j-2}^n) \\ &+ \quad \frac{x(x+\Delta x)(x+2\Delta x)}{6(\Delta x)^3} \\ &\qquad \times (\rho_j^n - 3\rho_{j-1}^n + 3\rho_{j-2}^n - \rho_{j-3}^n) \quad . \end{aligned} \tag{2.42}$$

The resulting finite difference scheme is unstable, as is the scheme that follows from interpolating with the quintic through the points $x_{j-5},,,x_{j+1}$

$$\begin{aligned} P_{5u}(x) \quad &= \quad \rho_j^n + \frac{x}{\Delta x}(\rho_{j+1}^n - \rho_j^n) + \frac{x(x-\Delta x)}{2(\Delta x)^2}(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \\ &+ \quad \frac{x(x^2-(\Delta x)^2)}{6(\Delta x)^3}(\rho_{j+1}^n - 3\rho_j^n + 3\rho_{j-1}^n - \rho_{j-2}^n) \\ &+ \quad \frac{x(x^2-(\Delta x)^2)(x+2\Delta x)}{24(\Delta x)^4}(\rho_{j+1}^n - 4\rho_j^n + 6\rho_{j-1}^n - 4\rho_{j-2}^n + \rho_{j-3}^n) \\ &+ \quad \frac{x(x^2-(\Delta x)^2)(x^2-4(\Delta x)^2)}{120(\Delta x)^5} \\ &\qquad \times (\rho_{j+2}^n - 5\rho_{j+1}^n + 10\rho_j^n - 10\rho_{j-1}^n + 5\rho_{j-2}^n - \rho_{j-3}^n) \quad . \end{aligned} \tag{2.43}$$

This is not surprising since, using these polynomials, the interpolation is consistently being performed towards one end of the interval of interpolation and it is well known (Kreyszig, 1983) that Lagrange interpolation is most accurate near the middle of the interval of interpolation and becomes unstable as the point at which the interpolation is being performed moves towards the extremities of the interval. For the same reason, using the quintic on $x_{j-4}, ..., x_{j+1}$ will also yield an unstable scheme. Furthermore, for reasons mentioned in the introduction to this section, there should be more points upstream of (x_j, t^n) than downstream. This means that the only stable, upwind schemes of order three and five, using equispaced gridpoints are those given by Eq. (2.37) and Eq. (2.39).

2.3.2 Fourth Order Upwind Schemes

Even though only odd order upwinding schemes possess negative feedback, it is worthwhile investigating fourth order schemes, if only to reinforce the conclusions made above. The first observation to be made, is that with even order schemes, there is a choice of possible stencils to use. This was seen with the second order schemes mentioned previously; one uses a "centred" stencil (giving the Lax-Wendroff or Leith scheme), the other an upstream bias in the stencil. When constructing fourth order schemes, a similar choice is possible.

If the points $x_{j-2}, ..., x_{j+2}$ are used, the interpolating polynomial is

$$P_4(x) = P_3(x) + \frac{x(x^2-(\Delta x)^2)(x-2\Delta x)}{24(\Delta x)^4} \times (\rho^n_{j+2} - 4\rho^n_{j+1} + 6\rho^n_j - 4\rho^n_{j-1} + \rho^n_{j-2}) \tag{2.44}$$

where $P_3(x)$ is given by Eq. (2.36). The resulting difference equation (in flux form) is

$$\begin{aligned}
\rho^{n+1}_j = \rho^n_j &- c_{j+\frac{1}{2}}\rho^n_j + c_{j-\frac{1}{2}}\rho^n_{j-1} \\
&- \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}})}{2}(\rho^n_{j+1} - \rho^n_j) \\
&+ \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}})}{2}(\rho^n_j - \rho^n_{j-1}) \\
&+ \frac{c_{j+\frac{1}{2}}(1-c^2_{j+\frac{1}{2}})}{6}(\rho^n_{j+1} - 2\rho^n_j + \rho^n_{j-1}) \\
&- \frac{c_{j-\frac{1}{2}}(1-c^2_{j-\frac{1}{2}})}{6}(\rho^n_j - 2\rho^n_{j-1} + \rho^n_{j-2}) \\
&- \frac{c_{j+\frac{1}{2}}(1-c^2_{j+\frac{1}{2}})(2+c_{j+\frac{1}{2}})}{24}(\rho^n_{j+2} - 3\rho^n_{j+1} + 3\rho^n_j - \rho^n_{j-1}) \\
&+ \frac{c_{j-\frac{1}{2}}(1-c^2_{j-\frac{1}{2}})(2+c_{j+\frac{1}{2}})}{24}(\rho^n_{j+1} - 3\rho^n_j + 3\rho^n_{j-1} - \rho^n_{j-2})
\end{aligned} \tag{2.45}$$

which is equivalent to the fourth order minimum amplitude error scheme of Rusanov (1970). The modified equivalent equation for this scheme is

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^4}{120}(2-c^2)(1-c^2)\frac{\partial^5 \rho}{\partial x^5} + O\{(\Delta x)^5\} \quad . \tag{2.46}$$

Alternatively, the points $x_{j-3},...,x_{j+1}$ could have been used to construct the polynomial

$$\begin{aligned} P_{4u}(x) &= \rho_j^n + \frac{x}{\Delta x}(\rho_{j+1}^n - \rho_j^n) + \frac{x(x-\Delta x)}{2(\Delta x)^2}(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \\ &- \frac{x((\Delta x)^2 - x^2)}{6(\Delta x)^3}(\rho_{j+1}^n - 3\rho_j^n + 3\rho_{j-1}^n - \rho_{j-2}^n) \\ &- \frac{x((\Delta x)^2 - x^2)(x+2\Delta x)}{24(\Delta x)^4}(\rho_{j+1}^n - 4\rho_j^n + 6\rho_{j-1}^n - 4\rho_{j-2}^n + \rho_{j-3}^n) \end{aligned} \tag{2.47}$$

yielding the difference equation

$$\begin{aligned} \rho_j^{n+1} = \rho_j^n &- c_{j+\frac{1}{2}}\rho_{j+1}^n + c_{j-\frac{1}{2}}\rho_j^n \\ &+ \frac{c_{j+\frac{1}{2}}(1+c_{j+\frac{1}{2}})}{2}(\rho_{j+1}^n - \rho_j^n) \\ &- \frac{c_{j-\frac{1}{2}}(1+c_{j-\frac{1}{2}})}{2}(\rho_j^n - \rho_{j-1}^n) \\ &+ \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)}{6}(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \\ &- \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)}{6}(\rho_j^n - 2\rho_{j-1}^n + \rho_{j-2}^n) \\ &+ \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)(2-c_{j+\frac{1}{2}})}{24}(\rho_{j+1}^n - 3\rho_j^n + 3\rho_{j-1}^n - \rho_{j-2}^n) \\ &- \frac{c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)(2-c_{j-\frac{1}{2}})}{24}(\rho_j^n - 3\rho_{j-1}^n + 3\rho_{j-2}^n - \rho_{j-3}^n) \end{aligned} \tag{2.48}$$

for which the modified equivalent equation is

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = -\frac{u(\Delta x)^4}{120}(3-c)(2-c)(1-c^2)\frac{\partial^5 \rho}{\partial x^5} + O\{(\Delta x)^5\} \quad . \tag{2.49}$$

Diagrams illustrating the wave propagation parameters and the results of the numerical test for Eq. (2.45) and Eq. (2.48) are given in Fig. 2.6(a-d) and Fig. 2.7(a-d), respectively. It is readily apparent that the accuracy of these two schemes is very similar, the major difference being that, again with the centred form, the oscillations lag the pulse but they precede the pulse when the upstream biassed scheme is used, as was found when comparing the two second order upwind schemes. This immediately suggests another method for calculating higher order schemes, by taking linear combinations of two schemes of equal order, (k), to produce a scheme

of order $k+1$. In this case, the linear combination gives fifth order upwinding. That this must be so, follows from the fact that fifth order upwinding is the unique difference scheme on the union of the two interpolation intervals $[x_{j-3}, x_{j+1}]$ and $[x_{j-2}, x_{j+2}]$ that has a modified equivalent equation of the form

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^5}{120}\eta_5(c)\frac{\partial^6 \rho}{\partial x^6} + O\{(\Delta x)^7\} \quad . \tag{2.50}$$

That is to say, if a weighted average of the two schemes is formed and the weight chosen to minimize the phase error (which is equivalent to removing the term involving $\partial^5\rho/\partial x^5$ from the resulting MEPDE), then the weight will be such that the combined scheme is exactly fifth order upwinding. Similarly the optimal combination of the two second order upwind schemes is third order upwinding. In fact, Noye (1987) has shown how this process can be used to construct "optimal" schemes, that is, schemes which are of maximum order for a specified stencil. This technique of taking two schemes of the same order and using a weighted average to construct a higher order scheme on the combined stencil will be used in the next chapter to obtain even higher order schemes .

The results of these two fourth order schemes slot in very well between third and fifth order upwinding. The relative phase speed and group velocities shown in Fig. 2.6(a,c) and Fig. 2.7(a,c) are very similar to those of third order upwinding shown in Fig. 2.4(a,c) with a zero crossing in the group velocity for small c in between $N_\lambda = 3$ and $N_\lambda = 4$, whereas for fifth order upwinding, shown in Fig. 2.5(c), the zero crossing is generally below $N_\lambda = 3$. The amplitude response of these two fourth order schemes, however, more closely resembles that of fifth order upwinding. This sort of behaviour is as expected from the MEPDE's of the four schemes. The modified equivalent equation, Eq. (2.40), for third order upwinding shows that this scheme has third order amplitude errors and fourth order phase errors, whereas Eq. (2.41) shows that fifth order upwinding has fifth order amplitude errors and sixth order phase errors. When compared with the modified equivalent equation for the two fourth order schemes, Eq. (2.46) and Eq. (2.49), this is as expected, i.e. the

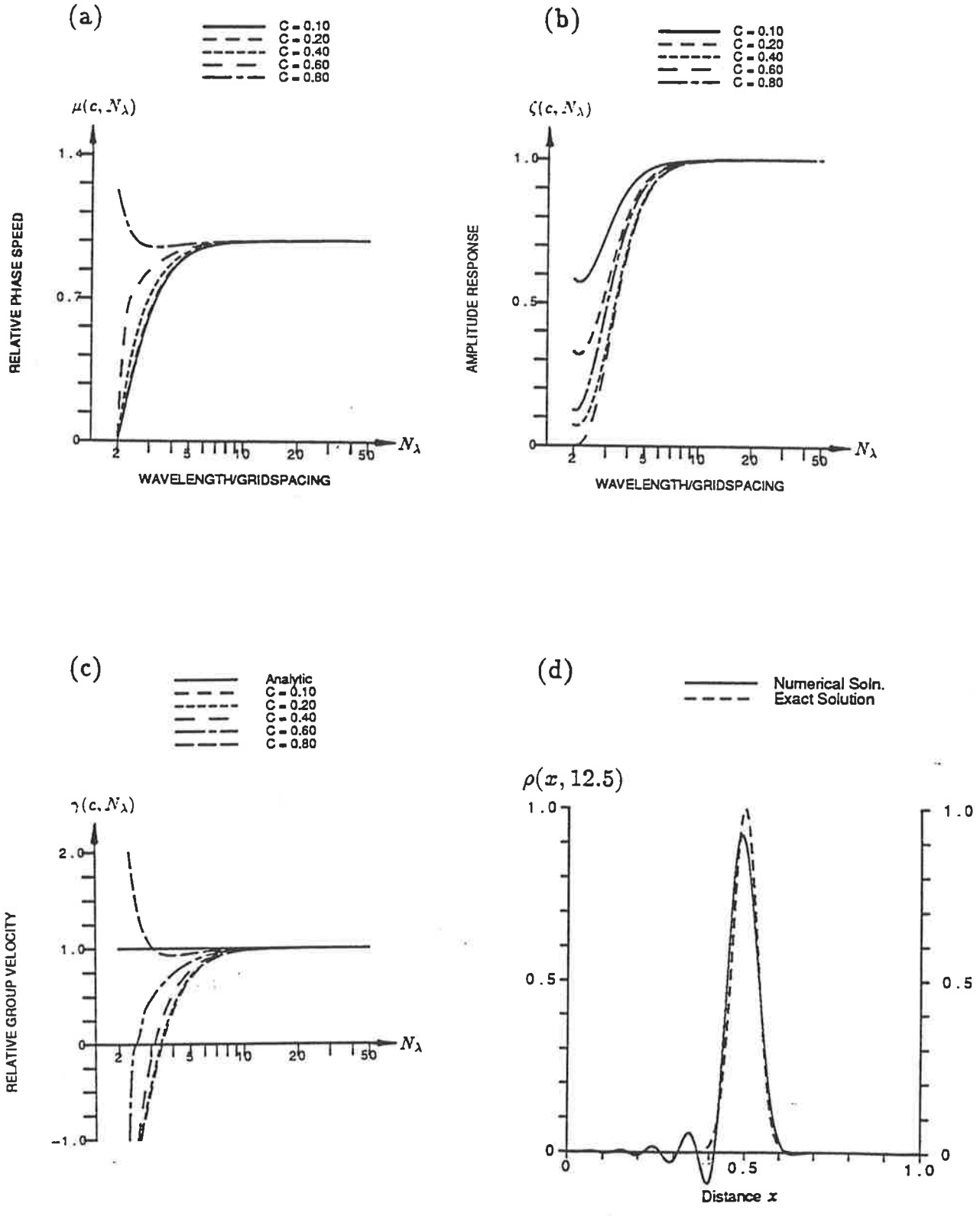

Figure 2.6: As in Fig. 2.1, but for Rusanov's fourth order scheme, Eq. (2.45). This scheme is stable for $c \leq 1$.

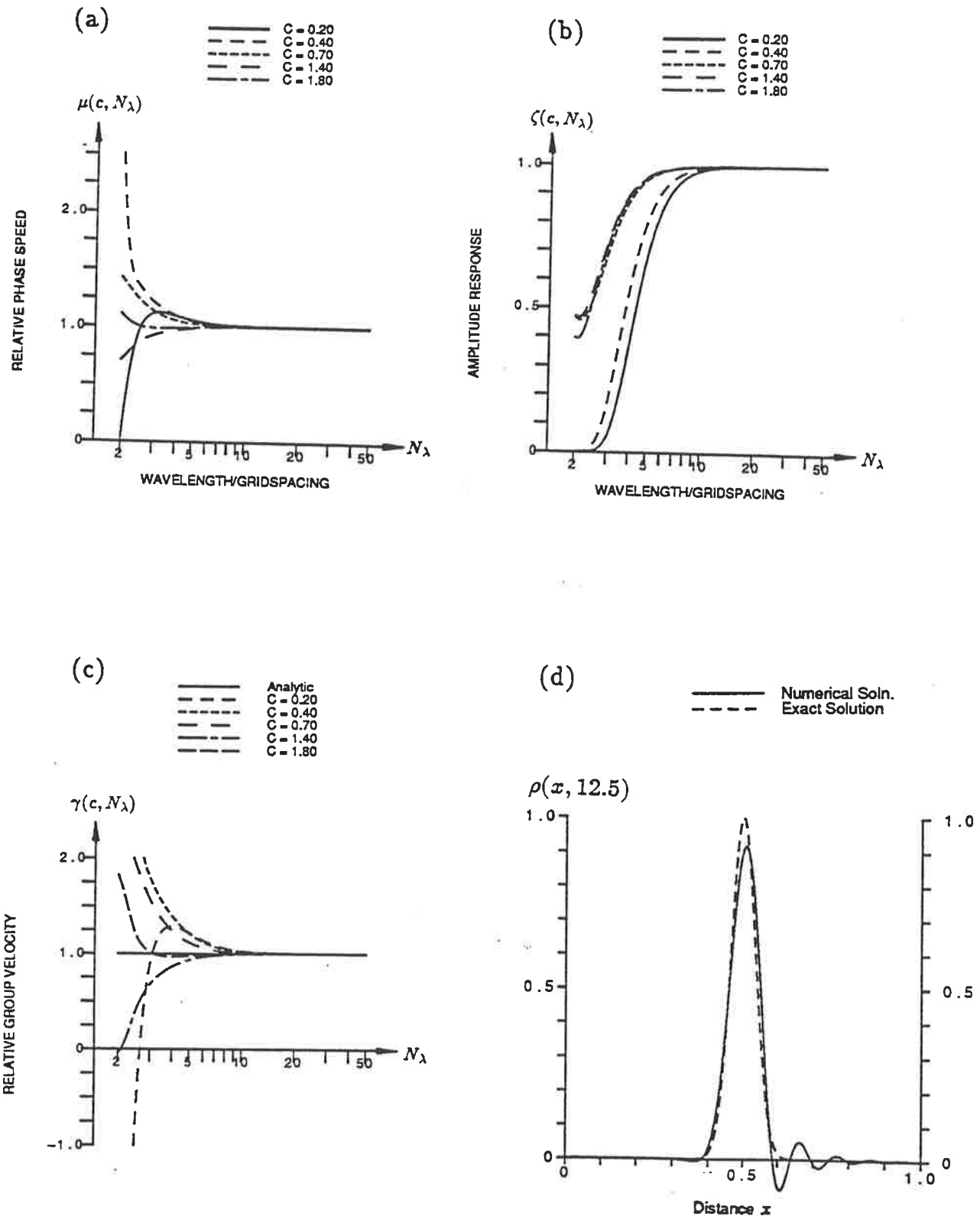

Figure 2.7: As in Fig. 2.1, but for the fourth order upstream biassed scheme, Eq. (2.48). This scheme is stable for $0 \leq c \leq 2$.

amplitude errors are similar to those of the fifth order scheme, but the phase errors are closer to those of third order upwinding.

2.3.3 Holly and Preissmann's Scheme

The scheme described by Holly and Preissmann (1977) uses a slightly different method to calculate the interpolating polynomials. It involves calculating the derivatives of ρ at each of the grid-points as well as ρ itself, and allows the use of third order interpolation, but only using information from the points (x_{j-1}, t^n) and (x_j, t^n). The scheme has an advantage over the previous schemes as the compact stencil makes it easier to handle irregular boundaries and boundary conditions, without significantly affecting the solution over the interior of the domain.

Again divided differences are used to obtain the interpolating polynomial for $\rho(x_j - w\Delta t, t^n)$. This cubic is differentiated to give a quadratic approximation, $\mathcal{R}$, to the derivative $\partial\rho/\partial x$. The resulting scheme is

$$\begin{aligned} \rho_j^{n+1} &= \rho_j^n - c^2(3-2c)(\rho_j^n - \rho_{j-1}^n) + \\ &\quad + c^2(1-c)\Delta x \mathcal{R}_{j-1}^n - c(1-c)^2\Delta x \mathcal{R}_j^n \\ \mathcal{R}_j^{n+1} &= -\frac{6c(c-1)}{\Delta x}(\rho_j^n - \rho_{j-1}^n) \\ &\quad - c(2-3c)\mathcal{R}_{j-1}^n + (1-c)(1-3c)\mathcal{R}_j^n \quad . \end{aligned} \tag{2.51}$$

To determine linear stability, substitute

$$\begin{bmatrix} \rho_j^n \\ \mathcal{R}_j^n \end{bmatrix} = \begin{bmatrix} \rho^\star \\ \mathcal{R}^\star \end{bmatrix} G^n e^{i\beta j} \tag{2.52}$$

into the difference equation, giving

$$(G\mathbf{I} - \mathbf{K}) \begin{bmatrix} \rho^\star \\ \mathcal{R}^\star \end{bmatrix} = 0 \tag{2.53}$$

where

$$\mathbf{K} = \begin{bmatrix} 1 + c^2(3-2c)(e^{-i\beta} - 1) & c(1-c)\Delta x \left(ce^{-i\beta} - (1-c) \right) \\ 6c(1-c)(1-e^{-i\beta})/\Delta x & -c(2-3c)e^{-i\beta} + (1-c)(1-3c) \end{bmatrix} \quad . \tag{2.54}$$

These equations only have non-trivial solutions if the determinant of the coefficient matrix is zero. This gives a quadratic equation for the eigenvalues of $\mathbf{K}$, namely G_1, G_2 and hence two solutions for ρ_j^n and $\mathcal{R}_j^n$. For stability, both of these values must have absolute value less than or equal to unity. From these two values of G, the wave propagation parameters may be constructed, with the two values giving two different modes, each with its own form of μ, ζ and γ. One of these two modes closely resembles the physical solution, the other is a spurious computational mode. The computational mode will be seen to decay very quickly for well-resolved components. It should be noted that the eigenvalues for Eq. (2.53) will not depend on Δx, since the eigenvalues are the solutions of

$$G^2 - (k_{1,1} + k_{2,2})\,G + k_{1,1}k_{2,2} - k_{1,2}k_{2,1} = 0 \tag{2.55}$$

where $k_{i,j}$ is the $(i,j)^{th}$ element of K. That is, the two terms involving Δx are multiplied together, giving an equation for G_1, G_2 which is independent of Δx.

The initial values for the derivatives can be obtained by standard approximations, e.g. centred or one-sided differences, such as

$$\begin{aligned} \mathcal{R}_j^n &\simeq \frac{\rho_{j+1}^n - \rho_{j-1}^n}{2\Delta x} \\ &\simeq \frac{\rho_{j+1}^n - \rho_j^n}{\Delta x} \quad . \end{aligned} \tag{2.56}$$

Holly and Preissmann (1977) noted that the overall performance of the scheme was virtually independent of how the derivatives are initially calculated and hence, the wave propagation parameters shown in Fig. 2.8(a-c) refer to the physical mode only. For the numerical test shown in Fig. 2.8(d), the derivatives were initially calculated using the centred differences. This scheme obviously produces excellent results, more closely representing fifth order upwinding than four-point third order upwinding, and the relative group velocity appears to be excellent. It must be remembered though, that this scheme requires twice the storage of the other schemes and so for a fairer comparison with four-point third order, the values for the wave propagation parameters for Holly and Preissmann's scheme should be compared with

those of twice the resolution for the four-point scheme, Eq. (2.37). That is, using "HP" to denote Holly and Preissmann's scheme and "UW3" to denote four-point third order upwinding, the amplitude response, $\zeta_{HP}(c, N_\lambda)$, should be compared with $\zeta_{UW3}(2c, 2N_\lambda)$ and similarly for the relative phase speed and group velocity. When this is done, the performance of the two schemes is quite similar. This is shown in Fig. 2.9(a-c) which illustrates the performance of four-point third order upwinding with twice the grid-spacing. The results show how well a particular Fourier component is advected by third order upwinding with twice the resolution. Hence, N_λ refers to the non-dimensional wavelength in the low resolution case, whereas the values of γ, μ and ζ refer to the high resolution case. For example, a component of length $2\Delta x$ in the low order solution is propagated with phase speed $\mu(c/2, 2)$ in the high resolution case (where c refers to the Courant number in the low resolution case). The diagrams here are much closer than those for Holly and Preissmann's scheme shown in Fig. 2.8(a-c). Another point about Holly and Preissmann's scheme is that it cannot be written in flux form and so when the velocity w varies with space this scheme is not necessarily conservative.

In one-dimensional problems, the issue of storage is generally not crucial, but in multi-dimensional computations it is significant, and for this reason it may be thought preferable to use third order upwinding which is still quite accurate.

Holly and Preissmann's approach does give a fast and accurate scheme, and if higher order forms can be found then these will be both compact and fast. There are three possible approaches. One is to advect the second spatial derivative as well as the first, however, this means tripling the storage of the scheme, which is undesirable. The other two possible approaches involve using a three-point scheme rather than a two-point scheme. For example, using values at x_{j-1}, x_j and x_{j+1} gives

$$\begin{aligned}\rho_j^{n+1} &= \rho_j^n - (c - \frac{5c^3}{2} + \frac{3c^5}{2})\Delta x \mathcal{R}_j^n \\ &+ \frac{c^2}{2}(2 - c^2)(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) - \frac{c^3}{4}(5 - 3c^2)(\rho_{j+1}^n - \rho_{j-1}^n)\end{aligned}$$

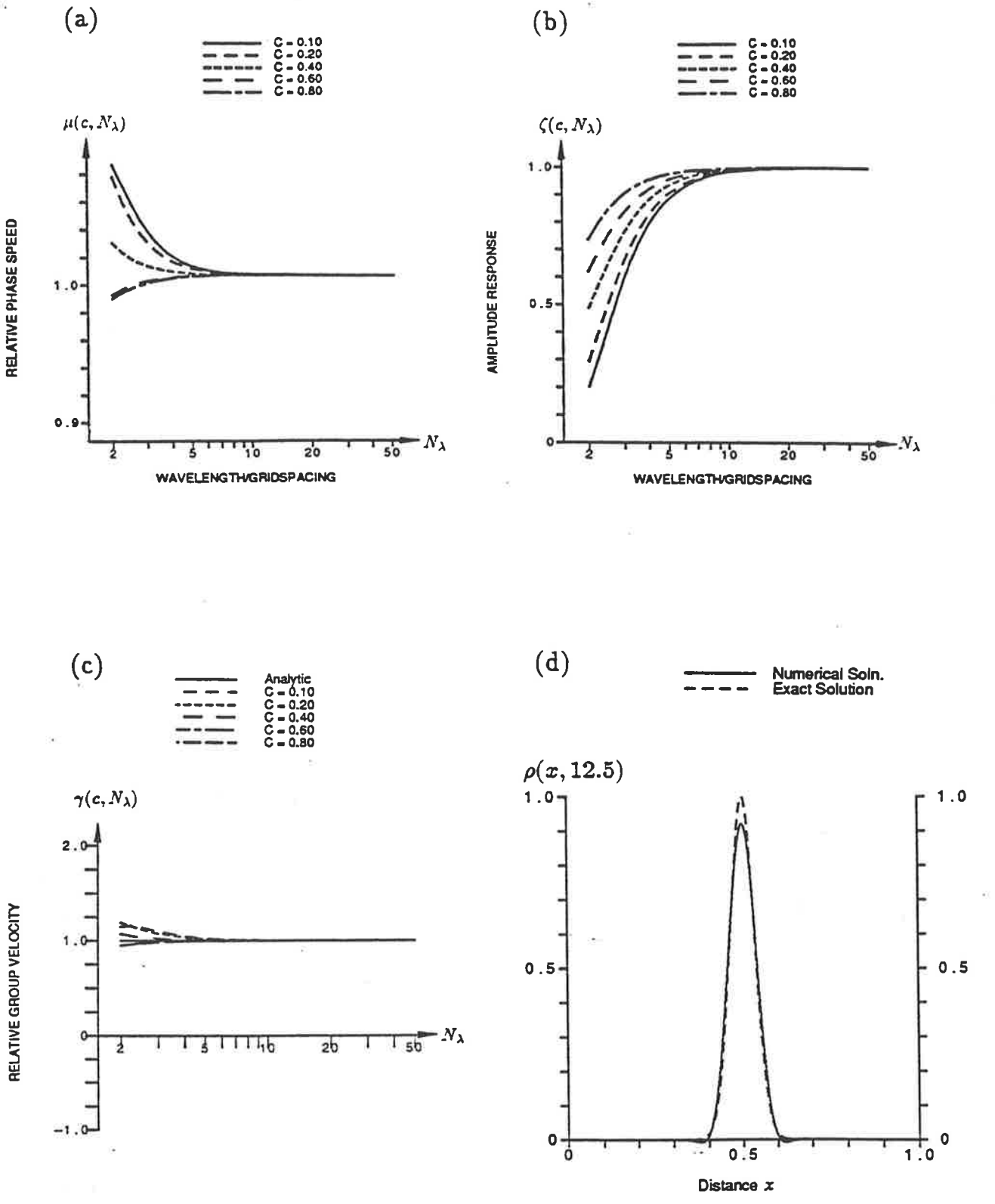

Figure 2.8: As in Fig. 2.1, but for Holly and Preissmann's scheme, Eq. (2.51). This scheme is stable for $c \leq 1$.

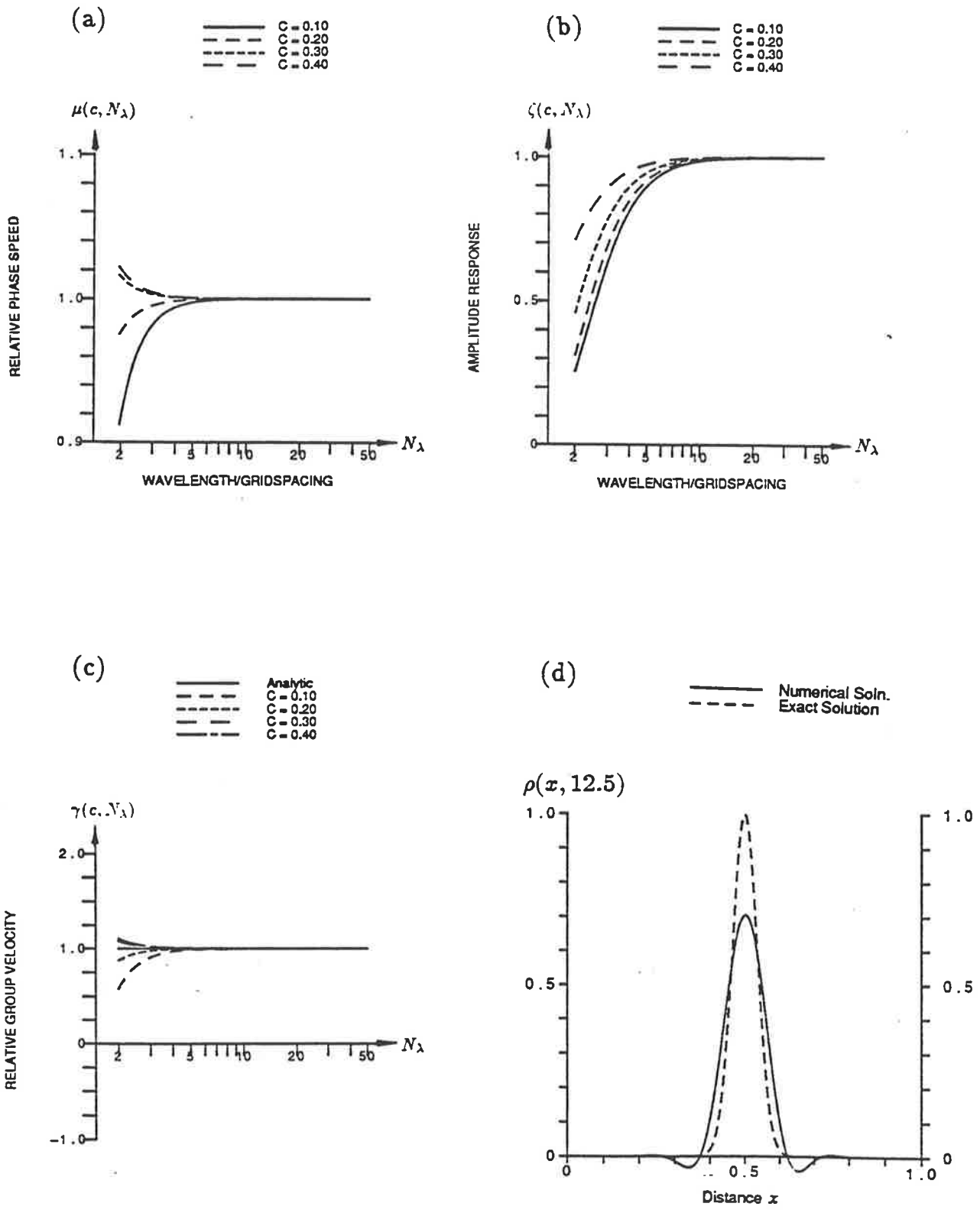

Figure 2.9: Effect of resolution on third order upwinding and Holly and Preissmann's scheme. Diagrams (a-c) are as in Fig. 2.4(a-c) but with twice the resolution. Diagram (d) illustrates the results from using Holly and Preissmann's scheme with half the resolution, to be compared with Fig. 2.4(d).

$$
\begin{aligned}
&- \frac{c^2\Delta x}{4}(1-c^2)(\mathcal{R}^n_{j+1}-\mathcal{R}^n_{j-1}) \\
&+ \frac{c^3\Delta x}{4}(1-c^2)(\mathcal{R}^n_{j+1}-2\mathcal{R}^n_j+\mathcal{R}^n_{j-1})
\end{aligned}
\tag{2.57}
$$

$$
\begin{aligned}
\mathcal{R}^{n+1}_j &= \mathcal{R}^n_j - (\frac{9c^2}{2}-\frac{15c^4}{4})\mathcal{R}^n_j - \frac{2c}{\Delta x}(1-c^2)(\rho^n_{j+1}-2\rho^n_j+\rho^n_{j-1}) \\
&+ \frac{15c^2}{4\Delta x}(1-c^2)(\rho^n_{j+1}-\rho^n_{j-1}) + \frac{c}{2}(1-2c^2)(\mathcal{R}^n_{j+1}-\mathcal{R}^n_{j-1}) \\
&- \frac{c^2}{4}(3-5c^2)(\mathcal{R}^n_{j+1}-2\mathcal{R}^n_j+\mathcal{R}^n_{j-1})
\end{aligned}
$$

or using values at x_{j-2} , x_{j-1} and x_j gives

$$
\begin{aligned}
\rho^{n+1}_j &= \rho^n_j - c\Delta x\mathcal{R}^n_j - c(1-c)(\Delta x\mathcal{R}^n_j - \rho^n_j + \rho^n_{j-1}) \\
&- c(1-c)^2\left[\Delta x(\mathcal{R}^n_j+\mathcal{R}^n_{j-1}) - 2\rho^n_j + 2\rho^n_{j-1}\right] \\
&- \frac{c(1-c)^2(2-c)}{4}\left[\Delta x(2\mathcal{R}^n_j+4\mathcal{R}^n_{j-1}) - 5\rho^n_j + 4\rho^n_{j-1} + \rho^n_{j-2}\right] \\
&- \frac{c(1-c)^2(2-c)^2}{4}\left[\Delta x(\mathcal{R}^n_j+4\mathcal{R}^n_{j-1}+\mathcal{R}^n_{j-2}) - 3\rho^n_j + 3\rho^n_{j-2}\right]
\end{aligned}
\tag{2.58}
$$

$$
\begin{aligned}
\mathcal{R}^{n+1}_j &= \mathcal{R}^n_j + (1-2c)\left[\mathcal{R}^n_j - \frac{1}{\Delta x}(\rho^n_j-\rho^n_{j-1})\right] \\
&+ (1-c)(1-3c)\left[\mathcal{R}^n_j+\mathcal{R}^n_{j-1} - \frac{2}{\Delta x}(\rho^n_j-\rho^n_{j-1})\right] \\
&- \frac{(1-c)(4c^2-8c+2)}{4}\left[2\mathcal{R}^n_j+4\mathcal{R}^n_{j-1} + \frac{1}{\Delta x}(-5\rho^n_j+4\rho^n_{j-1}+\rho^n_{j-2})\right] \\
&- \frac{(1-c)(2-c)(9c^2-9+2)}{4} \times \\
&\qquad \left[\mathcal{R}^n_j+4\mathcal{R}^n_{j-1}+\mathcal{R}^n_{j-2} - \frac{3}{\Delta x}(\rho^n_j-\rho^n_{j-2})\right] .
\end{aligned}
$$

Unfortunately both of these schemes are unconditionally unstable and so this is not a suitable approach for obtaining very high order schemes.

2.4 Conclusion

A variety of explicit finite difference schemes have been discussed in this chapter in order to provide a benchmark against which other schemes can be compared. These schemes were derived here using polynomial interpolation to approximate

$\rho(j\Delta x - w\Delta t, n\Delta t)$, although when originally developed, many of them were derived by other methods. It was pointed out that using Lagrange polynomials to perform the interpolation was equivalent to obtaining a modified equivalent equation of the highest possible order for an equivalent stencil. All these schemes can be obtained directly using the modified equivalent equation of Noye and Hayman (1986) as discussed earlier in this chapter.

The derivation using interpolation was used as it provides some additional insight into which stencils are candidates to produce stable schemes. The behaviour of high order polynomial interpolation towards the ends of the interpolation interval restricts the upstream bias of the possible stencils. The modified equivalent equation has, however, been shown to give additional information on the likely performance of the scheme due to the equivalence between it and the wave propagation parameters.

In addition to the likely accuracy of the scheme, the modified equivalent equation also provides an insight into the nature of the dominant errors, i.e. whether the errors are predominantly due to numerical damping or numerical dispersion.

The validity of using the wave propagation parameters (or the modified equivalent equation) as a basis for comparing different schemes is demonstrated by comparing the test case of advecting a thin Gaussian peak. That is, improving the order as $\Delta x \rightarrow 0$ of the wave propagation parameters (or the modified equivalent equation) not only provides greater accuracy in the limit as $\Delta x \rightarrow 0$, but also improves accuracy in general. This has only been shown for a case where the initial conditions are everywhere infinitely differentiable; the effect of using more general initial conditions will be discussed in the next chapter.

There remains the problem of how to quantify accuracy. Table 2.1 presents a comparison of different error measures for the schemes discussed in this chapter. The errors were calculated using the standard example of advecting a Gaussian pulse for ten periods and these illustrate several different features of each scheme's performance. From this, it is possible to gain an impression of the overall performance

Table 2.1:Error measures for standard difference schemes for the test problem with a Gaussian pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Minimum $\{\rho_j^n\}$	Sum Neg. Values	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
First order Upwinding	2.2×10^{-1}	8.6×10^{-1}	3.7×10^{-1}	0	-7.5×10^{-3}	42.30	1.2×10^{-3}	-0.1	0.14
Lax-Wendroff scheme	2.0×10^{-1}	5.7×10^{-1}	-3.0×10^{-1}	-3.7×10^{-1}	-5.5×10^{-1}	29.16	7.9×10^{-2}	-6.6	0.89
2^{nd} order Upstream biased	1.5×10^{-1}	4.7×10^{-1}	-2.8×10^{-1}	-4.0×10^{-1}	5.3×10^{-2}	28.74	-1.1×10^{-3}	7.4	0.75
Holly & Preissmann	1.8×10^{-2}	7.9×10^{-2}	-8.0×10^{-3}	-5.2×10^{-4}	2.5×10^{-13}	0.14	-5.5×10^{-6}	0.0	0.92
- with half the resolution	7.5×10^{-2}	2.9×10^{-1}	-4.3×10^{-2}	-2.2×10^{-1}	-1.0×10^{-5}	2.00	-2.9×10^{-5}	0.1	0.71
3^{rd} order Upwinding	6.7×10^{-2}	2.6×10^{-1}	-3.7×10^{-2}	-3.9×10^{-1}	2.8×10^{-6}	1.62	1.4×10^{-5}	0.0	0.74
Rusanov's 4^{th} order scheme	3.9×10^{-2}	1.4×10^{-1}	-8.7×10^{-2}	-4.5×10^{-1}	2.7×10^{-3}	2.01	6.1×10^{-4}	-0.8	0.93
4^{th} order Upstream biased	3.9×10^{-2}	1.4×10^{-1}	-8.8×10^{-2}	-4.3×10^{-1}	-6.5×10^{-4}	1.72	-3.5×10^{-4}	0.8	0.92
5^{th} order Upwinding	1.1×10^{-2}	4.4×10^{-2}	-3.4×10^{-3}	-1.6×10^{-2}	-2.2×10^{-8}	0.05	4.0×10^{-6}	0.0	0.96

of each scheme. The errors are all calculated by examining the numerical solution after the final time-step only, so that the minimum value of ρ_j^n is taken over the values of ρ_1^{2500} to ρ_{100}^{2500}. One problem in calculating some of these error measures, is locating the actual position of the centre of the pulse. Since no schemes propagate every Fourier component with the correct phase speed, the peak of the numerical solution may lag the peak of the analytical solution by a fraction of a grid spacing. If the values from the analytic solution are then compared directly with the numerical solution, the resulting error measures may not be truly representative of the scheme's performance. This is especially true when considering the peak height or the second and third order moments of the numerical solution.

There are many ways of attempting to "recover" the numerical peak; the simplest being to use polynomial interpolation near the maximum of the numerical solution. From this it is a simple matter to obtain approximations to the peak height and peak position. Since all the schemes here are based on polynomial interpolation, this is an appropriate method to recover the "true" numerical peak. If the polynomial used to perform the interpolation matches the order of the scheme, however, it is possible that this will introduce a favourable bias towards high order schemes, since high order polynomials can allow sharper and higher estimations of the peak height. In order to maintain consistency between the error measures for different order schemes, quadratic interpolation is used. Having established a fair and representative method for estimating the peak height and position, it is a straightforward matter also to estimate the centred second and third moments of the numerical solution.

Table 2.1 also presents error measures for Holly and Preissmann's two-point third order scheme, with the grid-spacing doubled (or the resolution halved). As mentioned before, this scheme produces excellent results when used with the same grid-spacing which is as expected since it requires twice the storage and effectively uses two values from each grid-point, (i.e. the scheme is running at twice the resolution of the other schemes, although this extra data originates from the same initial

condition as the other schemes used). The two sets of error measures for this scheme are presented, one with $\Delta x = 0.01$, the other with $\Delta x = 0.02$ to illustrate the overall performance of the method and to demonstrate that this is due to the advection of the derivatives. For the case where $\Delta x = 0.01$, Holly and Preissmann's scheme is comparable to fifth order upwinding, as is expected from comparing Fig. 2.8(d) with Fig. 2.5(d). The errors for the halved resolution are comparable to those of four-point third order upwinding, however, confirming earlier conclusions based on Fig. 2.9(a-c).

The error measures presented in Table 2.1 can be used to quantify the ability of each scheme to reproduce specific features of the analytic solution. For example, the minimum value of the numerical solution and the sum of negative values, give measures of the size of the spurious oscillations (provided the minimum value is negative). The two second order schemes have relatively high values for both of these error measures, as would be expected since both schemes produce such large oscillations. If the minimum value is positive, such as with first order upwinding, then the scheme must clearly be overly diffusive since the pulse has been smeared over the entire domain.

The two fourth order schemes produce oscillations of larger amplitude but shorter wavelength, in comparison with third order upwinding. This can be explained by examining the ratios of the errors for either of the fourth order schemes to third order upwinding. The ratio of the minima of the numerical solution is about 2.4 in each case, indicating that the fourth order schemes have side lobes more than twice as large as those of third order upwinding. The ratio of the sums of the negative values, however, are only about 1.2 in either case, indicating that the side lobes have close to the same area in the two cases. Thus the side lobes of the fourth order solution are approximately half the wavelength of those of third order upwinding. This is useful information when it comes to smoothing the solution as higher frequency noise can be damped out with less distortion of the pulse than lower frequency noise. So the

oscillations of the two schemes may be considered to be of similar severity since, for the fourth order scheme, short components must be damped more strongly, but not so many components require damping. For third order upwinding, on the other hand, the damping need not be as strong but more components must be damped.

The first, second and third moments of a distribution measure the centre of mass, spread and skewness, respectively. The sensitivity of the first two moments is best illustrated by comparing the errors for the three low order schemes. First order upwinding gives a very large error in the second moment due to large numerical damping and the two second order schemes give comparatively large errors in the first moment due to the large numerical oscillations of the latter schemes. The second moment is to some extent, also affected by dispersive errors, as is shown in the large errors for the two second order schemes, relative to the higher order schemes. The sensitivity of the third moment is best demonstrated in the error measures for the two fourth order schemes where the error in the centre of mass of the numerical solution is of opposite sign to the error in the peak position. This is explained by noting that the third moment is quite large in comparison with schemes of comparable accuracy (third and fifth order upwinding) indicating a definite tilt in the solution. This skewness is also evident in the illustrations Fig. 2.6(d) and Fig. 2.7(d).

The peak shift and peak height are both calculated by quadratic interpolation for the reasons stated above. These give a direct indication of how well the extremum itself is advected. Clearly the higher order schemes handle the peak better than the low order schemes. Furthermore, it is again apparent that even order schemes are better at retaining the amplitude of the wave (the second and fourth order peak heights are comparable to the third and fifth order peak heights). Whereas the odd order schemes are far superior at propagating the peak at the correct speed, the peak shifts for the odd order schemes are considerably better than for the low order schemes.

The two remaining error measures, the maximum absolute error and the root mean square (RMS) error, provide useful indications of the overall accuracy of the schemes. Of the two, the RMS error provides the better guide in that the maximum error is somewhat overly sensitive to errors in overall speed of the solution. The RMS error indicates that the first and second order schemes yield almost equally poor results, which confirms the impressions gained from examining Figs. 2.1, 2.3. For a particular application, some of these error measures may be of more importance than others, but since it is intended to develop a general scheme, it is the RMS error that will be used to quantify the accuracy of subsequent schemes. Reference to the other error measures will still be made to demonstrate any special features of a particular scheme.

The error measures presented in Table 2.1 clearly demonstrate the improvement in accuracy due to improved order, as was expected from the wave propagation parameters. This test cannot distinguish between schemes of equivalent order, as for example, in the case of the two fourth order schemes, where the error measures are very close. This is to be expected, from the wave propagation parameters which were very similar in nature, as shown in Fig. 2.6(a-c) and Fig. 2.7(a-c) (the phase lead of the upstream biassed scheme being very close in magnitude to the phase lag of Rusanov's scheme). The main conclusion to be made from this comparison between the different methods is that the theoretical order in the limit as $\Delta x \rightarrow 0$ is closely related to the overall accuracy of a scheme in practice. In problems more complex than the simple test case presented here, other factors will also be important. It is clear, however, that one of the major factors in determining overall accuracy has been isolated.

Besides the need for comparisons of computational expense and tests involving initial conditions that are not infinitely differentiable, the effect of velocity being fully variable requires some explanation. The derivations discussed here (Lagrange interpolation and modified equivalent equation) both assume that the velocity is

constant. If the velocity w is a function of space and/or time, then terms involving derivatives of w appear in the modified equivalent equation introducing low order terms. Alternatively, the derivation based on Lagrange interpolation requires that the point where the characteristic containing the point (x_j, t^{n+1}) crosses the line $t = t^n$ in (x,t) space is known, i.e. the point $(x_j - w\Delta t, t^n)$ is known. If this point is known, then $w\Delta t$ is immediately defined and a constant value of the Courant number c can be used throughout the difference equation for ρ_j^{n+1}. In general, it is not possible to exactly backtrack along the characteristic and so the value of $x_j - w\Delta t$ is possibly of lower order than the rest of the scheme.

Such a process is not necessarily conservative. An alternative is to regard the difference equations to be in flux form $\rho_j^{n+1} = \rho_j^n - f_j + f_{j-1}$ and the order of the scheme to be related to the order of the fluxes f_j and f_{j-1} across the boundaries of the j^{th} grid cell (which is defined as the interval $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$). The value of $f_{j+\frac{1}{2}}$ may then be regarded as the flux out of grid cell j. In such a case, the velocity involved with each flux is the velocity at the grid cell boundary, and so $f_{j+\frac{1}{2}}$ depends only on the velocity $w_{j+\frac{1}{2}}$, and if w varies in time, then some form of mean over the time-step should be used. This approach allows for a simple extension to the case of variable velocity and is guaranteed to be conservative since the net amount of material within any region of the domain equals the difference between the amount of material entering and the amount leaving. For these reasons, the difference schemes were given in flux form as it will be in this form that the schemes are most amenable for general use.

Chapter 3

Implicit Finite Difference Schemes

In the previous chapter, a selection of explicit finite difference schemes for solving the one-dimensional constant coefficient advection equation were discussed. It was shown how they could all be developed by the use of the modified equivalent equation. A method of comparing the different schemes based on wave propagation parameters and a numerical test was also described. The relative merit of the different schemes determined by these comparisons, closely corresponded to the relative order of the different modified equivalent equations. The comparisons clearly demonstrate that higher order schemes provide greater overall accuracy. There are, however, two problems with these methods: firstly, to obtain high order schemes requires a wide stencil, and secondly, there are still stability restrictions on the results. By utilizing implicit finite difference schemes, it is possible to overcome these problems. All the schemes discussed in this chapter will be unconditionally stable in the von Neumann sense, however, there still may be restrictions on the time-step due to the stability of the matrix inversion. The schemes that do have restrictions on the time-step will be shown, however, to be highly accurate. A measure of efficiency must involve the overall accuracy of a scheme as well as the time spent obtaining the results. This means that there are two ways in which a scheme may be considered efficient, either by generating moderately accurate results quickly or by taking longer but providing highly accurate results. Schemes of both types will be discussed in this chapter.

As all the numerical experiments were performed on a Pyramid 9820 computer, a scalar machine, the matrix inversion was performed by an elimination algorithm of the type discussed by Thomas (1949). This algorithm is easily modified to cater for different types of boundary condition and penta-diagonal systems. The use of such algorithms influences the timings of the numerical experiments, since these algorithms cannot be vectorized. If a vector machine is to be used then other inversion techniques, such as conjugate gradient methods are required. Since new algorithms for matrix inversion on vector machines are continually being developed, it is beyond the scope of this thesis to include all the possible effects of using vector machines. Attention will be confined to scalar computers.

3.1 Implicit Central Differencing

The most direct method of ensuring that a finite difference scheme is at least second order, is to use only second order approximations to the derivatives in the model equation. The simplest way of achieving this is to use central differences in time as well as space, an approach which has several advantages over others as outlined below.

In order to derive modified equivalent equations and wave propagation parameters, it will again be assumed that $w(x,t) = u$, a positive constant. Later, some numerical damping tests will be presented for cases where the velocity is a function of x and it will be shown that conclusions reached under the assumption of constant velocity will apply to the more general case.

For the constant velocity case, the general form of a centred two-level finite difference scheme is

$$\sum_{p=-R}^{R} \theta_p \frac{\rho_{j+p}^{n+1} - \rho_{j+p}^{n}}{\Delta t} + u \sum_{q=1}^{R} \frac{\xi_q}{2} \left\{ \frac{\rho_{j+q}^{n+1} - \rho_{j-q}^{n+1}}{2q\Delta x} + \frac{\rho_{j+q}^{n} - \rho_{j-q}^{n}}{2q\Delta x} \right\} = 0 \tag{3.1}$$

with $\theta_p = \theta_{-p}$, $p \geq 1$, for some integer, R. Furthermore, for the difference equation

to be consistent with the advection equation,

$$\sum_{p=-R}^{R} \theta_p = \sum_{q=1}^{R} \xi_q \tag{3.2}$$

This equation can be rewritten in the form

$$\sum_{p=-R}^{R} A_p \rho_{j+p}^{n+1} = \sum_{p=-R}^{R} B_p \rho_{j+p}^{n} \tag{3.3}$$

where

$$A_p = \begin{cases} \theta_p + c\xi_p/(4p) & p > 0 \\ \theta_p - c\xi_p/(4p) & p < 0 \\ \theta_0 & p = 0 \end{cases} \tag{3.4}$$

and

$$B_p = \begin{cases} \theta_p - c\xi_p/(4p) & p > 0 \\ \theta_p + c\xi_p/(4p) & p < 0 \\ \theta_0 & p = 0 \end{cases} \tag{3.5}$$

so $B_{-p} = A_p$. One of the most important consequences of this "*diagonal differencing*" is that the von Neumann amplification factor lies on the unit circle for all values of N_λ and c. This is seen by using the relation between the coefficients A_p and B_p and substituting $\rho_j^n = G^n exp(i\beta j)$ into Eq. (3.2) giving

$$\left[A_0 + \sum_{k>0} \left\{ 2\theta_k \cos(\beta k) + i\frac{2c}{4k}\xi_k \sin(\beta k) \right\} \right] G = A_0 + \sum_{k>0} \left\{ 2\theta_k \cos(\beta k) - i\frac{2c}{4k}\xi_k \sin(\beta k) \right\} \tag{3.6}$$

which is of the form

$$(a + ib)G = a - ib \quad , \qquad a,\ b\ real \tag{3.7}$$

and hence $|G| = 1$ for all N_λ and c. So not only are such schemes unconditionally stable but they also have an amplitude response of unity for all (N_λ, c). This can be generalized to apply to methods involving more than two time levels as discussed in Noye (1987).

The modified equivalent equation for such schemes will only contain even order terms (i.e. odd order derivatives) as a direct consequence of using centred difference approximations. It then follows from Section 2.2 that the only source of error in these schemes will be numerical dispersion. This may, however, lead to larger errors

than those of the explicit schemes (discussed in Chapter Two) because the short wavelength components that are out of phase with the longer wavelengths will have larger amplitude than before, due to the absence of any numerical damping.

3.1.1 Crank Nicolson scheme

This is the best known three-point, two-level implicit finite difference scheme and is also the simplest equation of the form Eq. (3.2). It is obtained by putting $R = 1$, $\theta_0 = 1$, $\theta_{\pm 1} = 0$ and $\xi_1 = 1$, giving in the general case

$$\begin{aligned} \rho_j^{n+1} &+ \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^{n+1} + \rho_j^{n+1}) - \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^{n+1} + \rho_{j-1}^{n+1}) \\ &= \rho_j^n - \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^n + \rho_j^n) + \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^n + \rho_{j-1}^n) \quad . \end{aligned} \tag{3.8}$$

The modified equivalent equation for this scheme is

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = -\frac{u(\Delta x)^2}{3!}(\frac{c^2}{2} + 1)\frac{\partial^3 \rho}{\partial x^3} + O\{(\Delta x)^4\} \quad . \tag{3.9}$$

Although this scheme is unconditionally stable in the von Neumann sense, there remains the problem of stability of the other set of calculations involved with this scheme, namely the solution of the tri-diagonal system of linear algebraic equations. A sufficient condition for this is that the system of equations is diagonally dominant, i.e.

$$|A_0| \geq |A_{-1}| + |A_1| \tag{3.10}$$

which reduces to $c \leq 2$ for this scheme. This guarantees that the Crank Nicolson scheme is useful for $c \leq 2$.

The results of this scheme are shown in Fig. 3.1(c) illustrating the problems of dispersive errors, untempered by diffusive errors. The diagram is similar to that for the Lax-Wendroff scheme, Fig. 2.2(d). In fact, the results for the Crank Nicolson scheme are slightly worse as can be seen from Table 3.1, which lists the same error measures which were presented in Table 2.1.

The deterioration of results is evident in the larger and sharper secondary peak in the graph of the numerical results. This is a direct consequence of the absence of

any diffusive errors to cancel out the dispersive errors. If numerical damping were present, then the diffusion of the high frequency components should smooth out the secondary maximum making it more like that shown in the illustration of the test of the Lax-Wendroff scheme.

The relative phase speed and group velocity are shown in Fig. 3.1(a,b). It is worth noting that there is little change in these as the Courant number varies, so that the dispersion seen in Fig. 3.1(c) is typical of the scheme. That is, the propagation of a component is almost solely determined by the resolution of the component (N_λ) and virtually independent of the velocity field or time-step.

3.1.2 Linear Finite Element Crank Nicolson

A suggested improvement on the Crank Nicolson scheme was first obtained by improving the order of a finite element scheme (Cathers and O'Connor, 1985). The improvement comes from using all six computational points in the approximation to the temporal derivative. The optimal weighting was found to be $\theta_0 = 2/3$, $\theta_1 = \theta_{-1} = 1/6$, giving the linear finite element Crank Nicolson difference equation

$$\begin{aligned}
\rho_j^{n+1} &+ \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^{n+1} + \rho_j^{n+1}) - \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^{n+1} + \rho_{j-1}^{n+1}) \\
&+ \frac{1}{6}(\rho_{j+1}^{n+1} - \rho_j^{n+1}) - \frac{1}{6}(\rho_j^{n+1} - \rho_{j-1}^{n+1}) \\
&= \rho_j^n + \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^n + \rho_j^n) - \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^n + \rho_{j-1}^n) \\
&\quad + \frac{1}{6}(\rho_{j+1}^n - \rho_j^n) - \frac{1}{6}(\rho_j^n - \rho_{j-1}^n) \qquad (3.11)
\end{aligned}$$

with a modified equivalent equation

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} &= -\frac{u(\Delta x)^2}{3!}\frac{c^2}{2}\frac{\partial^3 \rho}{\partial x^3} \\
&\quad + \frac{u(\Delta x)^4}{5!}\left(\frac{2}{3} - \frac{3c^4}{2}\right)\frac{\partial^5 \rho}{\partial x^5} + O\{(\Delta x)^6\} \quad . \qquad (3.12)
\end{aligned}$$

As can be seen, the scheme is still only second order, however, the leading term is now proportional to c^2 rather than $1 + c^2$ and so for small c at least, an appreciable improvement over the standard Crank Nicolson scheme is obtained. In fact,

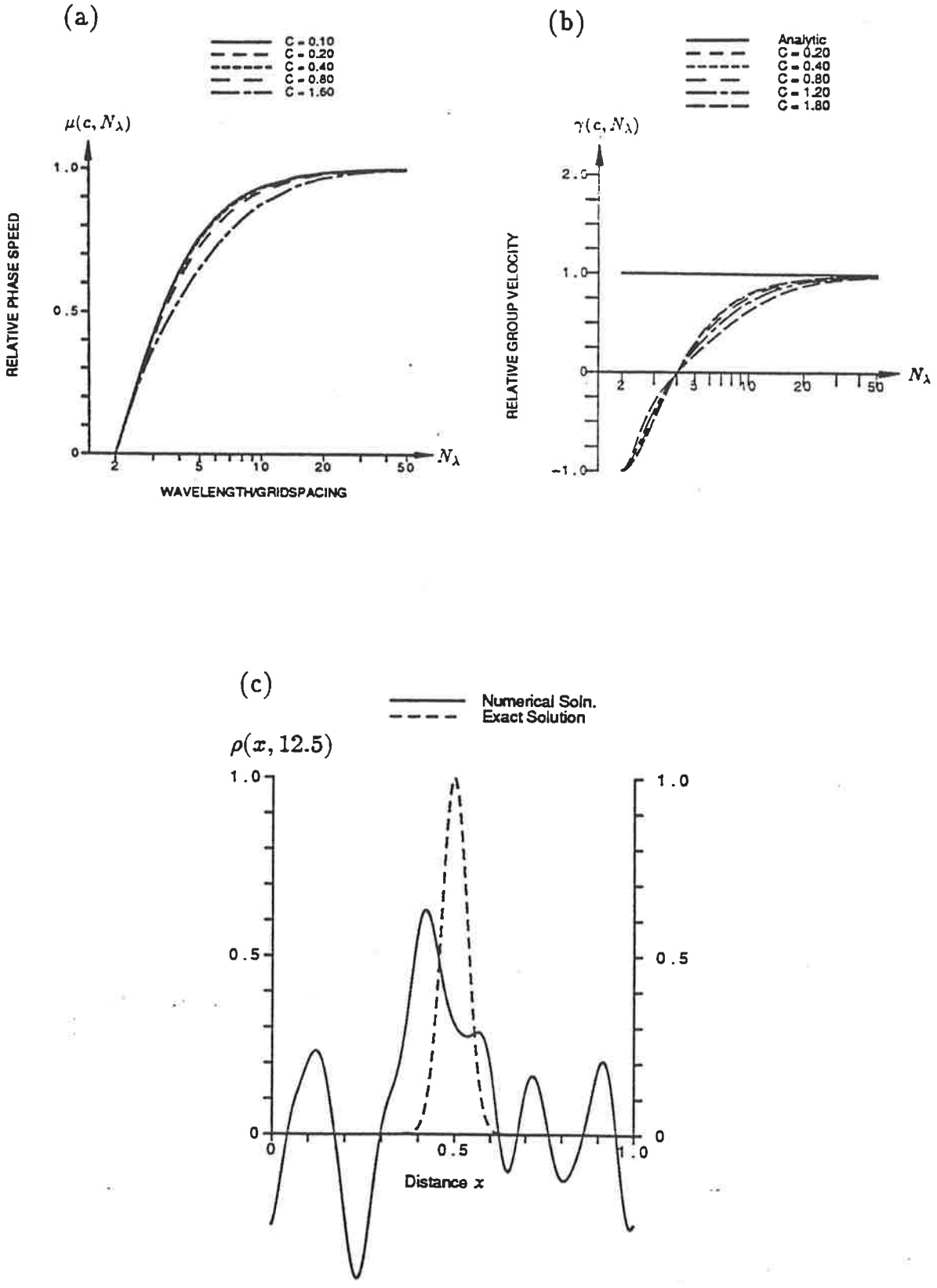

Figure 3.1: Illustration of the performance of Crank Nicolson scheme, Eq. (3.8). The diagrams show (a) the relative phase speed, μ, (b) the relative group velocity, γ, and (c) the results of the the numerical test case with a Gaussian peak as the initial condition and cyclic boundary conditions.

the wave propagation characteristics shown in Fig. 3.2(a,b) demonstrate that this improvement (also shown in the numerical results in Fig. 3.2(c)) is obtained for all values of the Courant number. Again, the weak variation of the wave propagation characteristics with Courant number shows the comparisons can be expected to hold for most values of Courant number, despite only being calculated for $c = 0.4$.

From Table 3.1 it can be seen that this scheme is slightly worse than third order upwinding but is a significant improvement on the Lax-Wendroff and Crank Nicolson schemes. The errors are due to the shift in the position of the peak caused by the phase lag of the short wavelength Fourier components. This lag also introduces a skew to the peak as shown in the error of the third moment.

3.1.3 Fourth Order Centred-Time Centred-Space (CTCS)

The Linear Finite Element Crank Nicolson Scheme, as discussed in the previous section, is fourth order in Δx but only second order in Δt. Since it uses a six-point computational stencil it should be possible to produce a completely fourth order scheme by use of the modified equivalent equation. Since any centred scheme will automatically comprise only even powers of Δx and Δt, if one free parameter is introduced into the discretization then this could be used to eliminate all the second order terms leaving only the fourth and higher order terms. One such way is to allow $\theta_{-1} = \theta_1 = \theta$ and $\theta_0 = 1 - 2\theta$ in Eq. (3.2). That the sum $\theta_{-1} + \theta_0 + \theta_1$ must equal unity follows directly from the consistence condition, Eq. (3.2). Converting the difference equation into a modified equivalent equation gives

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = -\frac{u(\Delta x)^2}{3!}\left(-2\theta + \frac{1}{3} + \frac{c^2}{6}\right)\frac{\partial^2 \rho}{\partial x^2} + O\{(\Delta x)^4\} \quad . \tag{3.13}$$

Putting $\theta = (2 + c^2)/12$ gives the fourth order scheme

$$\begin{aligned} \rho_j^{n+1} &+ \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^{n+1} + \rho_j^{n+1}) - \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^{n+1} + \rho_{j-1}^{n+1}) \\ &+ \frac{2 + c_{j+\frac{1}{2}}^2}{12}(\rho_{j+1}^{n+1} - \rho_j^{n+1}) - \frac{2 + c_{j-\frac{1}{2}}^2}{12}(\rho_j^{n+1} - \rho_{j-1}^{n+1}) \\ = \rho_j^n &- \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^n + \rho_j^n) - \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^n + \rho_{j-1}^n) \end{aligned}$$

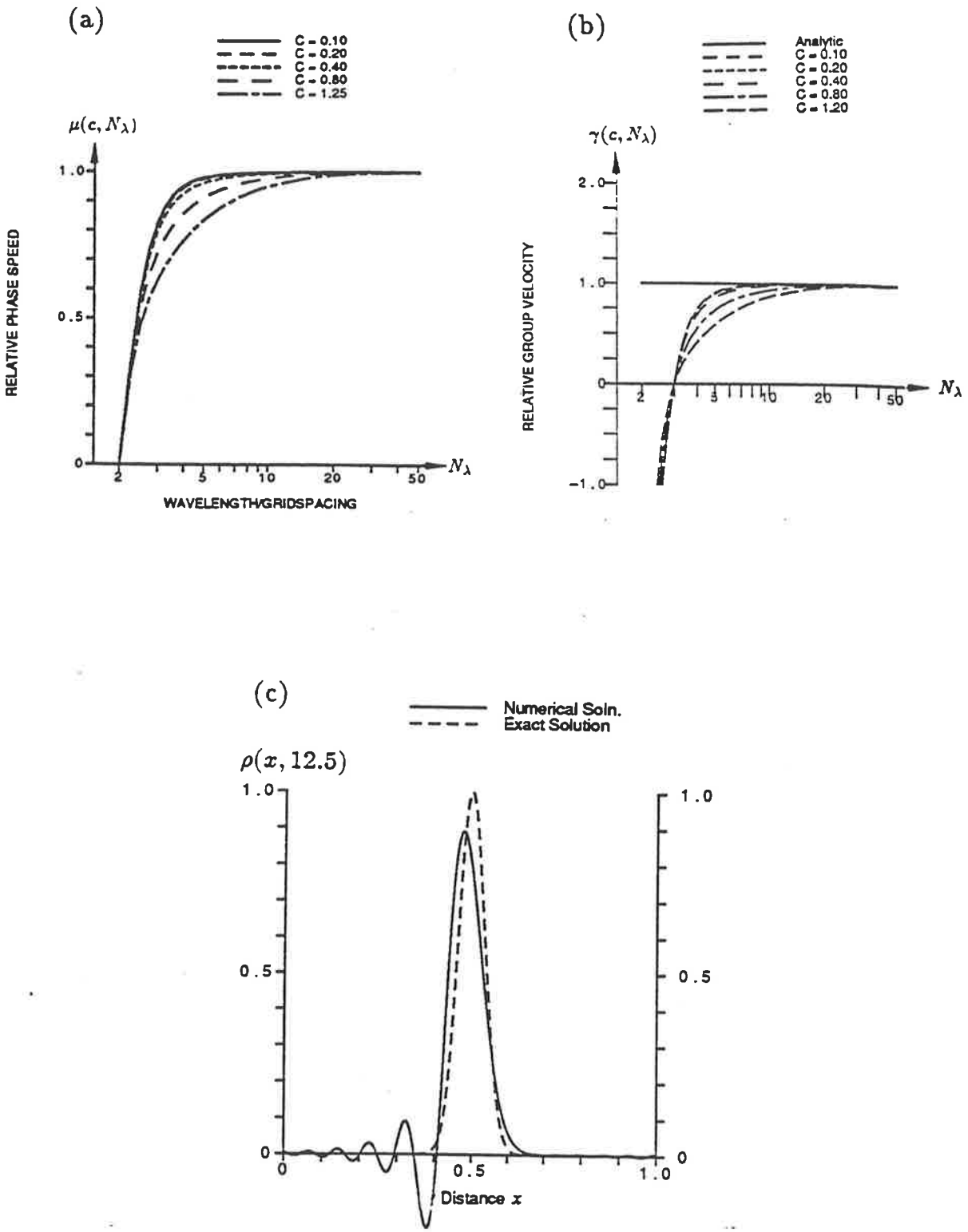

Figure 3.2: As in Fig. 3.1 but for the Linear Finite Element Crank Nicolson scheme, Eq. (3.11).

$$+ \quad \frac{2+c_{j+\frac{1}{2}}^2}{12}(\rho_{j+1}^n - \rho_j^n) - \frac{2+c_{j-\frac{1}{2}}^2}{12}(\rho_j^n - \rho_{j-1}^n) \tag{3.14}$$

with modified equivalent equation

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u(\Delta x)^4}{720}(4-c^2)(1-c^2)\frac{\partial^5 \rho}{\partial x^5} + O\{(\Delta x)^6\} \quad . \tag{3.15}$$

This scheme requires slightly more computational time than the second order Crank Nicolson scheme due to the more complicated coefficients. The bulk of the computational effort is spent in the solution of the tri-diagonal system, which is the same for any scheme. There are, however, significant gains in accuracy to be made using this scheme, as can be seen in Fig. 3.3(c).

This scheme is unconditionally von Neumann stable, however it is only diagonally dominant provided $|c| \leq 1$. While lack of diagonal dominance does not necessarily mean that the solution of the system of equations is unstable, it does in this case. By simply using this scheme with a Courant number slightly greater than one, it becomes readily apparent that the calculations are unstable, but since the scheme is stable in the von Neumann sense, it must be the inversion that has become unstable. This scheme may be rewritten in terms of values at x_{j-2}, x_{j-1} and x_j rather than x_{j-1}, x_j and x_{j+1} and the scheme may then be marched. It is not obvious that this marching technique is in fact stable, but it is not difficult to show that this is the case, by considering the propagation of errors. Consider a general three-point recursion relation of the form

$$A_{-2}\rho_{j-2} + A_{-1}\rho_{j-1} + A_0\rho_j = d_j \tag{3.16}$$

where A_{-2}, A_{-1}, A_0 and d_j are all known, and let ξ_j be the difference between the computed solution and the true solution. Then ξ_j satisfies the homogeneous equation

$$A_{-2}\xi_{j-2} + A_{-1}\xi_{j-1} + A_0\xi_j = 0 \quad . \tag{3.17}$$

The general solution of this equation is given by substituting $\xi_j = \lambda^j$, where λ is some constant. This gives a quadratic in λ, and the solution is

$$\xi_j = A\lambda_1^j + B\lambda_2^j \tag{3.18}$$

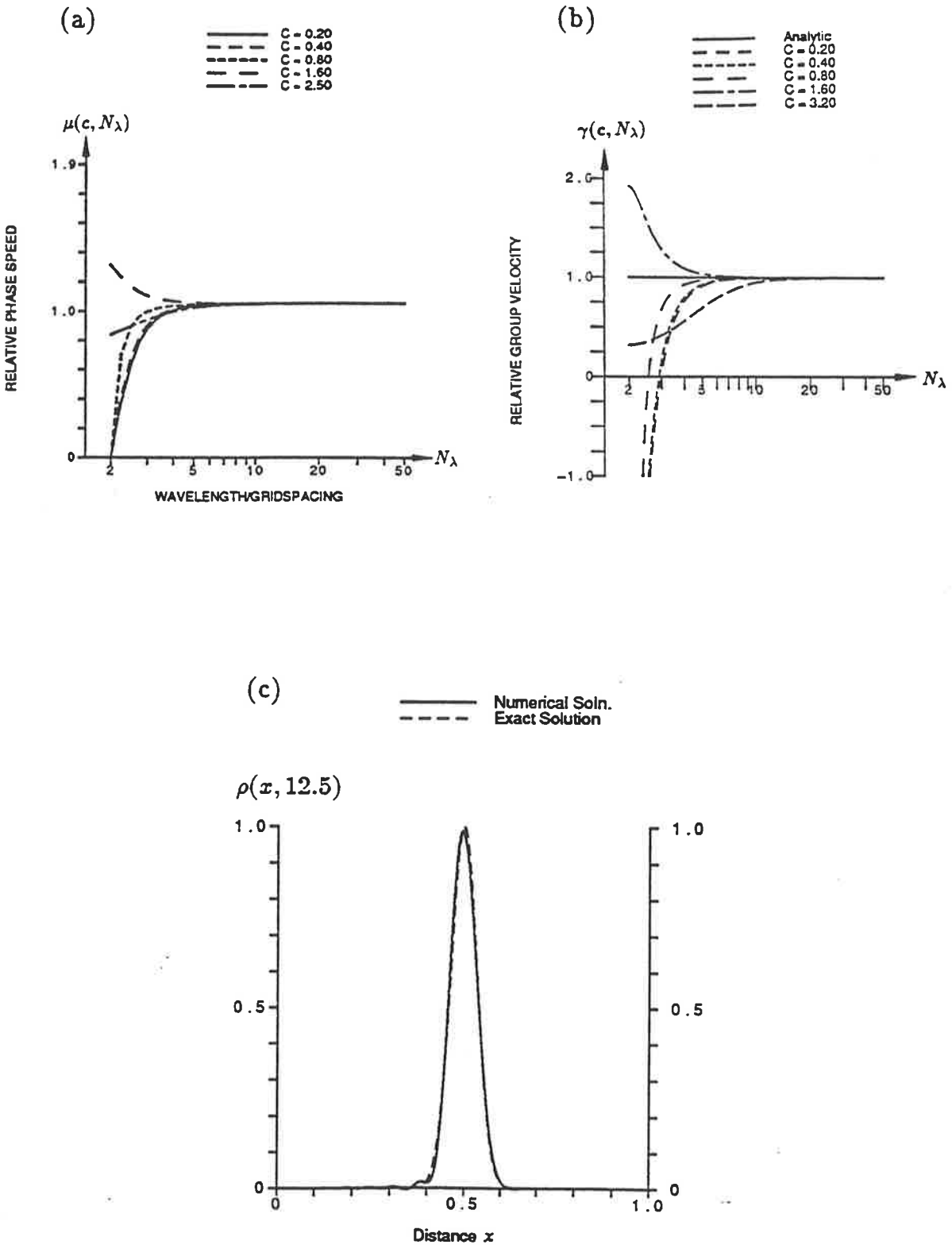

Figure 3.3: As in Fig. 3.1 but for the Fourth Order CTCS method, Eq. (3.14). This scheme is unconditionally stable.

where λ_1 and λ_2 are the two roots of the quadratic for λ, or in the case where $\lambda_1 = \lambda_2$,

$$\xi_j = (Aj + B)\lambda^j \tag{3.19}$$

The marching scheme requires knowledge of two points, ρ_0^{n+1} and ρ_1^{n+1}, a priori, and the method by which these two values are calculated determines the values of A and B. If λ_1 and λ_2 are distinct then ξ_j will be bounded provided λ_1 and λ_2 lie within or on the unit circle. In the case of equal roots, the roots must lie strictly within the unit circle. In the case of the fourth order scheme Eq. (3.14), the roots are distinct and lie within the unit circle provided $c \geq 1$. For the case where $c \leq -1$ the scheme may also be marched, except now Eq. (3.14) is marched in the reverse direction, from right to left. So, this scheme may be used for any value of Courant number provided the two methods of solution can be matched together.

The matching of the two schemes poses little difficulty. Where there is a transition from the marching scheme to the inversion scheme, the marched solution can provide a Dirichlet type boundary condition for use with the inversion scheme. Where there is a transition from the inversion scheme to the marching scheme, another method for determining the value with which to close off the inversion scheme is required. This can be any other scheme, for example, the linear Finite Element Crank Nicolson scheme can be used to give two equations for the last three unknowns in the system of linear equations, allowing one to be eliminated. Alternatively, one of the explicit schemes can be used as the overall accuracy of the scheme does not appear to be particularly sensitive to the scheme used. The marching scheme can be used from this point on. If the transition point is particularly significant for some other reason (which may be due to the physics of the problem) then a slight change in the time-step will move the transition point to a less significant position.

The Crank Nicolson and Linear Finite Element Crank Nicolson schemes discussed earlier also have restricted regions of diagonal dominance, namely $c \leq 2$ and $c \leq 4/3$, respectively, however, they are never stable when used as marching schemes.

In these two cases, diagonal dominance does not appear to be important since both schemes have been run with c greater than six for more than ten thousand time-steps without any indication of instability. The matrix inversions associated with the use of these three-point implicit schemes all appear to be unconditionally stable but this can only be formally shown for the fourth order scheme. The fourth order scheme has the additional advantage that when it is marched it has the effective computational speed of an explicit scheme.

As is shown in the graph of the numerical solution, Fig. 3.3(c), and in the comparison of errors, Table 3.1, the main source of the error is the slight shifting of the peak due to the phase lag of the Fourier components. This is also evident in the wave propagation characteristics shown in Fig. 3.3(a,b).

Comparisons between these three-point implicit schemes shows a consistent improvement as the order of the modified equivalent equation is increased, with the fourth order scheme being a significant improvement on the original Crank Nicolson scheme. The fourth order scheme is also seen to be an improvement on the two explicit fourth order schemes discussed in Chapter Two. This is apparent from comparing Figs. 2.7, 2.6 with Fig. 3.3 and the error measures in Table 3.1. This is also reflected in the modified equivalent equations. The functions $\eta_5(c)$ for the fourth order upwinding, Rusanov's fourth order scheme and fourth order centred-time, centred-space are:

$$\eta_5^{UW4}(c) = -(3-c)(2-c)(1-c) \tag{3.20}$$

$$\eta_5^{RUS}(c) = (2-c^2)(1-c^2) \tag{3.21}$$

$$\eta_5^{CN4}(c) = \frac{(4-c^2)(1-c^2)}{6} \tag{3.22}$$

and it can be shown that $\left|\eta_5^{CN4}(c)\right| \leq \left|\eta_5^{RUS}(c)\right|$ for all $|c| \leq 1$, corresponding to the range of stability for Rusanov's fourth order scheme and $\left|\eta_5^{CN4}(c)\right| \leq \left|\eta_5^{UW4}(c)\right|$ for $0 \leq c \leq 1.52$, which is over most of the stable range of fourth order upwinding.

3.2 Five-Point Implicit Schemes

The trailing oscillations produced by the fourth order centred-time, centred-space scheme, Eq. (3.14), are still larger than one would like, but to decrease the size of these oscillations using linear difference equations requires an even higher order scheme as demonstrated in Chapter Two. In the same chapter it was also shown, via the use of the modified equivalent equation, that a scheme using six grid-points can be at best fourth order. The derivation of this result is still valid for implicit schemes and so to obtain further improvement on these results requires a wider computational stencil. For this reason some five-point implicit schemes will be investigated.

3.2.1 Khaliq and Twizell's Schemes

The schemes discussed here were originally developed for use with fixed boundary conditions and have been modified to allow for cyclic boundary conditions.

Khaliq and Twizell (1982) showed that if the standard two-point central difference formula is used to approximate the spatial derivative in Eq. (2.2) then the following system of ordinary differential equations is obtained,

$$\frac{d}{dt}\underset{\sim}{\hat{\rho}} = -\frac{1}{2}\frac{w}{\Delta x}\mathbf{B}\underset{\sim}{\hat{\rho}} + O\left\{(\Delta x)^2\right\} \tag{3.23}$$

where

$$\underset{\sim}{\hat{\rho}} = [\hat{\rho}(\Delta x, t), \hat{\rho}(2\Delta x, t), \ldots, \hat{\rho}(J\Delta x, t)]^t \quad , \tag{3.24}$$

with $[]^t$ denoting transpose and

$$\mathbf{B} = \begin{bmatrix} 0 & 1 & & & & & 0 & -1 \\ -1 & 0 & 1 & & & & & 0 \\ & -1 & 0 & 1 & & & & \\ & & & \ddots & & & & \\ & & & & \ddots & & & \\ & & & & & \ddots & & \\ & & & -1 & 0 & 1 & & \\ 0 & & & & -1 & 0 & 1 & \\ 1 & 0 & & & & -1 & 0 & \end{bmatrix} \quad . \tag{3.25}$$

Using Eq. (3.23), Khaliq and Twizell derived the recurrence relation

$$\underset{\sim}{\rho}(t + \Delta t) = \underset{\sim}{\rho}(t) \exp\left(-\frac{w\Delta t}{2\Delta x}\mathbf{B}\right) \quad . \tag{3.26}$$

Using Padé approximants to the exponential matrix function gives rise to a family of finite difference methods of various orders of accuracy in Δt, although they are only ever second order in Δx.

Any Padé approximant to the exponential function may be used, in theory, to produce a two-level finite difference scheme but, in practice, there is a limited range. The form of the matrix $\mathbf{B}$ means that the (m,p) approximant leads to a numerical scheme that involves $2m+1$ values at the t^{n+1} level and $(2p+1)$ values at the level t^n. To derive a high order approximation, m or p must be greater than one, but if either exceeds three then the scheme becomes computationally expensive as well as giving rise to difficulties near boundaries. For this reason Khaliq and Twizell chose the (2,0), (2,1) and (2,2) Padé approximants to the exponential function. Using the (2,0) approximant gives

$$\begin{aligned}\rho_j^{n+1} &+ \frac{c_{j+\frac{1}{2}}}{2}(\rho_{j+1}^{n+1}+\rho_j^{n+1}) \\ &- \frac{c_{j-\frac{1}{2}}}{2}(\rho_j^{n+1}+\rho_{j-1}^{n+1}) \\ &+ \frac{c_{j+\frac{1}{2}}^2}{8}(\rho_{j+2}^{n+1}+\rho_{j+1}^{n+1}-\rho_j^{n+1}-\rho_{j-1}^{n+1}) \\ &- \frac{c_{j-\frac{1}{2}}^2}{8}(\rho_{j+1}^{n+1}+\rho_j^{n+1}-\rho_{j-1}^{n+1}-\rho_{j-2}^{n+1}) = \rho_j^n \end{aligned} \tag{3.27}$$

which is an approximation to Eq. (2.2) of $O\{(\Delta x)^2,(\Delta t)^2\}$. If $\rho_j^{(1)}$ is the approximation to $\hat{\rho}_j^{n+2}$ obtained by using Eq. (3.27) twice (once to go from time-level t^n to t^{n+1} and then again to time-level t^{n+2}) and $\rho_j^{(2)}$ is the approximation to the same value but obtained by using Eq. (3.27) once with double the time-step (and hence also the Courant number), then these two values may be extrapolated to give

$$\rho_j^{(E)} = \frac{4}{3}\rho_j^{(1)} - \frac{1}{3}\rho_j^{(2)} \tag{3.28}$$

where $\rho_j^{(E)}$ is now an $O\{(\Delta x)^2,(\Delta t)^4\}$ approximation to $\hat{\rho}_j^{n+2}$. This extrapolation is similar to the "deferred approach to the limit" of Richardson and Gaunt (1927).

Using the (2,1) Padé approximant gives

$$\rho_j^{n+1} + \frac{c_{j+\frac{1}{2}}}{3}(\rho_{j+1}^{n+1}+\rho_j^{n+1}) - \frac{c_{j-\frac{1}{2}}}{3}(\rho_j^{n+1}+\rho_{j-1}^{n+1})$$

$$
\begin{aligned}
&+ \frac{c_{j+\frac{1}{2}}^2}{24}(\rho_{j+2}^{n+1} + \rho_{j+1}^{n+1} - \rho_j^{n+1} - \rho_{j-1}^{n+1}) \\
&- \frac{c_{j-\frac{1}{2}}^2}{24}(\rho_{j+1}^{n+1} + \rho_j^{n+1} - \rho_{j-1}^{n+1} - \rho_{j-2}^{n+1}) \\
&= \rho_j^n - \frac{c_{j+\frac{1}{2}}}{6}(\rho_{j+1}^n + \rho_j^n) + \frac{c_{j-\frac{1}{2}}}{6}(\rho_j^n + \rho_{j-1}^n)
\end{aligned} \tag{3.29}
$$

with the corresponding extrapolation being

$$
\rho_j^{(E)} = \frac{8}{7}\rho_j^{(1)} - \frac{1}{7}\rho_j^{(2)} \tag{3.30}
$$

which is of $O\{(\Delta x)^2, (\Delta t)^5\}$. The (2,2) approximant yields

$$
\begin{aligned}
\rho_j^{n+1} \quad + \quad & \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^{n+1} + \rho_j^{n+1}) - \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^{n+1} + \rho_{j-1}^{n+1}) \\
& + \frac{c_{j+\frac{1}{2}}^2}{48}(\rho_{j+2}^{n+1} + \rho_{j+1}^{n+1} - \rho_j^{n+1} - \rho_{j-1}^{n+1}) \\
& - \frac{c_{j-\frac{1}{2}}^2}{48}(\rho_{j+1}^{n+1} + \rho_j^{n+1} - \rho_{j-1}^{n+1} - \rho_{j-2}^{n+1}) \\
= \rho_j^n \quad & - \frac{c_{j+\frac{1}{2}}}{4}(\rho_{j+1}^n + \rho_j^n) + \frac{c_{j-\frac{1}{2}}}{4}(\rho_j^n + \rho_{j-1}^n) \\
& + \frac{c_{j+\frac{1}{2}}^2}{48}(\rho_{j+2}^n + \rho_{j+1}^n - \rho_j^n - \rho_{j-1}^n) \\
& - \frac{c_{j-\frac{1}{2}}^2}{48}(\rho_{j+1}^n + \rho_j^n - \rho_{j-1}^n - \rho_{j-2}^n)
\end{aligned} \tag{3.31}
$$

with

$$
\rho_j^{(E)} = \frac{16}{15}\rho_j^{(1)} - \frac{1}{16}\rho_j^{(2)} \tag{3.32}
$$

being the associated $O\{(\Delta x)^2, (\Delta t)^6\}$ extrapolation. The equivalent continuous forms of the unextrapolated operators are

$$
\begin{aligned}
\mathcal{L}^{(KT0)}\{\rho_j^n\} \equiv & \left[\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x}\right. \\
& \left. - \frac{u(\Delta x)^2}{6}(c^2 - 1)\frac{\partial^3 \rho}{\partial x^3} + O\{(\Delta x)^3\}\right]_j^n
\end{aligned} \tag{3.33}
$$

$$
\begin{aligned}
\mathcal{L}^{(KT1)}\{\rho_j^n\} \equiv & \left[\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x}\right. \\
& \left. + \frac{u(\Delta x)^2}{6}\frac{\partial^3 \rho}{\partial x^3} + \frac{u(\Delta x)^3}{72}c^3\frac{\partial^4 \rho}{\partial x^4} + O\{(\Delta x)^4,\ (\Delta t)^4\}\right]_j^n
\end{aligned} \tag{3.34}
$$

and

$$
\mathcal{L}^{(KT2)}\{\rho_j^n\} \equiv \left[\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x}\right. \tag{3.35}
$$

$$+\frac{u(\Delta x)^2}{6}\frac{\partial^3\rho}{\partial x^3}+O\{(\Delta x)^4\}\Big]_j^n \quad ,$$

confirming that these approximations are $O\{(\Delta x)^2,(\Delta t)^2\}$, $O\{(\Delta x)^2,(\Delta t)^3)\}$ and $O\{(\Delta x)^2,(\Delta t)^4\}$. The operators on the right hand side of the previous expressions refer to the finite difference operators, and so an equivalence sign is used rather than an equal sign when relating the difference operators to the equivalent continuous operators. It should be noted that the finite difference operators and the modified equivalent operators may differ by a multiplicative constant. This comes from the coefficients in the finite difference operators being rescaled to remove common factors in each term. These rescaling factors only become important when other terms are included in the p.d.e being modelled, and are readily computed as they are simply the coefficient of $\partial\rho/\partial t$ in the equivalent p.d.e.

The modified equivalent equation for these schemes is obtained by letting these operators equal zero. The unextrapolated forms are unconditionally stable, but they are only diagonally dominant for $c \leq \sqrt{3}-1, \sqrt{10}-2$ and $\sqrt{21}-3$ respectively, although these appear to be overly conservative estimates on the stability of the matrix inversion since all three of these schemes have been run with $c > 6$ for more than ten thousand time steps with no sign of instabilities appearing.

Although the two integrations involved in this process are stable (in the von Neumann sense), it does not automatically imply that the extrapolations Eqs. (3.28, 3.30, 3.32) are stable. It was assumed by Khaliq and Twizell that this was the case, but as demonstrated by Noye and Steinle (1986), this is not always so. Since the extrapolation, by its very nature, is not a convex combination of the two solutions, $\rho_j^{(1)}$ and $\rho_j^{(2)}$, there must be a potential for instability. The calculation of the effect of the extrapolation on the stability of the process is described below.

Consider an arbitrary Fourier component of the solution of wavenumber m, and let $\beta = m\Delta x$. Let $G_1(c,\beta)$ and $G_2(c,\beta)$ be the von Neumann amplification factors of $\rho_j^{(1)}$ and $\rho_j^{(2)}$, respectively, then

$$G_2(c,\beta) = G_1(2c,\beta) \tag{3.36}$$

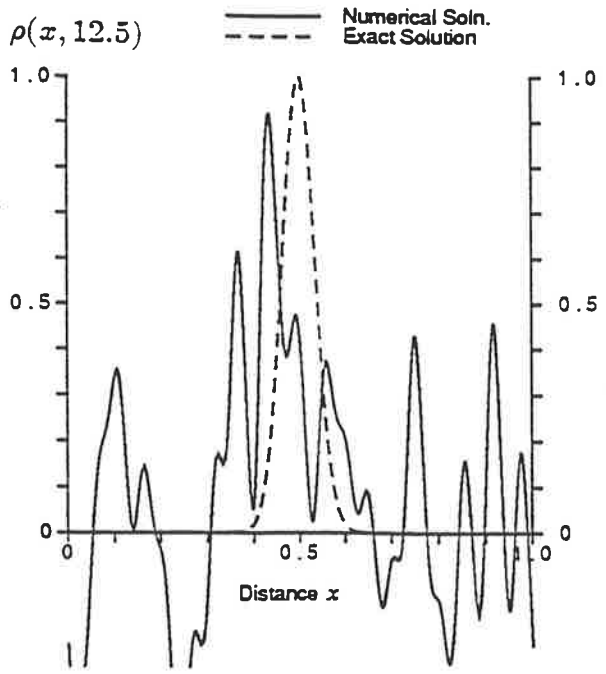

Figure 3.4: Illustration of Khaliq and Twizell's (2,0) extrapolated scheme Eq. (3.28) applied to the Gaussian test case.

and $G_E(c, \beta)$, the von Neumann amplification factors for $\rho_j^{(E)}$, is given by

$$G_E = \sqrt{\theta(G_1)^2 + (1-\theta)G_2} \tag{3.37}$$

where θ is the weight used in the extrapolation. Taking the square root normalizes G_E to be the net amplification factor after one time-step. So for stability, it is required that $|G_E(c, \beta)| \leq 1$ for all β which, in the case of the Eqs. (3.27, 3.28), is not true for any $c > 0$ although the other schemes Eqs. (3.29, 3.30) and Eqs. (3.31, 3.32) are unconditionally stable. The instability in Eq. (3.28) is quite weak and takes some time to grow sufficiently to dominate the solution in the constant coefficient case examined here. In non-linear cases, however, this instability may be more severe. The numerical results of this scheme, applied to the test case described in Chapter Two is shown in Fig. 3.4. By comparing this figure with Fig. 3.5(d) it is evident that the instability of the extrapolation is beginning to dominate the results after 2500 time-steps.

The wave propagation parameters and the numerical test are shown in Figs.3.6-3.9. These results are very similar to those presented for the Lax-Wendroff scheme, and the Crank Nicolson scheme, shown in Figs. 2.2 and 3.1, respectively. The er-

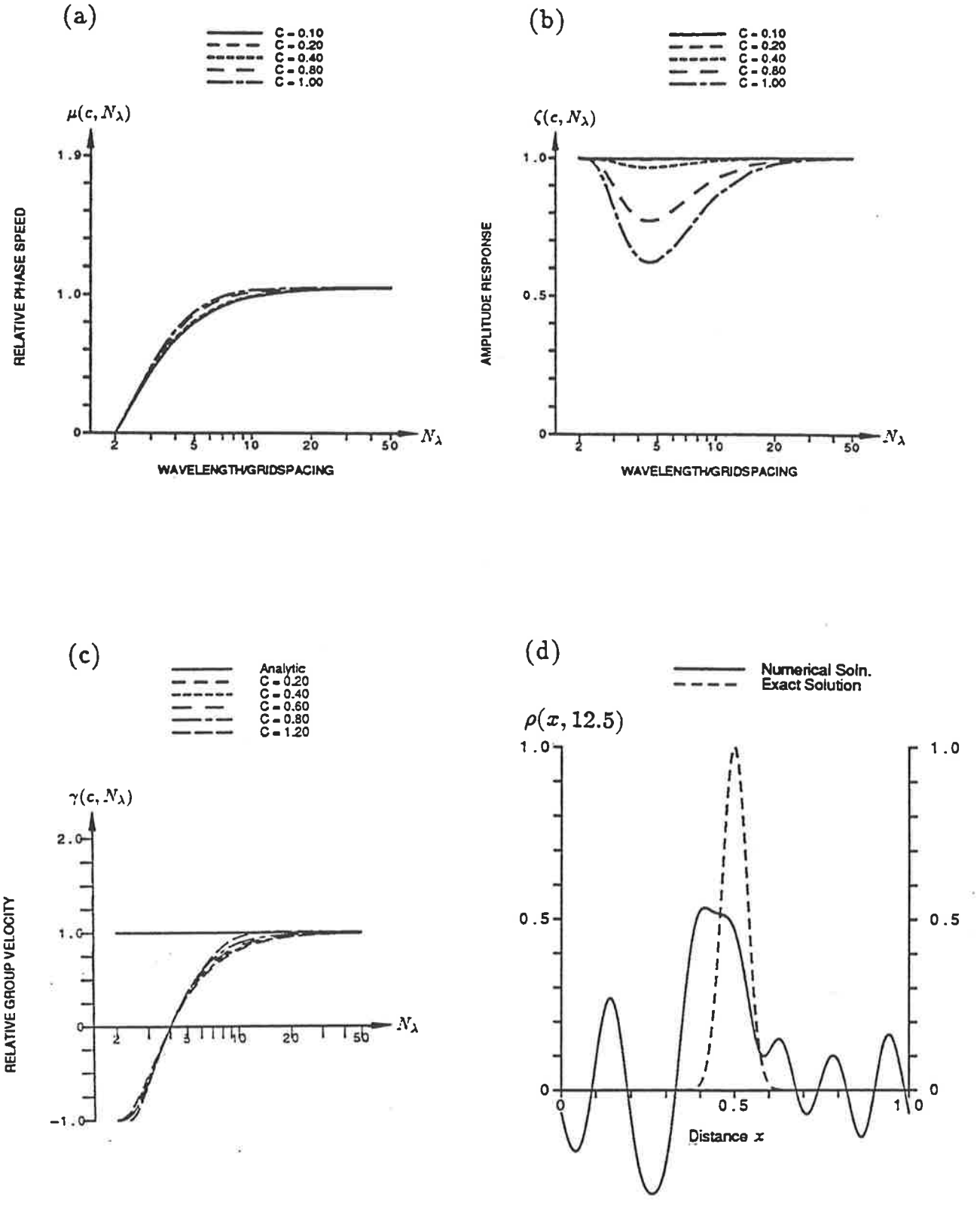

Figure 3.5: As in Fig. 2.1 but for Khaliq and Twizell's (2,0) scheme, Eq. (3.27).

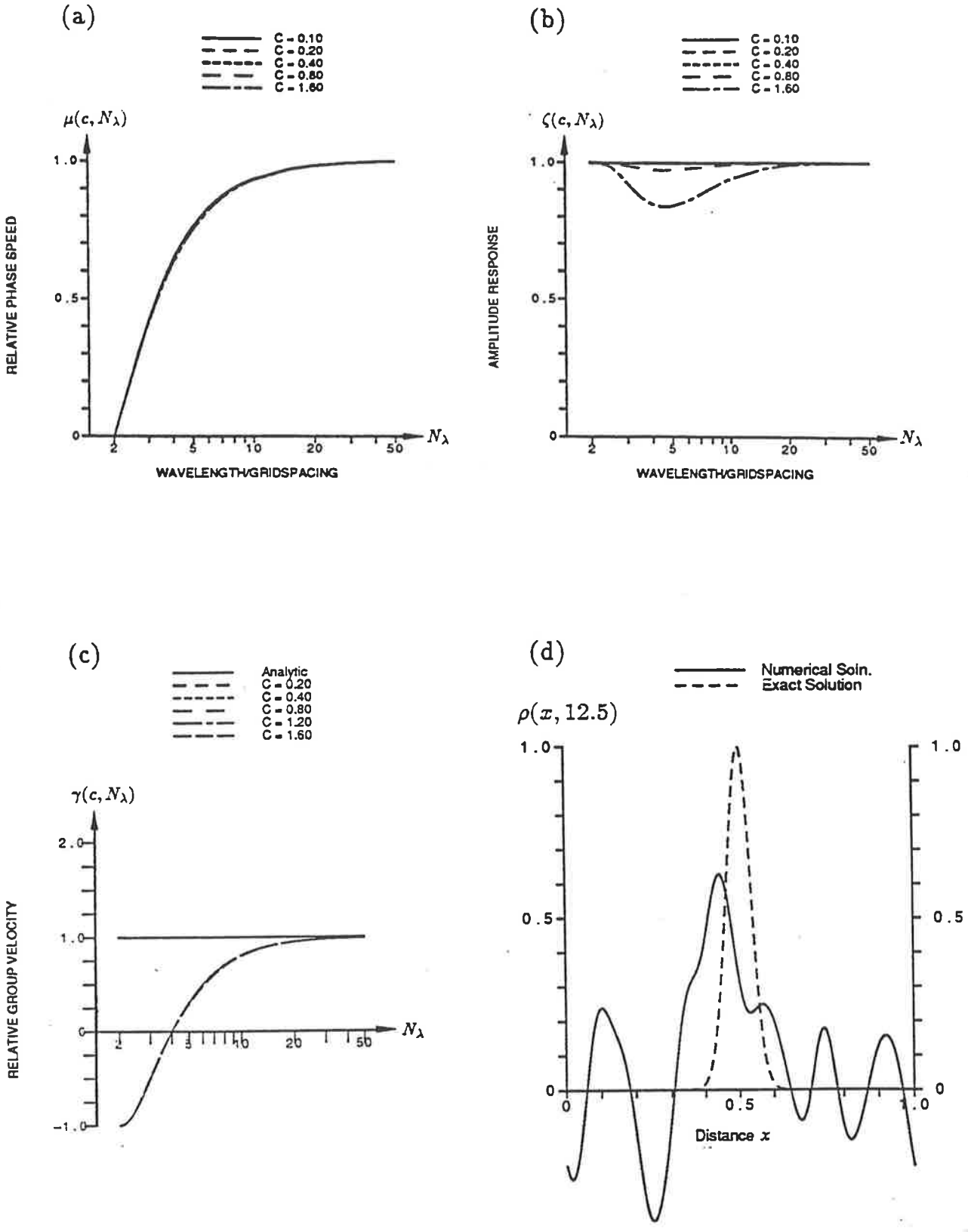

Figure 3.6: As in Fig. 2.1 but for Khaliq and Twizell's (2,1) scheme, Eq. (3.29).

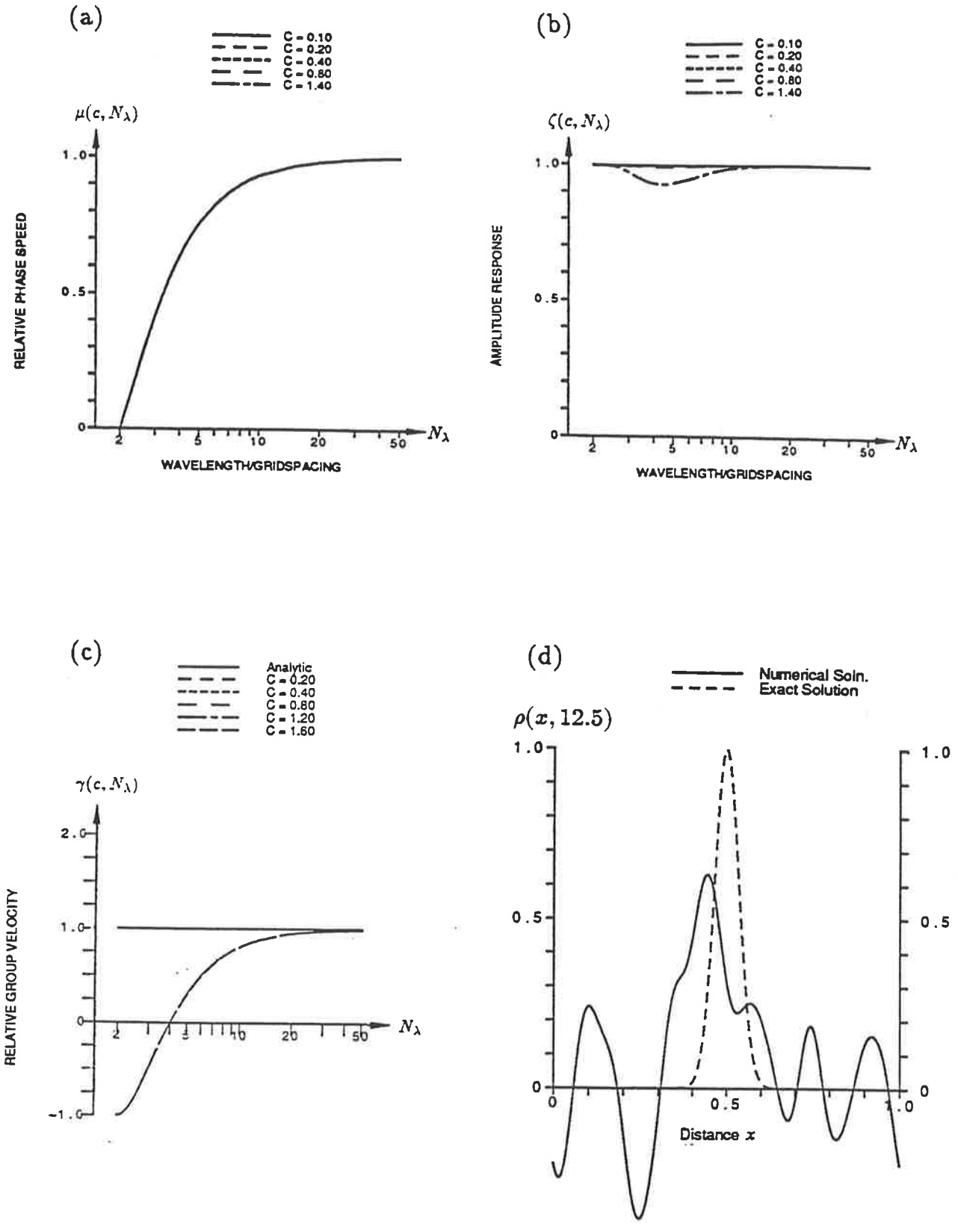

Figure 3.7: As in Fig. 2.1 but for Khaliq and Twizell's (2,1) extrapolated scheme, Eq. (3.30).

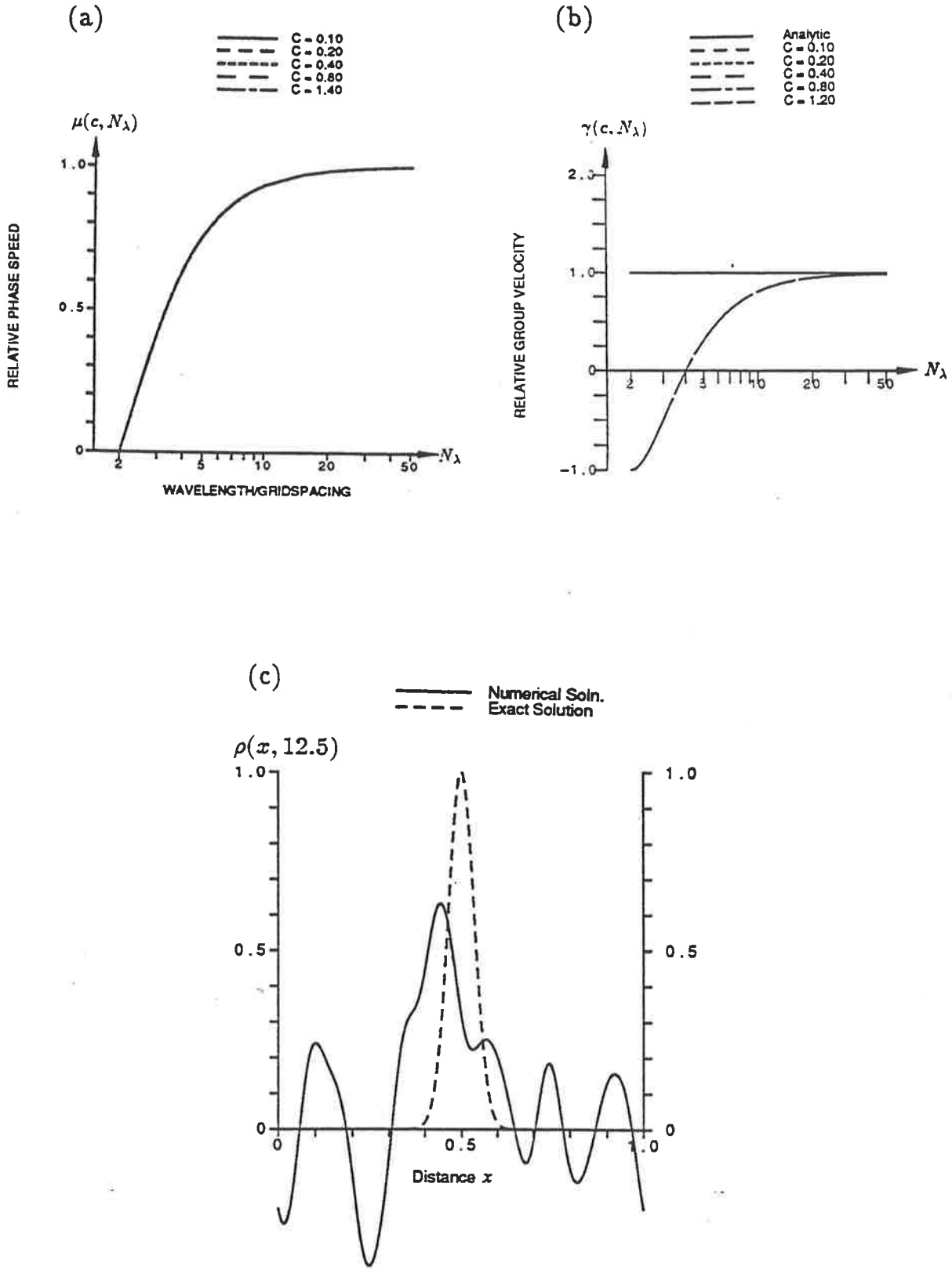

Figure 3.8: As in Fig. 3.1 but for Khaliq and Twizell's (2,2) scheme, Eq. (3.31).

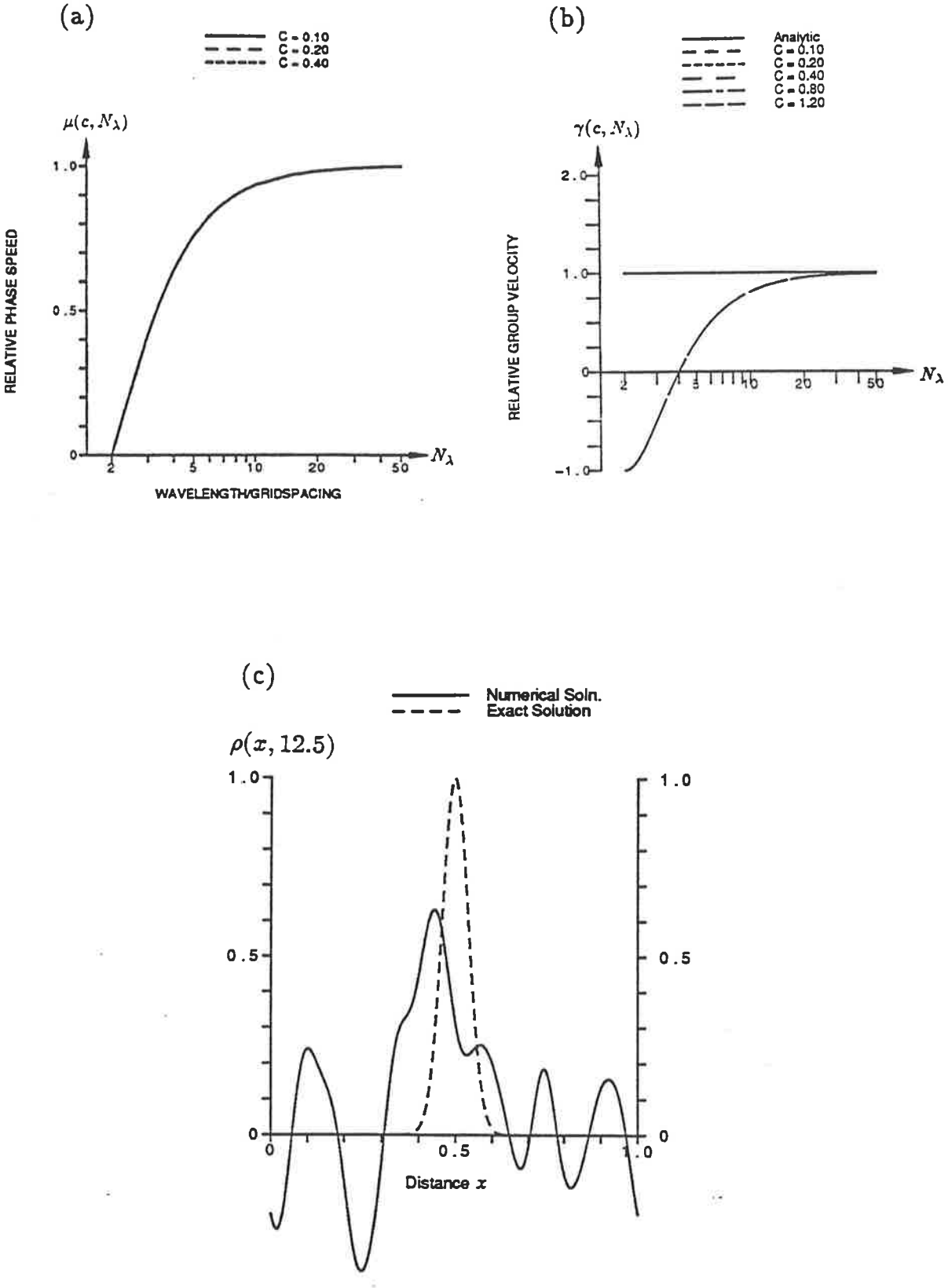

Figure 3.9: As in Fig. 3.1 but for Khaliq and Twizell's (2,2) extrapolated scheme, Eq. (3.32).

ror measures presented in Table 3.1 suggest these schemes are, if anything, slightly worse than the Lax-Wendroff scheme. Using the (2,1) Padé approximant gives some improvement over using the (2,0) approximant. This is as expected from the corresponding modified equivalent equations, since the leading residual term of Eq. (3.33) is greater than that in Eq. (3.34). This is in contrast with the use of the (2,2) Padé approximant rather than the (2,1) approximant, as now the change is negligible. Again, this is consistent with the modified equivalent equations for the two schemes, (Eq. (3.34) and Eq. (3.35)) in which the leading terms are identical. It should be noted that the results presented here are representative of the general performance of these schemes, since the wave propagation parameters shown in Figs. 3.5-3.7(a-c) and Figs. 3.8, 3.9 (a,b) show little variation with Courant number, c, and so the poor results are not due to the particular choice of parameters used in the test case.

Given the increase in computational effort due to the inversion of the penta-diagonal system of linear equations, these results seem quite poor in comparison with nearly all other schemes. The reason seems to be that although the approximations to the temporal derivative (involving terms in powers of Δt) are of high order, the approximations to the spatial derivatives are still only second order and these dominate to such an extent that results are typical of those of second order schemes.

It can be shown by this method that the three implicit methods derived by Khaliq and Twizell are only second order, despite the extrapolation. For example, if $\mathcal{L}^{(KT0)}$ is as defined in Eq. (3.33), the difference operator $\mathcal{L}^{(KT0e)}$ that corresponds to the extrapolation can be obtained explicitly, namely

$$\begin{aligned}\mathcal{L}^{(KT0e)}\{\rho_j^n\} \;=\; & c^4\rho_{j-4}^{n+2} - 8c^3\rho_{j-3}^{n+2} + 4c^2(6-c^2)\rho_{j-2}^{n+2} \\ & -24c(2-c^2)\rho_{j-1}^{n+2} + 6(8-8c^2+c^4)\rho_j^{n+2} \\ & +24c(2-c^2)\rho_{j+1}^{n+2} + 4c^2(6-c^2)\rho_{j+2}^{n+2} + 8c^3\rho_{j+3}^{n+2} \\ & +c^4\rho_{j+4}^{n+2} - 48\rho_j^n\end{aligned} \tag{3.38}$$

$$\equiv \left[\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \frac{u(\Delta x)^2}{6}\frac{\partial^3 \rho}{\partial x^3} - 4c^3\frac{u(\Delta x)^3}{24}\frac{\partial^4 \rho}{\partial x^4} + O\{(\Delta x)^4\}\right]_j^n \quad .$$

As expected the extrapolation, Eq. (3.28), leads to an increase in the order of the scheme in Δt but the second order terms in Δx remain. The other difference operators have modified equivalent equations of the form

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \frac{u(\Delta x)^2}{6}\frac{\partial^3 \rho}{\partial x^3} + O\{(\Delta x)^3\} = 0 \quad . \tag{3.39}$$

The extrapolations, Eqs. (3.30, 3.32), have no effect on the term involving $(\Delta x)^2$ as there is no refinement of the grid size, only the time-step is altered between the two results. The terms involving $(\Delta x)^2$ cannot be neglected or considered to be very much less than Δt, since to be of much practical use Δx must be of similar or lower order than Δt. To increase the order in Δx using Khaliq and Twizell's development, a higher order approximation to $\partial\hat{\rho}/\partial x$ must be used which would result in the matrix **B** having more off-diagonal elements requiring more than five values at the $(n+1)^{st}$ time-level, which is undesirable for the reasons given earlier.

3.3 High Order Implicit Schemes

As was discussed earlier, one way of developing higher order (and therefore more accurate) schemes is to remove each of the terms from the modified equivalent equation in turn. This can be done by taking the weighted average of two schemes of similar order, obtaining the modified equivalent equation for the combined scheme and selecting a value of the weight that removes the leading coefficient of the residual. This process is described in Noye and Hayman (1986) and will be used here to derive some high order and very accurate finite difference schemes as discussed in Noye and Steinle (1986).

The modified equivalent equation approach thus provides a method of using the difference schemes of Khaliq and Twizell and obtaining methods that are very high

order in Δx as well as Δt. To apply the modified equation approach, the difference schemes must be written in the form

$$\mathcal{L}\{\rho_j^n\} = 0 \tag{3.40}$$

where $\mathcal{L}$ is the corresponding difference operator, and all terms are expanded in a Taylor series about the point (x_j, t^n) to give the equivalent partial differential equation. From this, the modified equivalent partial differential equation is obtained.

Using the modified equation approach and ignoring the extrapolations, it is possible to construct schemes that are higher order in Δx but require no more values than those used in Eq. (3.31). The coefficients of these equations are more complicated, but since there is no longer any need for the extrapolations (which require the extra set of calculations every second time-step), the net saving in computational effort is significant.

The operators $\mathcal{L}^{(KT0)}$ and $\mathcal{L}^{(KT1)}$, defined in Eqs. (3.33, 3.34) are both $O\{(\Delta x)^2\}$, so by taking a linear combination of the two it is possible to eliminate the second order terms in the modified equivalent equation. It can be shown that the linear combination

$$\begin{aligned}
\mathcal{L}^{(NS1)}\{\rho_j^n\} &= \frac{1}{c}\left[3\mathcal{L}^{(KT0)} + (c^2-1)\mathcal{L}^{(KT1)}\right]\{\rho_j^n\} \qquad (3.41)\\
&\equiv \left[\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \right.\\
&\qquad \left. \frac{u(\Delta x)^3}{72}\frac{(2+c^2)^2}{c}\frac{\partial^4 \rho}{\partial x^4} + O\{(\Delta x)^4\}\right]_j^n
\end{aligned}$$

does precisely this. This third order operator gives rise to the difference equation

$$\begin{aligned}
24c_{j+\frac{1}{2}}\rho_j^{n+1} &+ 4(1+2c_{j+\frac{1}{2}}^2)(\rho_{j+1}^{n+1} + \rho_j^{n+1})\\
&- 4(1+2c_{j-\frac{1}{2}}^2)(\rho_j^{n+1} + \rho_{j-1}^{n+1})\\
&+ c_{j+\frac{1}{2}}(2+c_{j+\frac{1}{2}}^2)(\rho_{j+2}^{n+1} + \rho_{j+1}^{n+1} - \rho_j^{n+1} - \rho_{j-1}^{n+1})\\
&- c_{j-\frac{1}{2}}(2+c_{j-\frac{1}{2}}^2)(\rho_{j+1}^{n+1} + \rho_j^{n+1} - \rho_{j-1}^{n+1} - \rho_{j-2}^{n+1}) \qquad (3.42)\\
= \quad & 24c_{j+\frac{1}{2}}\rho_j^n + 4(1-c_{j+\frac{1}{2}}^2)(\rho_{j+1}^n + \rho_j^n)\\
&+ 4(1-c_{j-\frac{1}{2}}^2)(\rho_j^n + \rho_{j-1}^n) \quad .
\end{aligned}$$

Another third order scheme can be derived by taking a linear combination of the two equations Eqs. (3.29, 3.31), in order to eliminate the second order terms in their differential forms, namely

$$\begin{aligned}\mathcal{L}^{(NS2)}\{\rho_j^n\} &= \frac{1}{c}\left[6\mathcal{L}^{(KT0)} + (c^2-1)\mathcal{L}^{(KT2)}\right]\{\rho_j^n\} \\ &\equiv \left[\frac{\partial\rho}{\partial t} + u\frac{\partial\rho}{\partial x} + \right. \\ &\qquad \left. \frac{u(\Delta x)^3}{24}\frac{(2+c^2)}{c}\frac{\partial^4\rho}{\partial x^4} + O\{(\Delta x)^4\}\right]_j^n\end{aligned} \tag{3.43}$$

which corresponds to the difference equation

$$\begin{aligned}48c_{j+\frac{1}{2}}\rho_j^{n+1} &+ 12(1+c_{j+\frac{1}{2}}^2)(\rho_{j+1}^{n+1}+\rho_j^{n+1}) \\ &- 12(1+c_{j-\frac{1}{2}}^2)(\rho_j^{n+1}+\rho_{j-1}^{n+1}) \\ &+ c_{j+\frac{1}{2}}(5+c_{j+\frac{1}{2}}^2)(\rho_{j+2}^{n+1}+\rho_{j+1}^{n+1}-\rho_j^{n+1}-\rho_{j-1}^{n+1}) \\ &- c_{j-\frac{1}{2}}(5+c_{j-\frac{1}{2}}^2)(\rho_{j+1}^{n+1}+\rho_j^{n+1}-\rho_{j-1}^{n+1}-\rho_{j-2}^{n+1}) \\ = \quad & 48c_{j+\frac{1}{2}}\rho_j^n + 12(1-c_{j+\frac{1}{2}}^2)(\rho_{j+1}^n+\rho_j^n) \\ & -12(1-c_{j-\frac{1}{2}}^2)(\rho_j^n+\rho_{j-1}^n) \\ & -c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}}^2)(\rho_{j+2}^n+\rho_{j+1}^n-\rho_j^n-\rho_{j-1}^n) \\ & +c_{j-\frac{1}{2}}(1-c_{j-\frac{1}{2}}^2)(\rho_{j+1}^n+\rho_j^n-\rho_{j-1}^n-\rho_{j-2}^n) \quad .\end{aligned} \tag{3.44}$$

It is possible to continue this process of taking linear combinations of difference operators still further, since the last two schemes are both third order in Δx and Δt. Thus we may form the fourth order approximation

$$\begin{aligned}\mathcal{L}^{(NS3)}\{\rho_j^n\} &= \left[-6\mathcal{L}^{(NS1)} + (2+c^2)\mathcal{L}^{(NS2)}\right]\{\rho_j^n\} \\ &\equiv \left[\frac{\partial\rho}{\partial t} + u\frac{\partial\rho}{\partial x}\right. \\ &\qquad \left. +\frac{u(\Delta x)^4}{720}(34+10c^2+c^4)\frac{\partial^5\rho}{\partial x^5} + O\{(\Delta x)^6\}\right]_j^n\end{aligned} \tag{3.45}$$

from which we obtain the difference equation

$$48c_{j+\frac{1}{2}}\rho_j^{n+1} + 12c_{j+\frac{1}{2}}(\rho_{j+1}^{n+1}+\rho_j^{n+1}) - 12c_{j-\frac{1}{2}}(\rho_j^{n+1}+\rho_{j-1}^{n+1})$$

$$
\begin{aligned}
&+ \quad (2+c_{j+\frac{1}{2}}^2)(\rho_{j+2}^{n+1}+\rho_{j+1}^{n+1}-\rho_j^{n+1}-\rho_{j-1}^{n+1}) \\
&- \quad (2+c_{j-\frac{1}{2}}^2)(\rho_{j+1}^{n+1}+\rho_j^{n+1}-\rho_{j-1}^{n+1}-\rho_{j-2}^{n+1}) \qquad (3.46) \\
&= \quad 48c_{j+\frac{1}{2}}\rho_j^n - 12c_{j+\frac{1}{2}}(\rho_{j+1}^n+\rho_j^n) + 12c_{j-\frac{1}{2}}(\rho_j^n+\rho_{j-1}^n) \\
&\quad +(2+c_{j+\frac{1}{2}}^2)(\rho_{j+2}^n+\rho_{j+1}^n-\rho_j^n-\rho_{j-1}^n) \\
&\quad -(2+c_{j-\frac{1}{2}}^2)(\rho_{j+1}^n+\rho_j^n-\rho_{j-1}^n-\rho_{j-2}^n) \quad .
\end{aligned}
$$

The wave propagation parameters and results of using this scheme on the Gaussian pulse test case are presented for completeness in Fig. 3.10(a,c) and Table 3.1. Clearly the scheme is not as accurate as the fourth order, implicit scheme Eq. (3.14) discussed earlier. This is also reflected in the modified equivalent equation for this scheme, which can be obtained by setting Eq. (3.45) equal to zero. A comparison of the leading term in the residual of this equation

$$
\eta_5^{NS3}(c) = \frac{34+10c^2+c^4}{6} \qquad (3.47)
$$

with that of the fourth order three-point implicit scheme, namely

$$
\eta_5^{CN4}(c) = \frac{(4-c^2)(1-c^2)}{6} \qquad (3.48)
$$

shows that the leading term for the three-point scheme is always less than that for NS3.

To continue further, another fourth order scheme is required. This obviously cannot be obtained by taking other combinations of the previous methods since Eq. (3.46) would be obtained again. The fourth order difference operator $\mathcal{L}^{CN4}$ corresponding to the fourth order scheme Eq. (3.14) may be used to construct the sixth order scheme

$$
\begin{aligned}
\mathcal{L}^{(NS4)}\{\rho_j^n\} &= \left[(1-c^2)(4-c^2)\mathcal{L}^{(NS3)} - 4(34+10c^2+c^4)\mathcal{L}^{CN4}\right]\{\rho_j^n\} \\
&\equiv \left[\frac{\partial\rho}{\partial t} + u\frac{\partial\rho}{\partial x}\right. \qquad (3.49) \\
&+ \quad \left. u(\Delta x)^6\frac{(c^4+c^2-20)(c^2-1)}{30240}\frac{\partial^7\rho}{\partial x^7} + O\{(\Delta x)^8\}\right]_j^n \quad .
\end{aligned}
$$

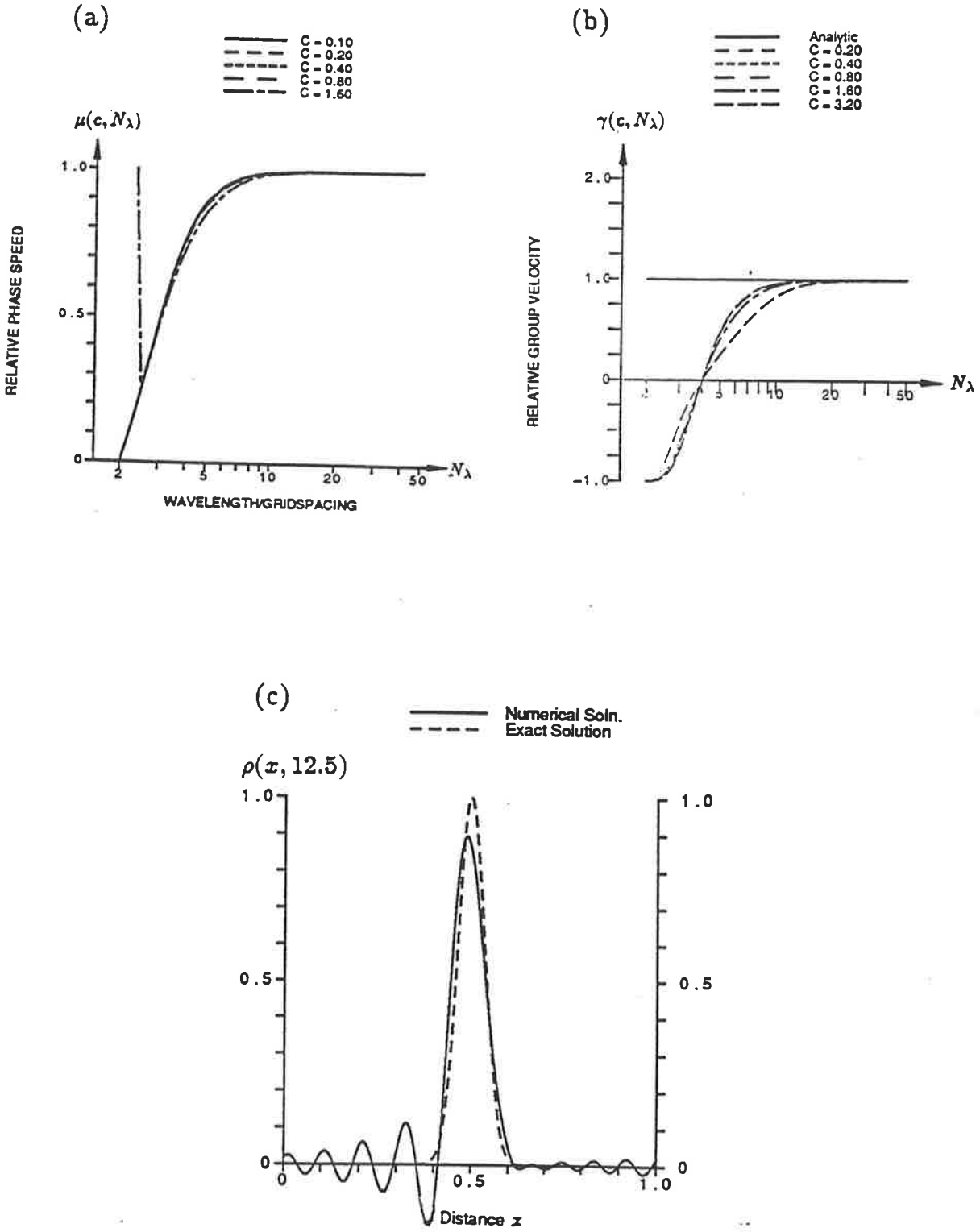

Figure 3.10: As in Fig. 3.1 but for the scheme NS3, Eq. (3.46).

This operator gives rise to the finite difference equation

$$
\begin{aligned}
720\rho_j^{n+1} \quad &+ \quad 180c_{j+\frac{1}{2}}(\rho_{j+1}^{n+1}+\rho_j^{n+1}) - 180c_{j-\frac{1}{2}}(\rho_j^{n+1}+\rho_{j-1}^{n+1}) \\
&+ \quad \left(136+4c_{j+\frac{1}{2}}^2(1+c_{j+\frac{1}{2}}^2)\right)(\rho_{j+1}^{n+1}-\rho_j^{n+1}) \\
&- \quad \left(136+4c_{j-\frac{1}{2}}^2(1+c_{j-\frac{1}{2}}^2)\right)(\rho_j^{n+1}-\rho_{j-1}^{n+1}) \\
&- \quad \left(4-c_{j+\frac{1}{2}}^2(5-c_{j+\frac{1}{2}}^2)\right)(\rho_{j+2}^{n+1}+\rho_{j+1}^{n+1}-\rho_j^{n+1}-\rho_{j-1}^{n+1}) \\
&+ \quad \left(4-c_{j-\frac{1}{2}}^2(5-c_{j-\frac{1}{2}}^2)\right)(\rho_{j+1}^{n+1}+\rho_j^{n+1}-\rho_{j-1}^{n+1}-\rho_{j-2}^{n+1}) \\
= \quad 720\rho_j^n \quad &-180c_{j+\frac{1}{2}}(\rho_{j+1}^n+\rho_j^n) + 180c_{j-\frac{1}{2}}(\rho_j^n+\rho_{j-1}^n) \\
&+ \quad \left(136+4c_{j+\frac{1}{2}}^2(1+c_{j+\frac{1}{2}}^2)\right)(\rho_{j+1}^n-\rho_j^n) \qquad (3.50) \\
&- \quad \left(136+4c_{j-\frac{1}{2}}^2(1+c_{j-\frac{1}{2}}^2)\right)(\rho_j^n-\rho_{j-1}^n) \\
&- \quad \left(4-c_{j+\frac{1}{2}}^2(5-c_{j+\frac{1}{2}}^2)\right)(\rho_{j+2}^n+\rho_{j+1}^n-\rho_j^n-\rho_{j-1}^n) \\
&+ \quad \left(4-c_{j-\frac{1}{2}}^2(5-c_{j-\frac{1}{2}}^2)\right)(\rho_{j+1}^n+\rho_j^n-\rho_{j-1}^n-\rho_{j-2}^n) \quad .
\end{aligned}
$$

Since these last equations are diagonally differenced, they are von Neumann stable for all values of c, however the system of algebraic equations is only diagonally dominant for $c \leq 1$. This scheme cannot be marched and for c greater than about 1.1, weak instabilities begin to appear after five to ten thousand time-steps. For $c \geq 1.3$, these instabilities appear after relatively few time-steps. The schemes derived, using the modified equation approach, are higher order in Δx but require no more values than those used in Eq. (3.31). The coefficients of these latter equations are more complicated, but since there is no longer any need for the extrapolations (which require the extra set of calculations every second time-step) the net saving in computational effort is significant. It should also be noted that since most of the computational time is spent solving the penta-diagonal system of equations, the effect of making the coefficients more complicated is not very significant.

Eliminating the two leading terms from the modified equivalent equations, Eqs. (3.33, 3.34, 3.35), produces a substantial improvement as shown in the wave propagation parameters for NS3 in Fig. 3.10(a-b). As N_λ increases, the relative phase speed now increases much more rapidly for small N_λ. This is reflected in the improved results in

the numerical test case shown in Fig. 3.10(c), where the oscillations are significantly smaller than those produced by the schemes of Khaliq and Twizell. By creating a scheme that is now sixth order, as is NS4, the oscillations are almost totally removed, as seen in Fig. 3.11(c), as now all Fourier components with wavelength greater than $N_\lambda = 4$ are advected almost perfectly. The oscillations still exist, as seen is Table 3.1 where the most negative number in the numerical solution is -3.6×10^{-5}, or less than 1% of the size of the oscillations produced by fifth order upwinding and five orders of magnitude less than the height of the pulse.

From the result in Chapter One, the modified equation approach applied to a ten point stencil is capable of producing an eighth order scheme. To derive such a scheme, approximate the time derivative by

$$\left.\frac{\partial \hat{\rho}}{\partial t}\right]_j^{n+\frac{1}{2}} \simeq \gamma T\rho \Big]_{j-2} + \theta T\rho \Big]_{j-1} + (1-2\theta-2\gamma)T\rho \Big]_j + \theta T\rho \Big]_{j+1} + \gamma \; T\rho \Big]_{j+2} \tag{3.51}$$

where

$$T\rho \Big]_j = (\rho_j^{n+1} - \rho_j^n)/\Delta t \quad . \tag{3.52}$$

The spatial derivative may be approximated by

$$\left.\frac{\partial \hat{\rho}}{\partial x}\right]_j^{n+\frac{1}{2}} = \frac{\beta}{2} S_2\rho\Big]^{n+1} + \frac{\beta}{2} S_2\rho\Big]^{n} + \frac{1-\beta}{2} S_4\rho\Big]^{n+1} + \frac{1-\beta}{2} S_4\rho\Big]^{n} \tag{3.53}$$

where

$$S_2\rho\Big]^n = \frac{\rho_{j+1}^n - \rho_{j-1}^n}{2\Delta x} \tag{3.54}$$

and

$$S_4\rho\Big]^n = \frac{\rho_{j-2}^n - 8\rho_{j-1}^n + 8\rho_{j+1}^n - \rho_{j+2}^n}{12\Delta x} \quad . \tag{3.55}$$

The first three terms of the residual are given by

$$\eta_3(c) = \frac{1}{6} - \frac{1}{6}\beta - 4\gamma - \theta + \frac{1}{12}c^2 \tag{3.56}$$

$$\eta_5(c) = \frac{1}{120} - \frac{1}{24}\beta + \frac{2}{3}\gamma\beta - 2\gamma + \frac{1}{6}\theta\beta - \frac{1}{4}\theta - \frac{1}{24}\beta c^2 + \frac{1}{24}c^2 + 16\gamma^2 + 8\theta\gamma - \gamma c^2 + \theta^2 \tag{3.57}$$

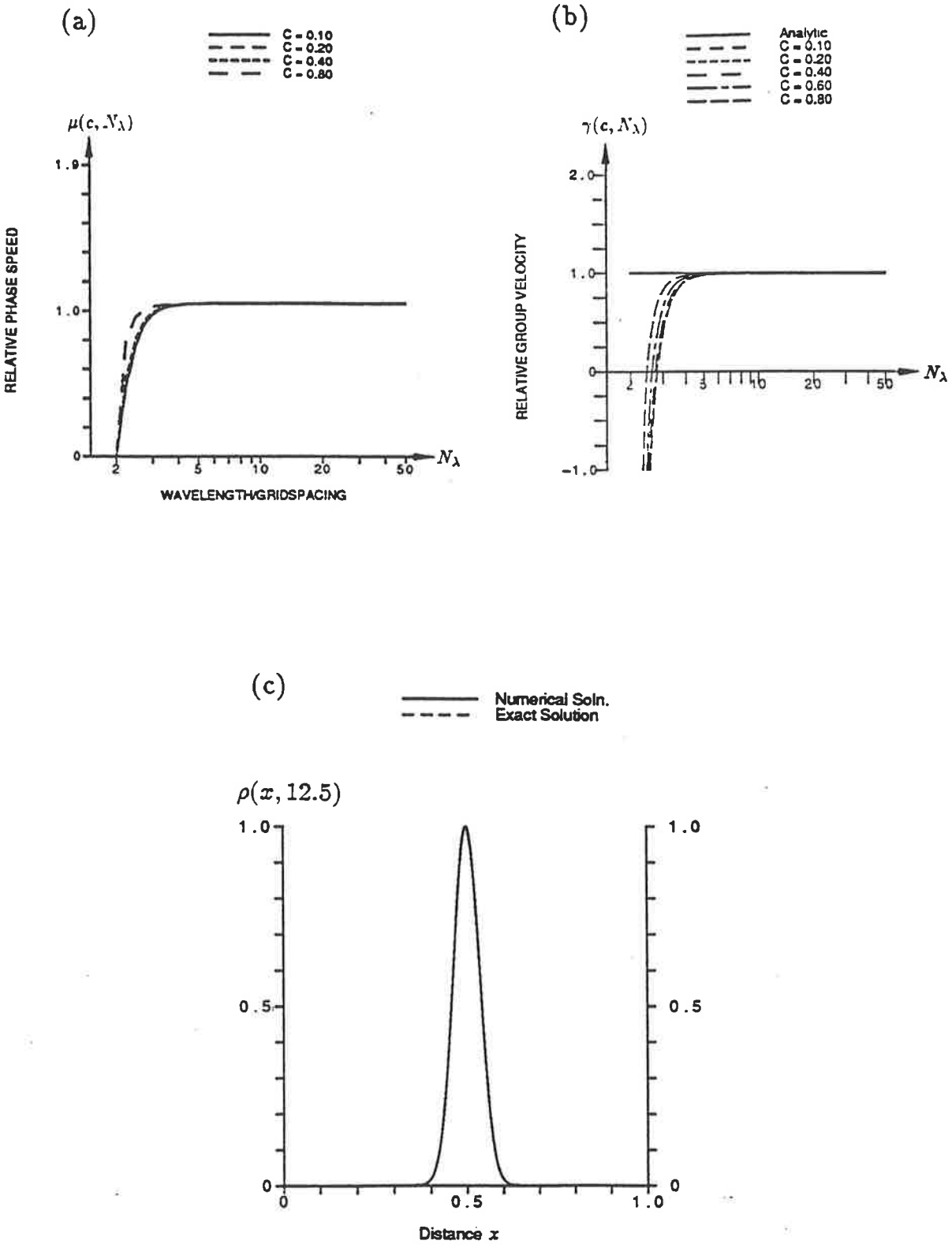

Figure 3.11: As in Fig. 3.1 but for the scheme NS4, Eq. (3.50). This scheme is unstable for $c > 1.1$.

$$
\begin{aligned}
& - \frac{1}{4}\theta c^2 + \frac{1}{80}c^4 \\
\eta_7(c) &= \frac{1}{5040} - \frac{1}{240}\beta + \frac{7}{18}\gamma\beta - \frac{13}{30}\gamma + \frac{1}{18}\theta\beta - \frac{1}{40}\theta \\
& - \frac{7}{288}\beta + \frac{13}{1440}c^2 + \frac{40}{3}\gamma^2 + \frac{14}{3}\theta\gamma - \frac{5}{6}\gamma c^2 + \frac{1}{3}\theta^2 \\
& - \frac{7}{48}\theta c^2 + \frac{1}{144}\beta^2 c^2 + \frac{1}{2}\gamma\beta c^2 + \frac{1}{8}\theta\beta c^2 - \frac{1}{96}\beta c^4 + \frac{1}{96}c^4 \\
& - \frac{8}{3}\gamma^2\beta - 64\gamma^3 - 48\theta\gamma^2 + 8\gamma^2 c^2 - \frac{4}{3}\theta\gamma\beta - 12\theta^2\gamma \\
& + 4\theta\gamma c^2 - \frac{1}{4}\gamma c^4 - \frac{1}{6}\theta^2\beta - \theta^3 + \frac{1}{2}\theta^2 c^2 - \frac{1}{16}\theta c^4 + \frac{1}{448}c^6 \quad .
\end{aligned} \tag{3.58}
$$

An eighth order method is found by solving the three simultaneous equations $\eta_3(c) = \eta_5(c) = \eta_7(c) = 0$ for the three weights β, γ and θ. This gives [1]

$$
\begin{aligned}
\theta &= -(c^4 - 10c^2 - 96)/420 \\
\gamma &= (c^4 + 35c^2 + 24)/1680 \\
\beta &= -(c^2 + 5)/7
\end{aligned} \tag{3.59}
$$

and the resulting scheme is

$$
\begin{aligned}
\rho_j^{n+1} &+ \gamma_{j+\frac{1}{2}}(\rho_{j+2}^{n+1} + \rho_{j+1}^{n+1} - \rho_j^{n+1} - \rho_{j-1}^{n+1}) \\
&- \gamma_{j-\frac{1}{2}}(\rho_{j+1}^{n+1} + \rho_j^{n+1} - \rho_{j-1}^{n+1} - \rho_{j-2}^{n+1}) \\
&+ \frac{c\beta_{j+\frac{1}{2}}}{24}(\rho_{j+2}^{n+1} + \rho_{j+1}^{n+1} - \rho_j^{n+1} - \rho_{j-1}^{n+1}) \\
&- \frac{c\beta_{j-\frac{1}{2}}}{24}(\rho_{j+1}^{n+1} + \rho_j^{n+1} - \rho_{j-1}^{n+1} - \rho_{j-2}^{n+1}) \\
&+ \theta_{j+\frac{1}{2}}(\rho_{j+1}^{n+1} - \rho_j^{n+1}) - \theta_{j-\frac{1}{2}}(\rho_j^{n+1} - \rho_{j-1}^{n+1}) \\
&+ \frac{c_{j+\frac{1}{2}}}{84}(16 + c_{j+\frac{1}{2}}^2)(\rho_{j+1}^{n+1} + \rho_j^{n+1}) \\
&- \frac{c_{j-\frac{1}{2}}}{84}(16 + c_{j-\frac{1}{2}}^2)(\rho_j^{n+1} + \rho_{j-1}^{n+1}) \\
= \rho_j^n &+ \gamma_{j+\frac{1}{2}}(\rho_{j+2}^n + \rho_{j+1}^n - \rho_j^n - \rho_{j-1}^n) \\
&- \gamma_{j-\frac{1}{2}}(\rho_{j+1}^n + \rho_j^n - \rho_{j-1}^n - \rho_{j-2}^n) \\
&- \frac{c\beta_{j+\frac{1}{2}}}{24}(\rho_{j+2}^n + \rho_{j+1}^n - \rho_j^n - \rho_{j-1}^n) \\
&+ \frac{c\beta_{j-\frac{1}{2}}}{24}(\rho_{j+1}^n + \rho_j^n - \rho_{j-1}^n - \rho_{j-2}^n)
\end{aligned} \tag{3.60}
$$

[1] The solution of these equations was found using the symbolic manipulation program, MACSYMA, a product of Symbolics Inc.

$$+ \quad \theta_{j+\frac{1}{2}}(\rho^n_{j+1} - \rho^n_j) - \theta_{j-\frac{1}{2}}(\rho^n_j - \rho^n_{j-1})$$
$$- \quad \frac{c_{j+\frac{1}{2}}}{84}(16 + c^2_{j+\frac{1}{2}})(\rho^n_{j+1} + \rho^n_j) + \frac{c_{j-\frac{1}{2}}}{84}(16 + c^2_{j-\frac{1}{2}})(\rho^n_j + \rho^n_{j-1}) \quad .$$

This scheme, NS5, is now eighth order in the case of constant velocity. This system of equations is only diagonally dominant for $c \leq 0.40129..$, but it has been run for more than ten thousand time-steps for values of c as large as 1.2 without any sign of instability. Beyond $c = 1.2$, instabilities begin to appear more quickly. It appears very likely that the two schemes NS4 and NS5 may both be used for $0 \leq c \leq 1$ and possibly slightly further.

The performance of this scheme is shown in Fig. 3.12(a-c). The relative phase speed for Fourier components resolved by three or more grid-points is almost ideal, while the relative group velocity for components resolved by four or more grid-points is also very close to unity and none of these components are in any way damped. The results of the Gauss pulse test case are almost perfect, with the errors being beyond the resolution of the diagrams. The errors are only evident in Table 3.1 which shows the oscillations to be about one-millionth of those produced by fifth order upwinding. The errors in the first and third moments have begun to be affected by machine precision.

3.4 Further Numerical Tests

So far, all schemes have been tested on infinitely differentiable initial conditions (a Gaussian pulse) and it is clear that these can be advected most successfully, provided that a sufficiently high order scheme is used. In many problems, however, abrupt changes in the solution are possible. For this reason, other initial conditions have been used such as the square wave (e.g. by Boris and Book, 1973) and cosine hill (e.g. Morton, 1985). These were chosen since the first and second derivatives become singular at certain points. Another initial condition that has been used (Steinle and Morrow, 1989 and Leonard, 1990) is a semi-ellipse, which has the advantage over a square wave in that there is still some structure to the solution away from the

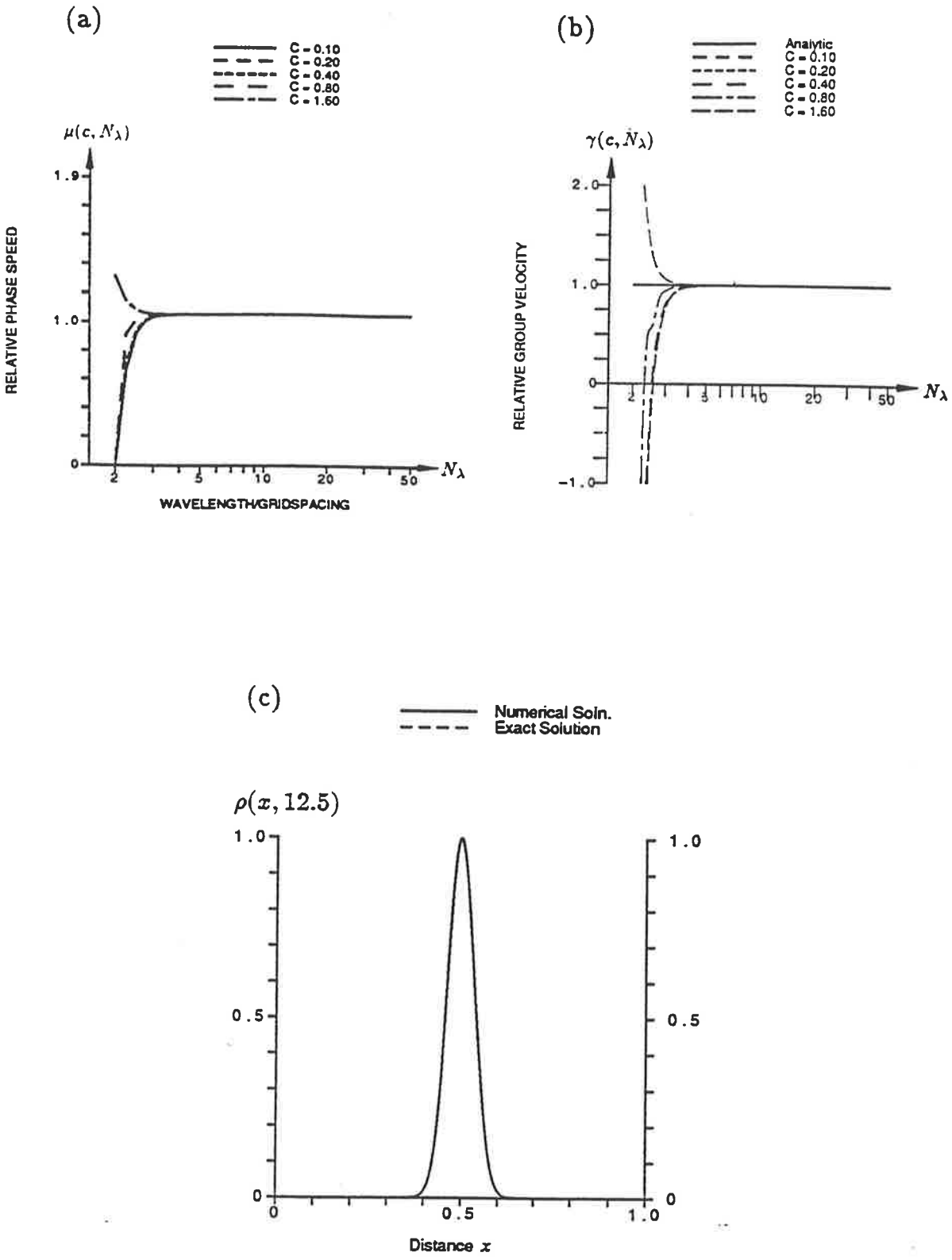

Figure 3.12: As in Fig. 3.1 but for the scheme NS5, Eq. (3.60). This scheme is unstable for $c > 1.2$.

Table 3.1 : Error measures for standard difference schemes for the test problem with a Gaussian pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Minimum $\{\rho_j^n\}$	Sum Neg. Values	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
First order Upwinding	2.2×10^{-1}	8.6×10^{-1}	3.7×10^{-1}	0	-7.5×10^{-3}	42.30	1.2×10^{-3}	-0.1	0.14
Lax-Wendroff scheme	2.0×10^{-1}	5.7×10^{-1}	-3.0×10^{-1}	-3.7×10^{-1}	-5.5×10^{-1}	29.16	7.9×10^{-2}	-6.6	0.89
3^{rd} order Upwinding	6.7×10^{-2}	2.6×10^{-1}	-3.7×10^{-2}	-3.9×10^{-1}	2.8×10^{-6}	1.62	1.4×10^{-5}	0.0	0.74
Rusanov's 4^{th} order scheme	3.9×10^{-2}	1.4×10^{-1}	-8.7×10^{-2}	-4.5×10^{-1}	2.7×10^{-3}	2.01	6.1×10^{-4}	-0.8	0.93
5^{th} order Upwinding	1.1×10^{-2}	4.4×10^{-2}	-3.4×10^{-3}	-1.6×10^{-2}	-2.2×10^{-8}	5.0×10^{-2}	4.0×10^{-6}	0.0	0.96
Crank Nicolson	2.5×10^{-1}	6.9×10^{-1}	-4.1×10^{-1}	-5.9×10^{-0}	-3.8×10^{-1}	50.20	-5.8×10^{-2}	-7.9	0.63
Linear F.E. Crank Nicolson	8.8×10^{-2}	3.3×10^{-1}	-2.0×10^{-1}	-1.1×10^{-0}	5.9×10^{-3}	4.15	2.3×10^{-3}	-2.1	0.89
4^{th} order CTCS	4.0×10^{-2}	1.8×10^{-1}	-4.2×10^{-2}	-1.9×10^{-1}	-4.7×10^{-3}	0.18	7.4×10^{-5}	-0.4	0.84
Khaliq and Twizell (2,0)	2.2×10^{-1}	5.3×10^{-1}	-3.0×10^{-1}	-5.1×10^{-0}	-8.7×10^{-1}	39.97	-1.2×10^{-1}	8.7	0.53
" (2,1)	2.5×10^{-1}	7.0×10^{-1}	-3.9×10^{-1}	-5.9×10^{-0}	4.6×10^{-1}	46.34	5.5×10^{-2}	-5.7	0.63
" (2,1) Extrapolated	2.5×10^{-1}	7.0×10^{-1}	-3.9×10^{-1}	-5.9×10^{-0}	4.6×10^{-1}	46.34	5.5×10^{-2}	-5.7	0.63
" (2,2)	2.5×10^{-1}	7.0×10^{-1}	-3.9×10^{-1}	-5.9×10^{-0}	4.6×10^{-1}	46.34	5.5×10^{-2}	-5.7	0.63
" (2,2) Extrapolated	2.5×10^{-1}	7.0×10^{-1}	-3.9×10^{-1}	-5.9×10^{-0}	4.6×10^{-1}	46.34	5.5×10^{-2}	-5.7	0.63
NS3	6.4×10^{-2}	2.1×10^{-1}	-1.6×10^{-1}	-1.4×10^{-0}	-4.3×10^{-2}	11.35	1.5×10^{-2}	-1.1	0.89
NS4	6.7×10^{-4}	3.0×10^{-3}	-3.6×10^{-5}	-9.3×10^{-5}	-3.0×10^{-7}	1.3×10^{-3}	8.8×10^{-6}	0.0	1.00
NS5	1.6×10^{-5}	6.8×10^{-5}	-4.1×10^{-9}	-1.3×10^{-8}	3.0×10^{-11}	1.3×10^{-3}	2.4×10^{-7}	0.0	1.00

discontinuities in the first derivative, and therefore provides a more stringent test. Since the first derivative is discontinuous, it is also more difficult to model than the cosine hill. To illustrate the performance of these schemes in advecting profiles which contain abrupt changes, the semi-ellipse test is used as a numerical example, namely, modelling the advection of the function

$$\hat{\rho}(x,0) = \begin{cases} 0 & 0 \leq x < 0.4 \\ 10\sqrt{(0.1)^2 - (x-0.5)^2} & 0.4 \leq x \leq 0.6 \\ 0 & 0.6 < x \end{cases} \tag{3.61}$$

with the same boundary conditions and parameters as for the Gauss pulse test. The graphs of these tests are shown in Figs. 3.13, 3.14, 3.15, 3.16, with the error measures presented in Table 3.2. The error measures for schemes which give similar results to those already included, have been omitted. The figures are arranged by the order of the modified equivalent equation for ease of comparison.

The illustrations of the semi-ellipse test show that the explicit schemes perform in much the same manner as for the Gauss pulse test, however the implicit schemes deteriorate in comparison with the earlier tests. This decrease in the accuracy of the implicit schemes is due to the absence of damping of the short wavelength components. Due to abrupt changes in the initial condition at $x = 0.4$ and $x = 0.6$, these components now have significantly greater amplitude than for the smooth Gaussian pulse. None of the schemes propagate components of wavelength $N_\lambda = 2$ at the correct phase speed. The implicit schemes all have $\mu(c, 2,) \equiv 0$ for all c and the explicit schemes (apart from Holly and Preissmann's scheme) have very low values of $\mu(c, 2)$. The energy associated with these waves is propagated at the group velocity and the relative group velocities, γ, of these schemes shows that this energy is transmitted backwards relative to the analytic solution. Further, the higher order implicit schemes transmit this energy faster in the reverse direction. This is endemic to using high order schemes since the derivative of μ with respect to N_λ near the Nyquist limit, $(N_\lambda = 2)$, increases with increasing order.

This problem with the short wavelengths is not so apparent with the explicit

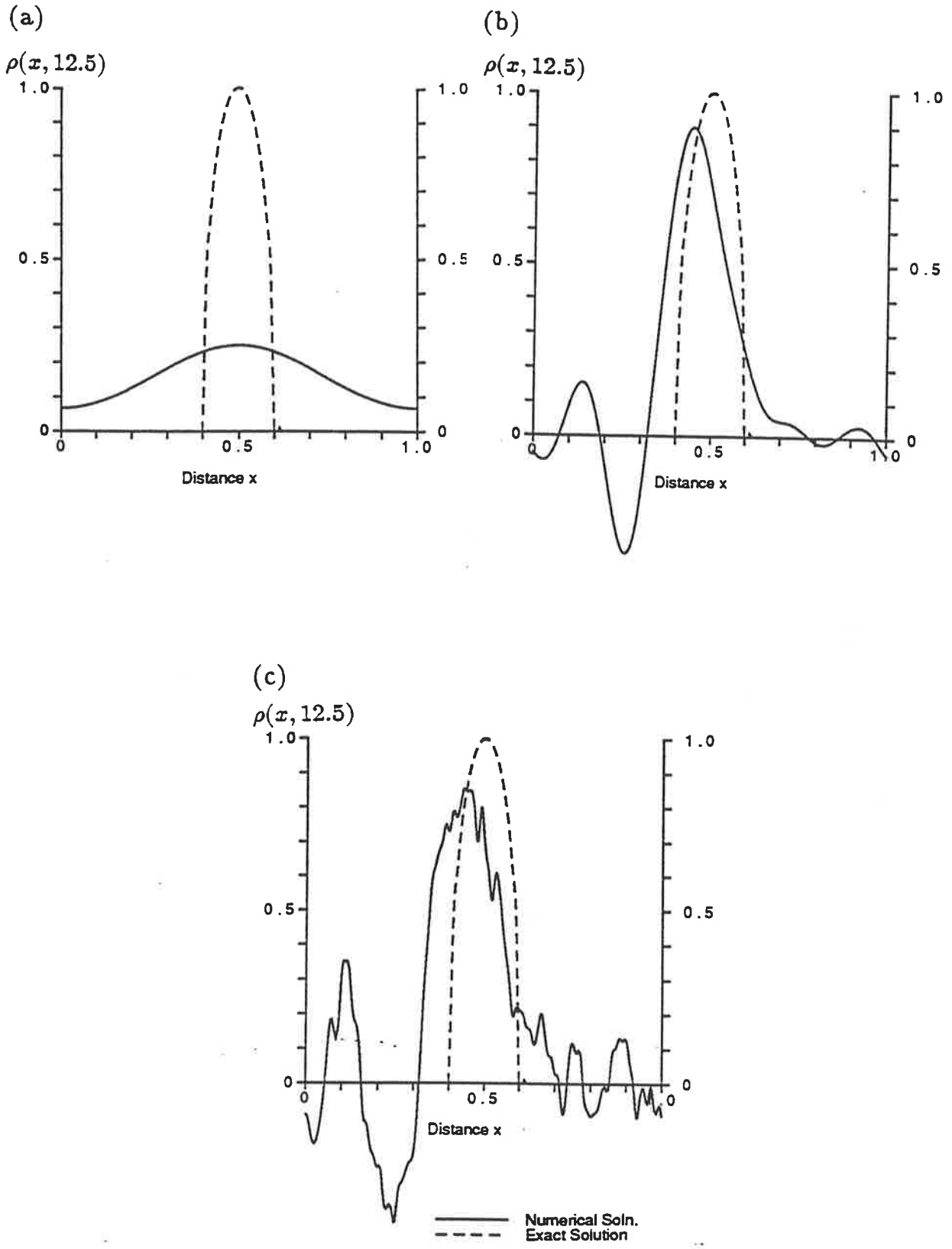

Figure 3.13: Illustration of numerical results obtained using (a) First order upwinding, (b) Lax-Wendroff and (c) Crank Nicolson schemes. The initial condition is a semi-ellipse and the boundary conditions are cyclic.

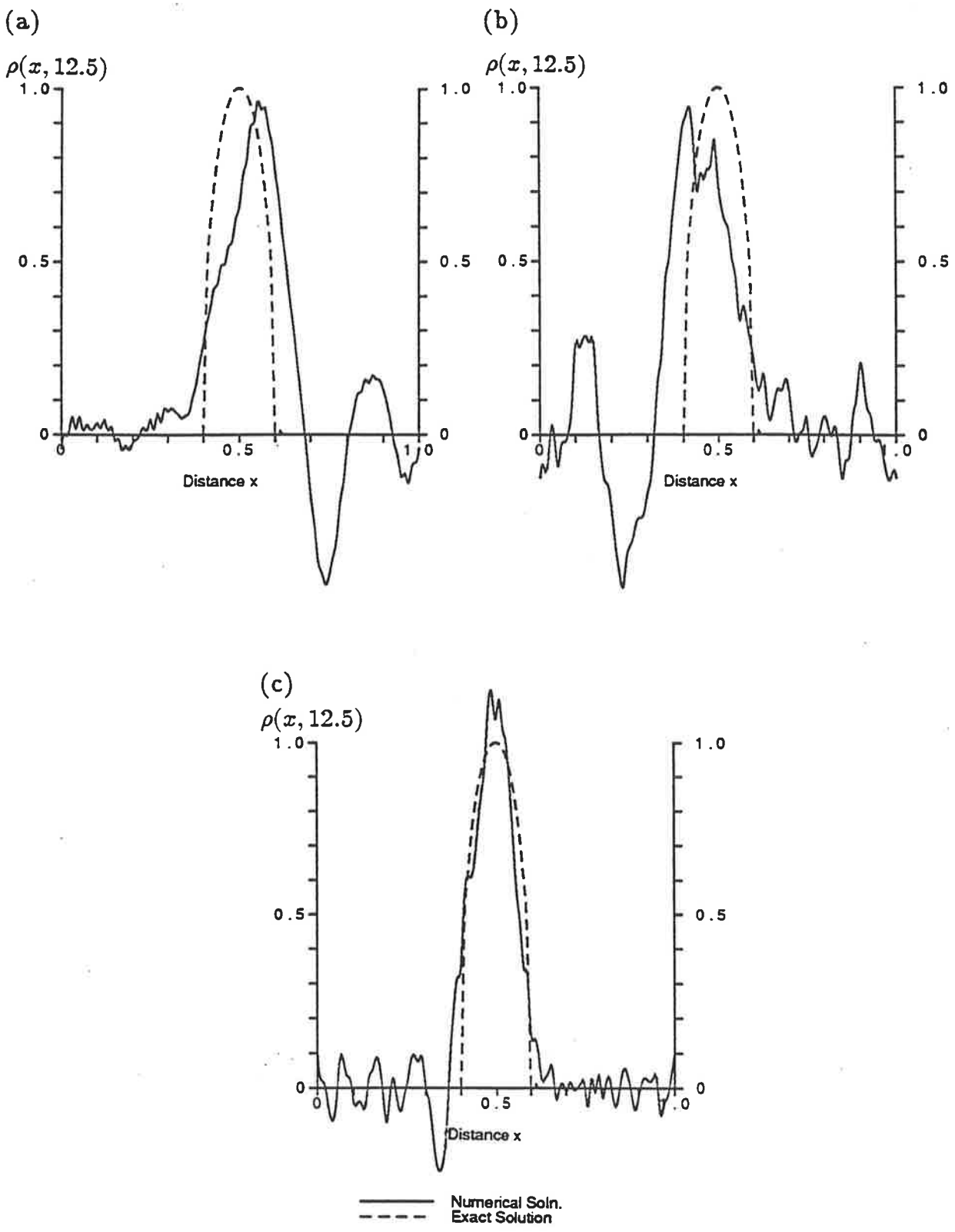

Figure 3.14: As in Fig. 3.13 but using (a) Khaliq and Twizell's (2,0) approximation, (b) Khaliq and Twizell's (2,2) approximation with extrapolation and (c) linear finite element Crank Nicolson schemes.

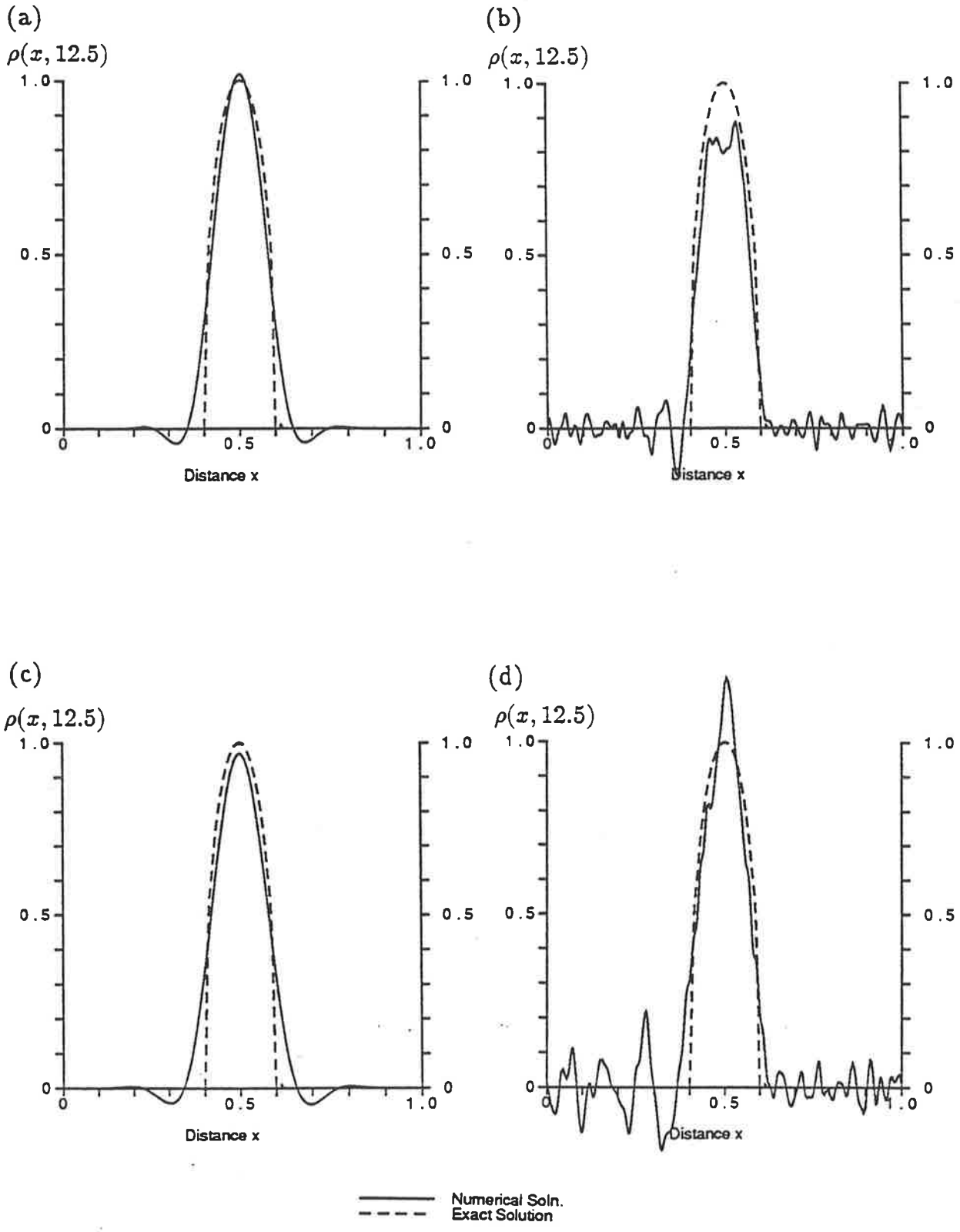

Figure 3.15: As in Fig. 3.13 but using (a) 3^{rd} order Upwinding, (b) 4^{th} order CTCS, (c) Rusanov's 4^{th} order and (d) NS3 schemes.

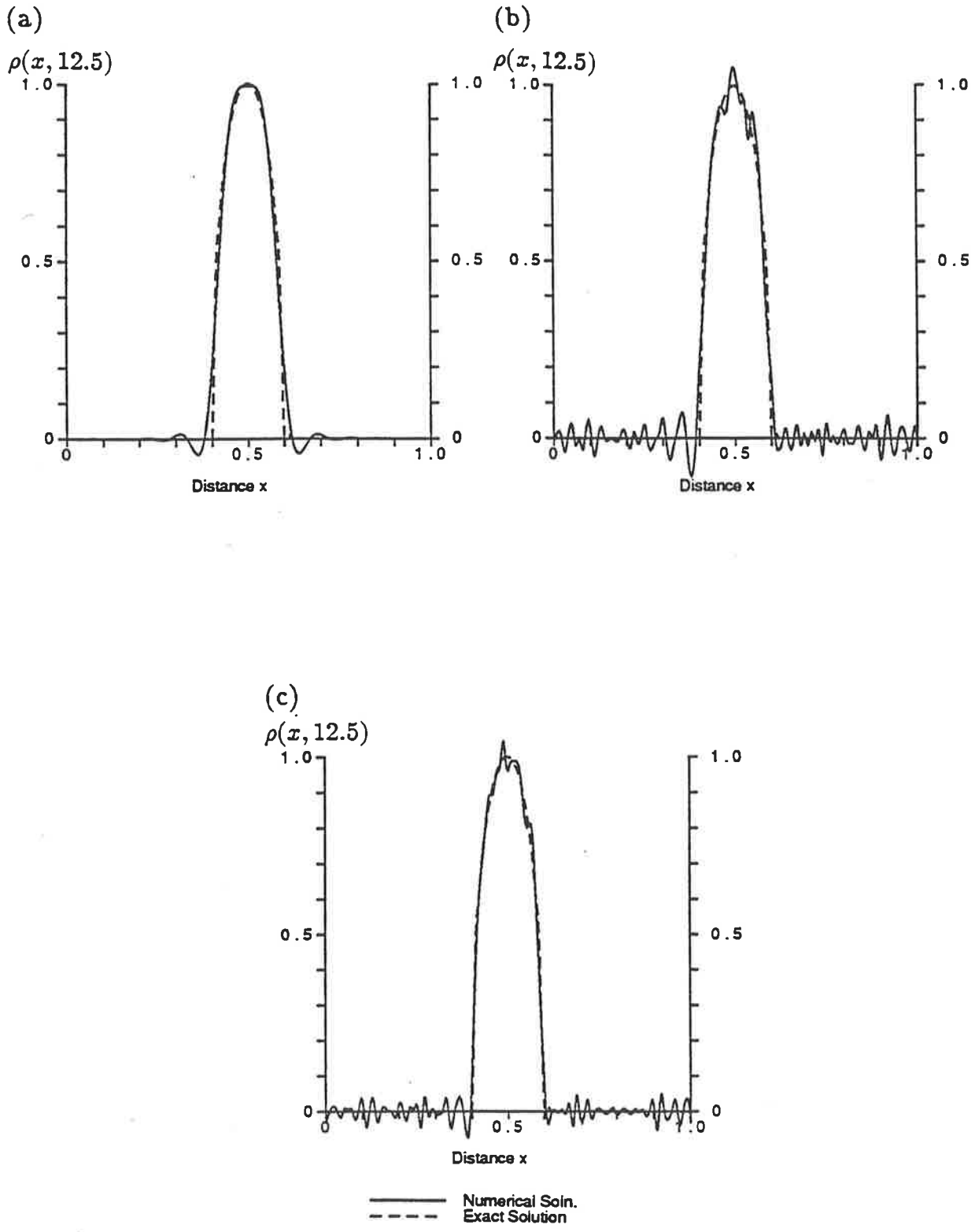

Figure 3.16: As in Fig. 3.16 but using (a) 5^{th} order Upwinding, (b) NS4 and (c) NS5 schemes.

Table 3.2 : Error measures for standard difference schemes for the test problem with a semi-elliptical pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Minimum $\{\rho_i^n\}$	Sum Neg. Values	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
First order Upwinding	2.9×10^{-1}	7.5×10^{-1}	6.7×10^{-2}	0	-1.3×10^{-2}	21.72	2.0×10^{-3}	-0.1	0.25
Lax-Wendroff scheme	2.2×10^{-1}	7.6×10^{-1}	-3.4×10^{-1}	-3.52	1.2×10^{-1}	8.45	3.4×10^{-2}	-5.6	0.90
3^{rd} order Upwinding	6.6×10^{-2}	3.0×10^{-1}	-4.3×10^{-2}	-0.48	8.5×10^{-6}	0.82	7.3×10^{-6}	0.0	1.02
Rusanov's 4^{th} order scheme	7.7×10^{-2}	3.3×10^{-1}	-4.6×10^{-2}	-0.61	1.1×10^{-4}	1.25	-3.7×10^{-5}	0.0	0.97
5^{th} order Upwinding	4.0×10^{-2}	2.1×10^{-1}	-4.5×10^{-2}	-0.28	2.1×10^{-6}	0.31	-6.2×10^{-4}	0.5	0.99
Crank Nicolson	2.6×10^{-1}	7.5×10^{-1}	-4.0×10^{-1}	-5.73	-1.4×10^{-1}	14.72	1.8×10^{-3}	-5.6	0.86
Linear F.E. Crank Nicolson	1.0×10^{-1}	3.4×10^{-1}	-2.4×10^{-1}	-2.07	4.1×10^{-3}	7.93	4.0×10^{-3}	-1.2	1.15
4^{th} order CTCS	7.8×10^{-2}	2.1×10^{-1}	-1.4×10^{-1}	-1.08	2.9×10^{-2}	5.75	7.3×10^{-4}	3.4	0.88
Khaliq and Twizell (2,0)	2.2×10^{-1}	7.5×10^{-1}	-4.3×10^{-1}	-4.05	-3.5×10^{-1}	9.77	-6.2×10^{-2}	5.2	0.96
" (2,2) Extrapolated	2.6×10^{-1}	8.9×10^{-1}	-4.4×10^{-1}	-5.16	-2.3×10^{-1}	14.19	-9.6×10^{-3}	-8.3	0.95
NS3	9.1×10^{-2}	3.1×10^{-1}	-1.7×10^{-1}	-2.11	1.1×10^{-1}	7.99	1.3×10^{-2}	0.6	1.18
NS4	4.6×10^{-2}	2.2×10^{-1}	-9.9×10^{-2}	-1.00	-4.1×10^{-3}	4.25	-4.3×10^{-4}	-0.1	1.05
NS5	3.1×10^{-2}	1.2×10^{-1}	-6.4×10^{-2}	-0.77	1.6×10^{-2}	3.53	2.8×10^{-3}	-1.0	1.04

schemes since they are also damped out very quickly, but as there is no such damping by the implicit schemes, these components produce a very noisy numerical solution. This noise also means that the values quoted in Table 3.2 for the peak shift and peak height are questionable for the implicit schemes as it is difficult to distinguish a genuine peak amongst the high frequency noise. This problem also affects the calculation of the moments of the solution. The values given are therefore only a guide to the performance of the implicit schemes, the high frequency noise in the numerical solutions making the error bounds for these calculations much larger than before.

That the upwind schemes are more successful, relative to the implicit schemes, at retaining the moments of the distribution is consistent with Martin (1975), where it was shown that these schemes are the ideal explicit finite difference schemes for preserving the moments of a distribution up to a given order (dependent on the order of the difference scheme).

Despite these problems, the conclusion based on the results of the Gauss pulse test, namely that high order schemes are more accurate, is still valid. The error measures quoted in Table 3.2 still exhibit general improvement in the overall accuracy with increasing order.

It is worth noting some of the features of the fifth or higher order schemes, shown in Fig. 3.16(a-d). The results from using fifth order upwinding, Fig. 3.16(a), show some flattening of the peak. In Chapter Four, this flattening will be seen to be important. It will be shown that techniques for obtaining smooth results also produce a flattening of extrema. This flattening of the peak is exacerbated if the high order scheme that forms the basis for the smoothing technique also displays this behaviour. The oscillations in the numerical solutions of NS4 and NS5, shown in Fig. 3.16(b,c) are also very high frequency, of only three or four grid-spacings in wavelength, and as such, are the type of oscillations that are most efficiently removed. The smoothing techniques discussed in the next chapter are based on

some form of low pass filter (generally non-linear), and so the higher the frequency of the noise, the more efficiently that noise is removed. If slightly lower frequencies ($N_\lambda = 7$ to 10) need to be moderated, then these filters begin to impact on the significant components of the solution.

While it has been established that on a given grid, for a given Courant number, the high order schemes provide superior results to the low order schemes, the effect of using the low order schemes on a finer grid has not been discussed. It has been acknowledged that under equivalent conditions, the high order schemes require considerably more CPU time, due to the more complicated coefficients and possible matrix inversion.

To demonstrate that high order schemes do obtain accurate results faster than low order schemes, Table 3.3 presents the RMS errors associated with all the schemes discussed so far on a variety of grids, ($J = 1/\Delta x = 100$, 200, 400 and 800) and repeating the previous test cases. The initial conditions, boundary conditions, Courant number and the velocity w have been left unchanged. Included in Table 3.3 are the CPU times for each scheme for the cases where $J = 100$ and 800. The CPU times required by each scheme quadruple for each halving of the grid-spacing since halving the number of grid-points requires the time-step to be halved as well. In the case where the initial condition is a Gaussian pulse, the efficiency of the high order schemes is immediately apparent. Not only are the high order schemes more accurate for $J = 100$ but the results improve at a faster rate with increasing resolution. The extreme cases are first order upwinding where the errors decrease by a factor of about two-thirds when the number of grid-points is changed from 100 to 800, whereas for NS5 the corresponding errors decrease to less than one-millionth of the size for the same change in resolution. This improvement in the rate of decrease of error with increased resolution is even seen when comparing first order upwinding with the Lax-Wendroff schemes where the errors for $J = 800$ are less than one-tenth of those for $J = 100$. Of the explicit schemes, fifth order upwinding is clearly the

most accurate, yet even this scheme requires approximately quadruple the CPU time to attain a similar accuracy as NS5.

When a semi-ellipse is used as an initial condition, the high order implicit schemes do not seem to be so great an improvement. The percentage decrease in RMS error for NS5 is still greater than all the other schemes but the improvement is less marked. Furthermore for this test, fifth order upwinding can now produce errors comparable to NS5 in less time. For the same grid-spacing, fifth order upwinding requires about one-quarter of the time of NS5, but the RMS error for NS5 with 100 grid-points is close to the midpoint of the RMS errors for fifth order upwinding with 100 and 200 grid-points, indicating that, in this case, fifth order upwinding can attain the same accuracy in roughly half the time as NS5.

A further point to be considered, arises from the discussion in Chapter Four; namely that smoothing techniques inherently produce results that do not contain discontinuous derivatives, effectively damping out high frequency noise. This is a direct consequence of all smoothing techniques introducing numerical diffusion in the presence of abrupt changes. Converting solutions such as the semi-elliptical pulse to a smoother waveform, favours the high order implicit schemes since it has been demonstrated that these schemes are excellent at advecting infinitely differentiable solutions.

3.5 Boundary Conditions for High Order Schemes

A further complication with using high order schemes is the question of computation of points adjacent to the boundary. Allied with this is the problem of how to compute the value at the downstream boundary. This is not difficult to overcome, however, especially if the flux forms of the difference equations are used. Firstly consider the upstream boundary. For schemes using stencils involving values at x_{j-2}, there is no equation for calculating the point adjacent to the boundary, ρ_1^{n+1}. The three basic choices are:

Table 3.3:Comparisons of improvement in R.M.S. error versus increase in CPU time due to increasing resolution. The test problems use either a Gaussian pulse or a semi-circle as the initial condition along with cyclic boundary conditions. The CPU Times are for the calculations with $J = 100$ and 800.

	CPU	Gaussian Pulse				Semi-elliptical Pulse				CPU
Scheme	$J = 100$		$J = 200$	$J = 400$	$J = 800$	$J = 100$	$J = 200$	$J = 400$	$J = 800$	
First Order Upwinding	1.8	2.2×10^{-1}	2.0×10^{-1}	1.8×10^{-1}	1.5×10^{-1}	2.9×10^{-1}	2.6×10^{-1}	2.2×10^{-1}	1.7×10^{-1}	110
Lax-Wendroff	4.4	2.0×10^{-1}	1.4×10^{-1}	6.0×10^{-2}	1.7×10^{-2}	2.2×10^{-1}	1.0×10^{-1}	6.8×10^{-2}	4.4×10^{-2}	280
2rd order Upwind biassed	4.4	1.5×10^{-1}	1.5×10^{-1}	6.6×10^{-2}	1.9×10^{-2}	1.0×10^{-1}	1.0×10^{-1}	7.1×10^{-2}	4.6×10^{-2}	280
3rd order Upwind biassed	7.8	6.7×10^{-2}	2.0×10^{-2}	3.4×10^{-3}	4.4×10^{-4}	6.6×10^{-2}	4.0×10^{-2}	2.3×10^{-2}	1.4×10^{-2}	500
Holly & Preissmann	10.9	1.8×10^{-2}	3.0×10^{-3}	4.0×10^{-4}	5.1×10^{-5}	4.2×10^{-2}	2.5×10^{-2}	1.4×10^{-2}	8.3×10^{-3}	700
4th order Upwind biassed	11.7	3.9×10^{-2}	4.3×10^{-3}	2.8×10^{-4}	1.8×10^{-5}	6.1×10^{-2}	3.3×10^{-2}	1.9×10^{-2}	1.1×10^{-2}	750
Rusanov's 4th order scheme	11.7	3.9×10^{-2}	4.0×10^{-3}	2.6×10^{-4}	1.6×10^{-5}	6.2×10^{-2}	3.4×10^{-2}	2.0×10^{-2}	1.1×10^{-2}	750
5th order Upwind biassed	14.8	1.1×10^{-2}	5.6×10^{-3}	1.8×10^{-5}	5.8×10^{-7}	4.0×10^{-2}	2.2×10^{-2}	1.2×10^{-2}	6.7×10^{-3}	950
Crank Nicolson	15.9	2.5×10^{-1}	1.6×10^{-1}	7.4×10^{-2}	2.2×10^{-2}	2.6×10^{-1}	1.2×10^{-1}	8.5×10^{-2}	5.3×10^{-2}	1020
Crank Nicolson (F.E.)	20.2	8.8×10^{-2}	2.6×10^{-2}	6.5×10^{-3}	1.6×10^{-3}	1.0×10^{-1}	5.7×10^{-2}	3.7×10^{-2}	2.3×10^{-2}	1290
Fourth order CTCS	22.7	1.1×10^{-2}	7.0×10^{-4}	4.3×10^{-5}	2.6×10^{-6}	5.9×10^{-2}	3.1×10^{-2}	1.8×10^{-2}	1.0×10^{-2}	1450
Khaliq & Twizell (2,0)	45.5	2.2×10^{-1}	1.5×10^{-1}	6.1×10^{-2}	1.6×10^{-2}	2.2×10^{-1}	1.0×10^{-1}	7.1×10^{-2}	4.6×10^{-2}	2850
" (2,1)	51.1	2.5×10^{-1}	1.6×10^{-1}	7.0×10^{-2}	2.0×10^{-2}	2.6×10^{-1}	1.2×10^{-1}	8.0×10^{-2}	5.2×10^{-2}	3200
" (2,1) + Extrap.	78.2	2.5×10^{-1}	1.6×10^{-1}	7.0×10^{-2}	2.0×10^{-2}	2.6×10^{-1}	1.2×10^{-1}	7.9×10^{-2}	5.2×10^{-2}	5040
" (2,2)	52.0	2.5×10^{-1}	1.6×10^{-1}	7.0×10^{-2}	2.0×10^{-2}	2.6×10^{-1}	1.2×10^{-1}	8.0×10^{-2}	5.2×10^{-2}	3330
" (2,2) + Extrap.	79.7	2.5×10^{-1}	1.6×10^{-1}	7.0×10^{-2}	1.0×10^{-2}	2.6×10^{-1}	1.2×10^{-1}	8.0×10^{-2}	5.2×10^{-2}	5100
NS3	46.6	6.4×10^{-2}	7.4×10^{-3}	4.8×10^{-4}	3.0×10^{-5}	9.1×10^{-2}	4.8×10^{-2}	2.8×10^{-2}	1.5×10^{-2}	2980
NS4	51.9	6.7×10^{-4}	1.0×10^{-5}	1.6×10^{-7}	2.5×10^{-9}	4.6×10^{-2}	2.0×10^{-2}	1.2×10^{-2}	6.5×10^{-3}	3320
NS5	55.2	1.6×10^{-5}	5.4×10^{-8}	2.1×10^{-9}	9.8×10^{-12}	3.1×10^{-2}	1.7×10^{-2}	8.6×10^{-3}	4.7×10^{-3}	3530

1. interpolation of the value ρ_1^{n+1} from surrounding values,

2. extrapolation of the solution to obtain an estimate of the fictitious value ρ_{-1}^{n} which can then be used directly in the equation for ρ_1^{n+1},

3. the use of another, more compact scheme to provide a value for ρ_1^{n+1}.

It was also mentioned in the introduction to subsection 2.4.2, that Khaliq and Twizell's derivation was originally developed for Dirichlet boundary conditions in particular and as such, provided a scheme to estimate ρ_1^{n+1}. The resulting boundary scheme is actually equivalent to setting ρ_{-1}^{n+1} equal to ρ_1^{n+1}. This can be very easily verified from the equations in the original paper. Putting $\rho_{-1}^{n+1} = \rho_1^{n+1}$ is equivalent to setting the first derivative of ρ to be zero (to second order in Δx).

To test these different approaches a simple numerical experiment was designed. Fifth order upwinding was used to advect a pulse of the form

$$\hat{\rho}(x,0) = 10\sqrt{\max(0,(0.1)^2 - (x+0.5)^2)} \quad , \tag{3.62}$$

namely, a semi-ellipse centred on $x = -0.5$, through the boundary. The same parameters as before were used ($\Delta x = 1/100$, $\Delta t = 1/200$, $c = 0.4$), but the pulse was advected for two hundred time-steps. At the conclusion of the calculations, the pulse should be centred at $x = 0.3$ and extending from $x = 0.2$ to $x = 0.4$. The combination of fifth order upwinding and a semi-ellipse was chosen since the scheme advects this waveform fairly cleanly, but if there is any problem with an algorithm, the semi-ellipse produces a stronger signal than would be obtained by using a smooth waveform such as a Gaussian pulse. Fifth order upwinding was also selected because it is one of the more simply programmed high order schemes. When using implicit schemes, some of the above techniques may cause the matrix inversion to become unstable, introducing an additional factor that may obscure any results. The problem with using fifth order upwinding is that it requires values at x_{j-3} and so this scheme also requires an independent estimation of ρ_2^{n+1}. This

estimation was done by using third order upwinding. This scheme also advects a semi-ellipse reasonably cleanly.

The importance of using schemes that do not produce noisy results, is that often errors in the specification of boundary conditions are manifested as oscillations entering the domain. These oscillations could be lost amongst the noise of the base scheme if one of the implicit schemes is used rather than fifth order upwinding. Care must also be taken during the calculation of the fluxes when using different schemes to calculate adjacent values. This is overcome by using the average flux of the two schemes. For example, between the second and third grid-point the flux may be given by $(f_{5/2}^{UW3} + f_{5/2}^{UW5})/2$ where $f_{5/2}^{UW3}$ and $f_{5/2}^{UW5}$ represent the fluxes between the grid-points of third and fifth order upwinding, respectively. This average flux is then used to calculate ρ_2^{n+1} via

$$\rho_2^{n+1} = \rho_2^n - f_{5/2} + f_{3/2} \quad . \tag{3.63}$$

A similar process can be used to match the fluxes between the first and second grid-points if necessary.

As a benchmark, the test was run with ρ_1^n being calculated directly from the analytic solution. While this type of overspecification of values can induce oscillations by itself, it will be assumed that these are not too serious. When compared with the analytic solution and the other experiments, this assumption is seen to be valid. One of the problems with such a test, is that conventional error measures that represent the overall performance of a scheme are not appropriate for measuring the performance of, what are essentially, "local" schemes, since the standard error measures taken over the entire domain are dominated by the errors of the base scheme; in this case, fifth order upwinding.

Despite the small overall contribution to a gross error measure by these different approximations in the vicinity of the boundary, the different boundary schemes provide the major source of error, even in such a simple test as this. The results are summarized in Table 3.4 which illustrates the RMS error taken over the eighteen

Table 3.4 : Error measures of different techniques for estimating $\hat{\rho}(\Delta x,\ n\Delta t)$. The test involved advecting a semi-ellipse through the boundary. The parameters used in the experiment are the same as the semi-ellipse test with cyclic boundary conditions but only 200 time-steps were calculated. The errors have been calculated using the eighteen points adjacent to the boundary. The *First Sig. Value* is the index of the first grid-point with absolute value greater than .0001.

Scheme	RMS. Error	Maximum $\lvert Error \rvert$	First Sig. Value
Using analytic solution	2.4×10^{-3}	7.3×10^{-3}	11
Linear Interpolation	2.6×10^{-2}	8.1×10^{-2}	4
Quadratic Interpolation	3.4×10^{-2}	1.1×10^{-1}	4
Cubic Interpolation	6.8×10^{-2}	1.8×10^{-1}	1
Linear Extrapolation	4.6×10^{-3}	1.6×10^{-2}	11
Quadratic Extrapolation	7.8×10^{-3}	2.7×10^{-2}	11
Cubic Extrapolation	6.1×10^{-3}	2.1×10^{-2}	9
As for Khaliq and Twizell	6.5×10^{-3}	2.5×10^{-2}	11
First order Upwinding	2.0×10^{-3}	6.9×10^{-3}	11
Lax-Wendroff	4.9×10^{-3}	1.6×10^{-2}	11
4^{th} order C.T.C.S.	5.0×10^{-3}	1.6×10^{-2}	11

grid-points adjacent to the boundary as well as listing the first grid-point with absolute value greater than 1×10^{-4}. Again these errors only involve values from the final time-step.

While care must be taken when drawing conclusions from such a simple test as this, some points are quite evident. Firstly, interpolation is not an appropriate method near boundaries, as shown by the substantially larger errors than other schemes. The index of the "first significant value" (the first value of ρ_j^{200} such that $\left|\rho_j^{200}\right| > 1 \times 10^{-4},\ j = 1, \cdots, 18$) is also much closer to the boundary in comparison to other techniques. This indicates that the trailing edge of the pulse is leaving a much more noticeable trail, which can be explained very simply by considering the estimated value of ρ_1^{n+1} when the trailing edge of the pulse is halfway between the first and second grid-points. The values of the analytic solution at the points in the vicinity of the boundary are $\hat{\rho}_0 = 0$, $\hat{\rho}_2 = 0.312$, $\hat{\rho}_3 = 0.527$ and $\hat{\rho}_4 = 0.661$. These values may be used as a guide to the values of the numerical solution at the same edge. When a value for ρ_1 is interpolated between these values, it is going to be

quite high in comparison to the analytic solution, ($\hat{\rho_1} = 0$). The interpolation will take many time-steps before the value of ρ_1^{n+1} goes close to zero due to the values of ρ_2^{n+1} being kept high (because ρ_1^n, ρ_2^n, ρ_3^n,... are too large), stopping the interpolated value of ρ_1^{n+1} approaching zero as quickly as it should. A further problem with interpolation is that it is not based on fluxes and so it is not conservative, as can be seen in the example above, where ρ_1 was incorrectly increased, even if exact values are given to the interpolation.

Of the remaining techniques there is little variation, the exception being when first order upwinding is used to generate the fluxes away from the boundary. This is perhaps not surprising since the flux away from the boundary as calculated by this scheme is just $c_{1/2}\hat{\rho_0^n}$, that is, it only involves values from the boundary. All the other schemes use information from the numerical solution and so they do not respond as quickly to changes in the boundary conditions. This rapid response more than compensates for the low order of this approximation to the outward flux. The values of ρ_1^{n+1} are also improved since the flux between the first and second grid-points is given by an average of first and third order upwinding which is an improvement on using only first order fluxes to estimate ρ_1^{n+1}. This scheme has the further advantage that it is very easy to implement and so it is used for the remainder of this thesis whenever estimates of ρ_1^{n+1} are required in numerical tests involving Dirichlet boundary conditions.

The remaining problem is the downstream, outflow boundary. Most of the schemes require a value for ρ_{j+1}^n. Strictly, the advection equation, Eq. (2.2), has no downstream boundary condition and the imposition of an artificial condition overspecifies the problem. This can lead to incorrect outward flux from the domain. In practice, however, diffusion is generally present, if only weakly, and the presence of a second order derivative requires the specification of another boundary condition and so in practice, the downstream boundary value can generally be obtained from this boundary condition.

As the advection equation is only being considered in the sense that it is the limit as diffusion tends to zero of the full transport equation, the specification of a downstream boundary condition does not present a problem. There still remains, however, the problem of calculating the point adjacent to the boundary for the high order schemes which require values at x_{j+2} to estimate ρ_j^{n+1}. This is resolved in a similar fashion to the estimation of ρ_1^{n+1} adjacent to the upstream boundary, with a low order scheme providing the flux out of the grid cell adjacent to the boundary and the high order scheme providing a flux into this grid cell.

For these reasons, when high order schemes which are tested on problems involving fixed boundary conditions are examined later in this thesis, the points adjacent to the boundary will be given by a lower order scheme.

3.6 Conclusion

Some new high order and very accurate schemes were developed in this chapter using the modified equation approach of Noye and Hayman (1986). Although these schemes generally require more CPU time per grid-point per time-step than other schemes discussed so far, they have been shown to produce results of such accuracy (particularly when the initial condition is infinitely differentiable) that other schemes require considerably more time to yield comparable results, confirming the conclusions of Chapter Two that high order schemes are more accurate in practice. That this is true in the limit as $\Delta x \rightarrow \infty$ is obvious, but other methods, such as numerical experiments and comparison of wave propagation parameters are required to demonstrate this for realistic Δx. These high order implicit schemes have not only been shown to be highly accurate, but also provide a greater improvement in accuracy (for a given increase in resolution) than low order schemes. That is, if efficiency is measured by the computational time required to obtain a specified accuracy, then these implicit schemes are very efficient.

It was also shown that to obtain the full advantages of high order schemes *both* the

spatial and the temporal differencing must be of high order. This is clear from the comparisons of Khaliq and Twizell's schemes with the other very high order schemes. The improvement in accuracy by increasing the order of the approximation to $\partial\rho/\partial t$ was negligible due to the presence of second order errors in the approximation to the spatial derivative. It is also clear that any scheme that contained approximations that were only first or second order accurate was clearly deficient in modelling the simplest form of advection, as was demonstrated by the results obtained from first order upwinding, Lax-Wendroff, Standard Crank Nicolson and the Khaliq and Twizell schemes.

The effect on initial conditions that contain discontinuous derivatives was also examined. The high order implicit schemes suffered from the production of high frequency noise which is inherent in all diagonally differenced (centred-time, centred-space) schemes. This noise is due to the lack of damping of the high frequency Fourier components, and so relative to the explicit schemes discussed in Chapter Two, performance of the implicit schemes was degraded.

The errors of the high order schemes NS4 and NS5 are almost entirely due to noise with a wavelength of only three of four grid-spacings and as will be seen in Chapter Four, this type of noise is most amenable to removal by smoothing techniques.

Chapter 4

Review of Smooth Numerical Schemes

It was mentioned in the Chapter One that in many problems it is crucial to obtain smooth results when modelling advective processes. This chapter presents a review of some techniques that are used to attain this goal. The problem of obtaining smooth results is closely linked with removing non-physical negative values, since these negative values are due to the same spurious oscillations that cause the roughness in the numerical solution. Such oscillations cannot just be ignored, since removing them from the solution by setting negative values to zero adds mass to the system. The solution must be smoothed over the entire domain since the source of the oscillations may be some distance from where they finally appear. For the moment there will be no strict definition of the smoothness of a solution except that there should be no "lumps" or "bumps" in the solution. Later in this chapter, various definitions of smoothness and their effect on numerical results will be discussed.

4.1 Linear Filtering

Unwanted negative values in what should be a positive solution, are caused by spurious oscillations. These in turn are due to the poor numerical phase speed of the short Fourier components. One procedure is to use a numerical filter to selectively damp out these components. These filters can have a sharp cutoff of short wavelengths as shown in Fig. 4.1(a) and so delete the high frequencies and thereby retain the

middle and low frequencies. This does not remove all the negative values from the numerical results, however, as even the mid-frequencies of the numerical solution still travel at an incorrect phase speed. If these frequencies are also removed then much of the detail of the distribution will also be lost. A further problem is that no linear filters have perfect response, that is, linear low pass filters have responses that are neither identically zero for high frequencies nor identically one for intermediate or lower frequencies. So linear filtering will not necessarily remove all the negative values, although it may make them acceptably small in some applications. Examples of using a filter in combination with NS5, Eq. (3.60) are provided in Fig. 4.1. The filter used was discussed by Shapiro (1970), namely

$$\begin{aligned} \rho_j^{n+1} &= \frac{186}{256}\rho_j^H + \frac{56}{256}\left(\rho_{j-1}^H + \rho_{j+1}^H\right) - \frac{28}{256}\left(\rho_{j-2}^H + \rho_{j+2}^H\right) \\ &\quad + \frac{8}{256}\left(\rho_{j-3}^H + \rho_{j+3}^H\right) - \frac{1}{256}\left(\rho_{j-4}^H + \rho_{j+4}^H\right) \end{aligned} \tag{4.1}$$

where ρ_j^H is the time-advanced, oscillatory solution.

The filter given by Eq. (4.1) has quite a sharp cutoff in comparison with other filters. This is seen in Fig. 4.1(c) where even though the filter has been applied 250 times during the calculations, the disruption to the overall shape of the peak is minimal. It is also interesting to note the flattening of the peak when the filter is applied every time step as shown in Fig. 4.1(b). There is a seven or nine point plateau that arises. This will be seen to be a common feature of non-linear filters as well and is one of the main criticisms levelled against some of the non-linear smoothing schemes. It seems from this example that a certain amount of peak flattening may well be unavoidable. The reason why the flattening occurs is that the diffusion flattens out the extrema to roughly a three point plateau, but the high order scheme then produces over-shoots at either end of this plateau. These over-shoots are themselves flattened out and so the plateau gradually widens. In this and several other cases, the peak was also slightly concave. This slight dip near the peak has two causes. Firstly the over-shoots take some time to occur and are therefore slightly less damped. Secondly, as the plateau widens, the slope on the sides of the

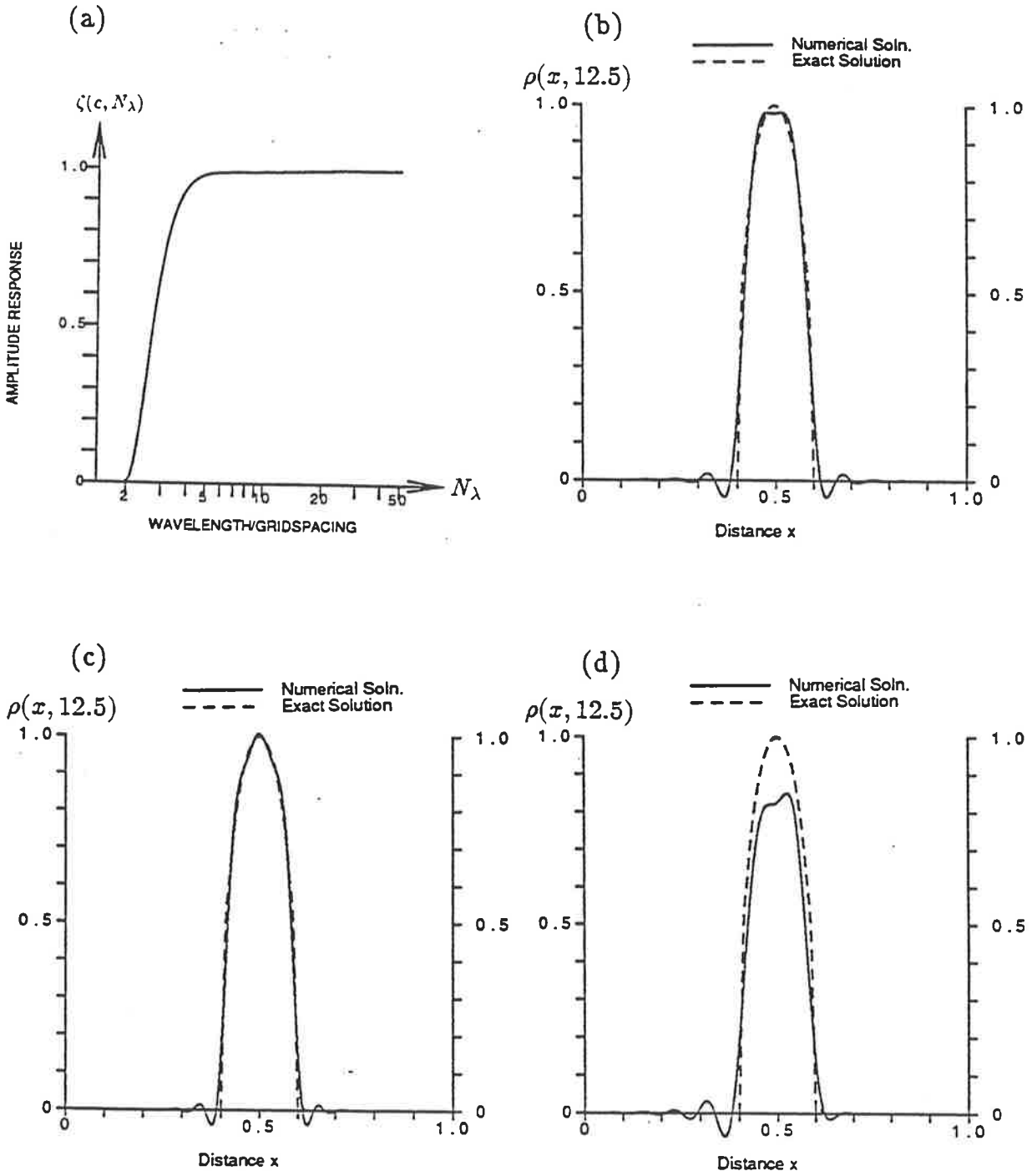

Figure 4.1: Illustration of the performance of the filter, Eq. (4.1). The diagrams show (a) the amplitude response, ζ, (b) the use of the filter every time-step with NS5, Eq. (3.60) in advecting a semi-elliptical pulse with cyclic boundary conditions and (c) the use of the filter every ten time-steps for the same test. Diagram (d) shows the results from filtering fourth order CTCS, Eq. (3.14), every time-step.

pulse becomes steeper, which in turn causes larger over-shoots.

The main conclusion from this example is that linear filtering will not provide smooth results without seriously affecting the amplitude of the pulse. It also shows in which regions the diffusion is most needed, namely where there was a sudden change in the first derivative, at the leading and trailing edges of the pulse. This example does flatter filtering somewhat in that the filter had quite a sharp cutoff and the high order scheme was very accurate. If a less accurate scheme was used, such as fourth order CTCS, Eq. (3.14), as shown in Fig. 4.1(d) the results are quite poor. Not only are the oscillations still present but the peak is also rather distorted. To remove these oscillations the filter must involve much stronger damping, but this will seriously degrade the pulse amplitude, in much the same manner as first order upwinding, Eq. (2.27).

This style of numerical filter is equivalent to an approximation to the diffusion equation

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(\alpha \frac{\partial \rho}{\partial x}\right) \tag{4.2}$$

with an appropriate choice of parameters to obtain the desired amplitude response. Nevertheless, each time the results are filtered, numerical diffusion is added to the solution and this diffusion accumulates over time. Filtering the solution too strongly results in too much residual diffusion being found in the final results. As was seen in Figs. 2.2, 3.1, if there is insufficient damping (i.e. the filtering is not done often enough) the oscillations can become so severe that they are comparable, in magnitude, to the original waveform. In such cases removing the oscillations by filtering once only will damp out most of the waveform.

The residual diffusion left by linear filters cannot be removed, a posteriori, by a direct inverse scheme, since any linear numerical model that accurately models the diffusion equation cannot be inverted in a manner that does not generate any oscillations. The non-existence of such an inverse is due to the "anti-diffusion"

equation

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left(-\alpha \frac{\partial \rho}{\partial x}\right) \quad , \quad \alpha > 0 \tag{4.3}$$

producing results that grow exponentially. This exponential growth will apply to roundoff errors in any linear scheme that approximates this equation. The amplification of these roundoff errors will generate oscillations in the "anti-diffused" solution.

A further problem with filtering is that of *"ringing"* as discussed by Shapiro (1970). This is where the filter produces oscillations of its own. These oscillations first appear at the grid-points adjacent to the boundaries of the solution domain. These oscillations may grow to be many orders of magnitude larger than the maximum value of the numerical solution. This phenomenon occurs when high order filters (i.e. filters that have a very sharp cut off) are used in problems with specified boundary conditions. This is due, in part, to inappropriate boundary conditions being applied to the filter. The boundary conditions must be those of the physical problem but these are not necessarily consistent with solving the diffusion equation and can thereby generate severe oscillations, defeating the purpose of filtering.

From this it can be concluded that linear filtering schemes are not appropriate as general techniques for smoothing numerical solutions, due to the difficulties associated with determining how strongly to diffuse the solution, how often the solution should be filtered and, overcoming ringing near the boundaries. An additional problem is that the continual addition of diffusion can have a significant impact on the overall behaviour of the solution, either by the accumulation of numerical diffusion or through the suppression of some non-linear interaction. By using non-linear filters it is possible to remove some of the main problems with linear filtering. Some of the more common non-linear filtering schemes are discussed below.

4.2 CIP

A method based on polynomial approximation was developed by Takewaki and Yabe (1987). The Cubic-Interpolated Pseudo-Particle Method is based on interpolating the value of $x_j - w\Delta t$ in much the same fashion as the explicit techniques discussed in Chapter Two, with the difference being that the interpolation is done in such a way as to avoid over- and under-shoots in the numerical results. The full form of the CIP method is the implicit difference scheme

$$\begin{aligned}
\mathcal{R}_j^{n+1} &= \rho_{j-2}^n - \frac{3}{2}\rho_j^n + \frac{4}{3}\rho_{j+1}^n \\
&\quad - \frac{c_{j+\frac{1}{2}}}{\Delta x}\left[\rho_{j+1}^n - \rho_j^n\right] + \frac{c_{j-\frac{1}{2}}}{\Delta x}\left[\rho_j^n - \rho_{j-1}^n\right] \\
&\quad + \frac{c_{j+\frac{1}{2}}^2}{2\Delta x}\left[\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n\right] \\
&\quad - \frac{c_{j-\frac{1}{2}}^2}{2\Delta x}\left[\rho_j^n - 2\rho_{j-1}^n + \rho_{j-2}^n\right]
\end{aligned} \tag{4.4}$$

$$\begin{aligned}
\frac{18}{192}\rho_{j-1}^{n+1} &+ \frac{156}{192}\rho_j^{n+1} + \frac{18}{192}\rho_{j+1}^{n+1} = \rho_j^n \\
&+ \frac{5\Delta x}{192}\left[\mathcal{R}_{j+1}^{n+1} - \mathcal{R}_{j+1}^n + \mathcal{R}_j^{n+1} - \mathcal{R}_j^n\right] \\
&- \frac{5\Delta x}{192}\left[\mathcal{R}_j^{n+1} - \mathcal{R}_j^n + \mathcal{R}_{j-1}^{n+1} - \mathcal{R}_{j-1}^n\right] \\
&+ \frac{18}{192}\left[\rho_{j+1}^n - \rho_j^n\right] - \frac{18}{192}\left[\rho_j^n - \rho_{j-1}^n\right] \\
&- \frac{c_{j+\frac{1}{2}}}{2}\left[\rho_{j+1}^n + \rho_j^n - \frac{\Delta x}{4}(\mathcal{R}_{j+1}^n - \mathcal{R}_j^n)\right] \\
&+ \frac{c_{j-\frac{1}{2}}}{2}\left[\rho_j^n + \rho_{j-1}^n - \frac{\Delta x}{4}(\mathcal{R}_j^n - \mathcal{R}_{j-1}^n)\right] \\
&- \frac{c_{j+\frac{1}{2}}^2}{2}\left[\frac{3}{2}(\rho_{j+1}^n + \rho_j^n) + \frac{\Delta x}{4}(\mathcal{R}_{j+1}^n + \mathcal{R}_j^n)\right] \\
&+ \frac{c_{j-\frac{1}{2}}^2}{2}\left[\frac{3}{2}(\rho_j^n + \rho_{j-1}^n) + \frac{\Delta x}{4}(\mathcal{R}_j^n + \mathcal{R}_{j-1}^n)\right] \\
&- \frac{c_{j+\frac{1}{2}}^3}{6}\left[\mathcal{R}_{j+1}^n - \mathcal{R}_j^n\right] + \frac{c_{j-\frac{1}{2}}^3}{6}\left[\mathcal{R}_j^n - \mathcal{R}_{j-1}^n\right] \\
&- \frac{c_{j+\frac{1}{2}}^4}{2}\left[\rho_{j+1}^n + \rho_j^n - \frac{\Delta x}{2}(\mathcal{R}_{j+1}^n + \mathcal{R}_j^n)\right] \\
&+ \frac{c_{j-\frac{1}{2}}^4}{2}\left[\rho_j^n + \rho_{j-1}^n - \frac{\Delta x}{2}(\mathcal{R}_j^n + \mathcal{R}_{j-1}^n)\right] .
\end{aligned}$$

Takewaki and Yabe also gave an alternative approximate form that is explicit, that closely approximates the implicit form.

Table 4.3 shows that this scheme is also very inefficient as the errors are comparable to those of first order upwinding when the difference in CPU time per time-step is taken into account. The RMS errors for the CIP method with $J = 100$ are similar to those of first order upwinding with $J = 200$ and 400.

That the scheme should be so diffusive, is consistent with the result of Godunov (1959) since the CIP method can be very closely approximated by an explicit (linear) scheme. The implicit form should not be of a high order of accuracy either. This can be seen in Fig. 4.2(b) which illustrates the poor amplitude response of this scheme, similar in fact, to that of first order upwinding shown in Fig. 2.1(b). The scheme requires the calculation of derivatives of ρ as does the scheme of Holly and Preissmann, but to guarantee smoothness, the derivatives must be calculated using the same cubic as used to estimate $\rho(x_j - w\Delta t,\ t^n)$.

If the same equation as Holly and Preissmann's scheme is used to estimate the derivatives, the amplitude response (and hence the accuracy) of this scheme is significantly improved, but the numerical solution then contains oscillations.

4.3 Smolarkiewicz's Scheme

Linear filtering, as discussed in the previous section, is based on taking an oscillatory numerical solution and attempting to produce smooth results. An alternative is to take an already overly diffused solution and attempt to remove this diffusion at the same time as the initial calculations are performed so that the solution remains smooth. One such scheme based on this idea was developed by Smolarkiewicz (1983).

Consider first order upwinding, which always produces smooth results. From the modified equivalent equation we obtain an expression for the amount of diffusion introduced by this scheme, namely

$$\alpha_{UW1} = \frac{1}{2}(|u|\,\Delta x - \Delta t u^2) \quad . \tag{4.5}$$

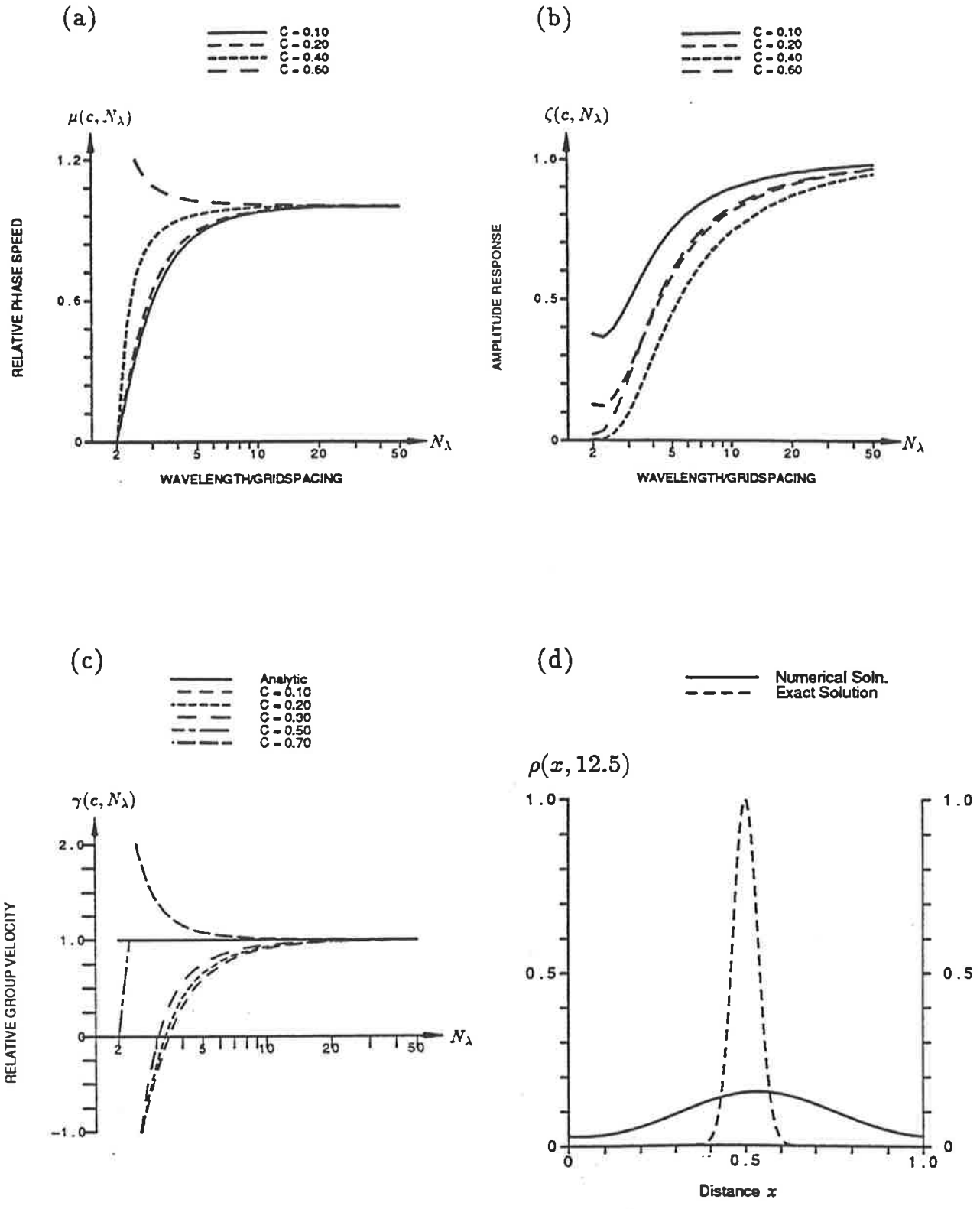

Figure 4.2: Illustration of the performance of the CIP scheme, Eq. (4.4). The diagrams show (a) the relative phase speed, μ, (b) the amplitude response, ζ, (c) the relative group velocity, γ, and (d) the results of the numerical test case with a Gaussian pulse as the initial condition and cyclic boundary conditions.

If the results of first order upwinding are used as initial conditions to the problem of solving

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}(\alpha_{UW1}\frac{\partial \rho}{\partial x}) \quad , \tag{4.6}$$

the final solution should have the diffusion removed from the upwinded solution. It should be noted that this does not remove all the numerical damping as the higher order amplitude errors corresponding to the higher order terms in the modified equivalent equation remain, but it does remove the dominant amplitude errors associated with this scheme. The obvious difficulty is how to solve Eq. (4.6) and retain the smooth nature of the solution, since any linear finite difference approximation will generate more oscillations as roundoff errors will be negatively diffused (i.e.magnified) along with the rest of the solution. Smolarkiewicz suggested rewriting Eq. (4.6) as

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}(K\rho) \tag{4.7}$$

where $K = \alpha(\partial\rho/\partial x)/\rho$ which is just the advection equation, Eq. (2.1), with velocity $w = K$. This equation can be solved by first order upwinding, giving a solution to Eq. (4.6) which is smooth if K can be accurately estimated. Smolarkiewicz suggested the approximation

$$K_{j+\frac{1}{2}} = \frac{(\left|u_{j+\frac{1}{2}}\right|\Delta x - \Delta t u^2_{j+\frac{1}{2}})}{\Delta x}\left(\frac{\rho^*_{j+1} - \rho^*_j}{\rho^*_{j+1} + \rho^*_j + \epsilon}\right) \tag{4.8}$$

where ρ^*_j is the value obtained from the first upwind step, and ϵ is some small parameter so that $K = 0$ if $\rho^*_{j+1} = \rho^*_j = 0$. This choice of K will also always give a stable scheme, as noted by Smolarkiewicz.

This scheme is obviously an improvement on linear filtering since it attempts to do something about the residual diffusion found in the numerical solution, however, it does suffer from some drawbacks. For instance, there is no reason to assume that it in any sense minimizes the residual diffusion. Indeed, Smolarkiewicz remarked that the residual diffusion can be successively reduced by repeatedly applying the antidiffusion step. The CPU time, however, will also increase linearly with the

number of additional applications of the antidiffusive step, and the return (measured by the improvement in accuracy) rapidly diminishes.

Applying the antidiffusive time-step three times does sharpen the peak to some extent as seen by comparing Fig. 4.3(a) with Fig. 4.3(c). The number of times the antidiffusive step is denoted by the variable ITER in the figures and tables. The iteration of the antidiffusive step both enlarges the upstream skew of the pulse and produces a sharper peak, whether it is warranted or not. The first side-effect explains the order of magnitude increase in the third moment of the numerical solution, given in Table 4.1. The second side-effect is shown in Fig. 4.4(a,c) which indicate how Smolarkiewicz's scheme produces a similar shaped pulse irrespective of the initial condition, with the curvature of the sides of the numerical pulse being opposite to that of the analytic pulse. Such behaviour is exacerbated by iterating the antidiffusive step, and this can be seen when other waveforms (square and triangular pulses for example) are used. The odd behaviour of the RMS errors with increasing resolution, as shown in Table 4.3, are also due to the Gaussian pulse being close to the favoured waveform of this scheme. The RMS errors for $J = 400$ and 800 or for the semi-ellipse test are more representative of the general performance of this scheme. This type of problem has been found to occur with several of the techniques discussed in this chapter, with different techniques having particular favoured waveforms. For this reason, comparisons will continue to be based on the Gaussian and the semi-ellipse test cases.

This systematic distortion of profiles can be seen by comparing Fig. 4.3(a) where the antidiffusive step is performed once per time-step and Fig. 4.3(c) where the scheme is applied three times per time-step. This is also apparent in the error measures given in Table 4.1. It should be noted that iterating the antidiffusive step also introduces a substantial lag in the position of the peak and shows the pulse upstream.

Each antidiffusive step still leaves some residual diffusion though, due to the fact

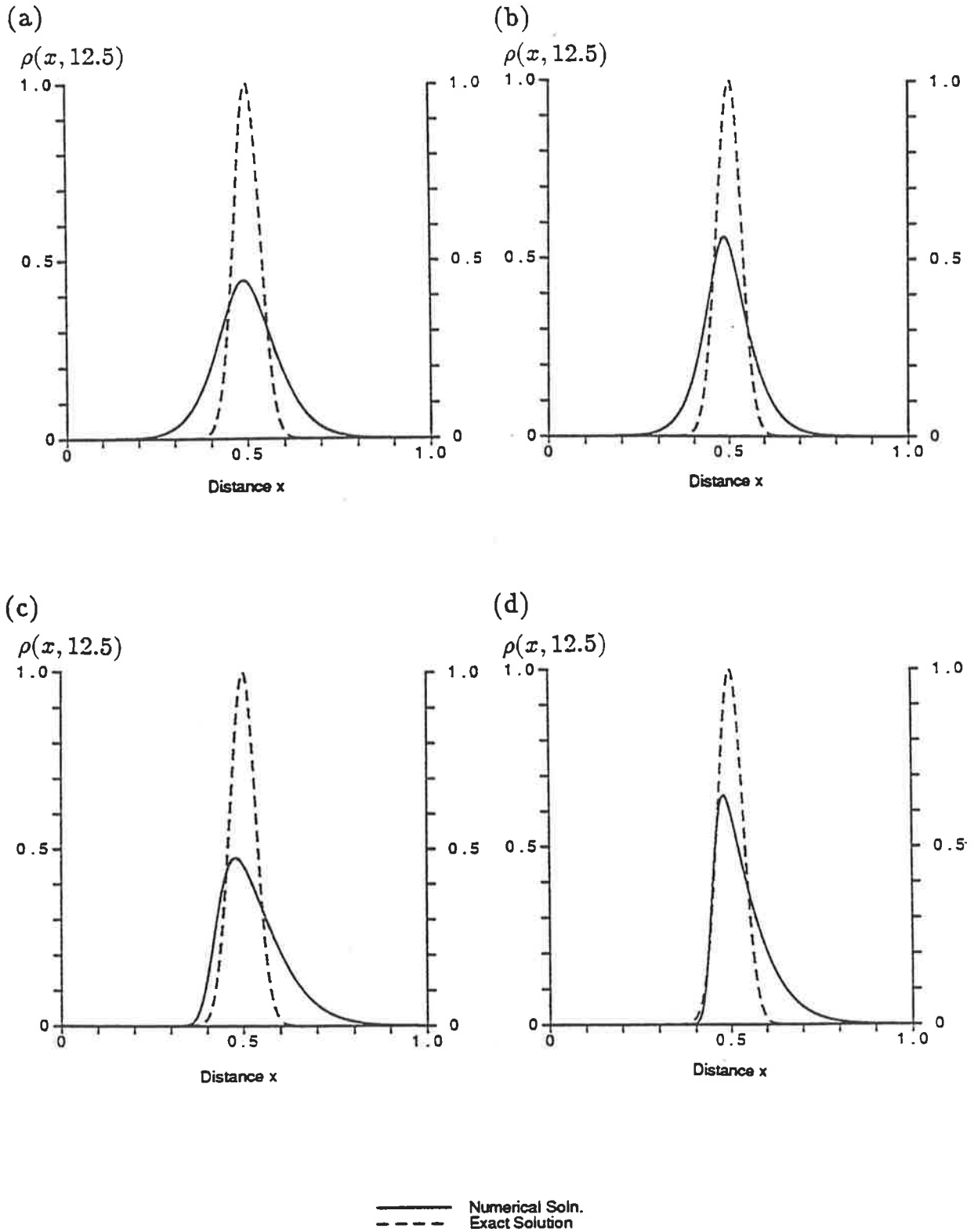

Figure 4.3: Illustration of the performance of Smolarkiewicz's scheme. The diagrams show (a) the basic scheme, (b) the effect of scaling the antidiffusive velocities by Sc=1.06, (c) the impact of iterating the antidiffusive step (ITER=3) and (d) the effect of using both Sc=1.06 and ITER=3. The initial condition is a Gaussian pulse with cyclic boundary conditions.

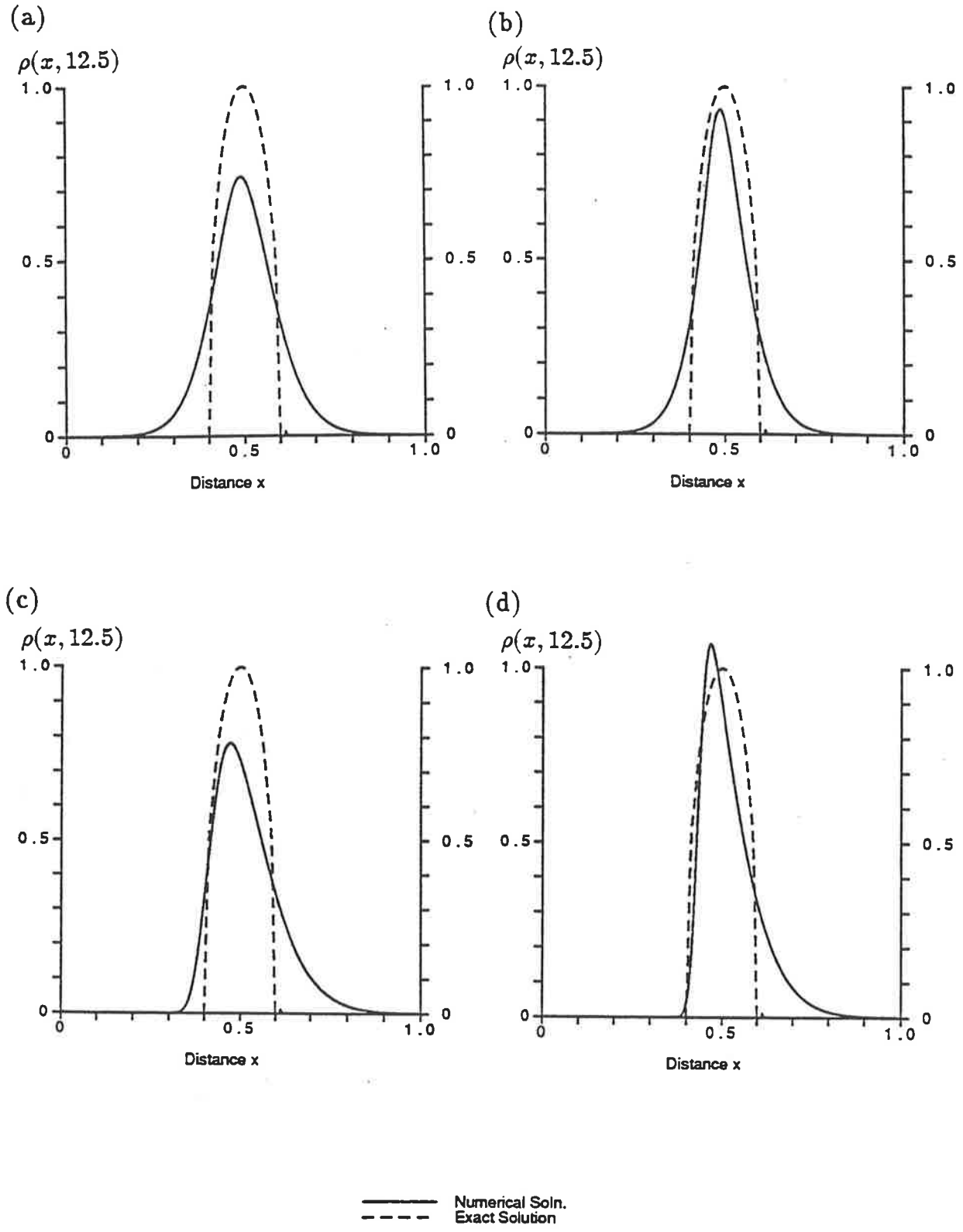

Figure 4.4: As for Fig. 4.3 but with a semi-elliptical initial condition. The diagrams show the performance of (a) the basic scheme, (b) using Sc=1.06, (c) using ITER=3, (d) using both Sc=1.06 and ITER=3.

that first order upwinding is being used here as well. A second suggestion was to scale the antidiffusive velocities $K_{j+\frac{1}{2}}$. Smolarkiewicz found empirically that the best scaling factor was about 1.06, because beyond this the calculations could become unstable. This scaling factor is denoted by Sc in the figures and tables. There was no real justification for this except that it seemed to work, in that a peak does become sharper. Scaling the antidiffusive velocities has the added advantage over repeating the antidiffusive step, in that it is obviously very quick. Magnifying the velocities has the disadvantage that peaks again become skewed as in Fig. 4.3(b), and the symmetry of the previous scheme is lost. If these two suggestions are combined then the faults of both schemes are compounded as shown in Fig. 4.3(d).

Another problem with this scheme is that it admits over-shoots in the numerical solution. This can be seen by examining a sudden jump in the solution, i.e. the point x_k where $\rho_j^n = 1$ to the right of x_k (i.e. $j \geq k$) and $\rho_j^n = 0$ to the left ($j < k$).

Using first order upwinding alone gives

$$\rho_j^* = \begin{cases} 1 & j > k \\ 1-c & j = k \\ 0 & j < k \end{cases} \tag{4.9}$$

from which the antidiffusive velocities are obtained, namely

$$K_{j+\frac{1}{2}} = \begin{cases} \Delta x(|c| - c^2)c/\left[(2-c)\Delta t\right] & j = k \\ 0 & j \neq k \end{cases} \tag{4.10}$$

and finally

$$\rho_k^{n+1} = 1 + c^2(1-c)^2/(2-c) \quad . \tag{4.11}$$

Stability of the scheme requires that $0 \leq c \leq 1$ so that there is a possible maximum over-shoot of 4.2% of the height of the jump at $c = (\sqrt{17}-3)/7 \sim 0.562$. This scheme will produce over-shoots but not under-shoots, because the antidiffusive velocities, $K_{j+\frac{1}{2}}$, will be directed towards the region of higher density, and hence the oscillations only take the form of over-shoots. Although it was claimed by Smolarkiewicz (1983) that this scheme preserves monotonicity of the solution, it is not actually the case, for if the results are examined after one or two time-steps, there are oscillations in the results. These oscillations do not produce negative values but they can cause the

numerical solution to exceed the least upper bound of the analytic solution, even in the constant velocity case of advection. The presence of these oscillations is easily verified by using a square wave as an initial condition. If such a test is performed, with the width of the pulse being twenty grid-spacings, then these oscillations persist for up to 240 time-steps. They eventually disappear when the numerical solution forms a definite peak and develops into the scheme's favoured waveform. While, in general, over-shoots do not cause as many problems as under-shoots, they are undesirable in algorithms for solving general non-linear problems.

Another reason for examining other approaches is that it is difficult to use Smolarkiewicz's approach to obtain more accurate smooth solutions, since of the many schemes devised by the author and co-workers, none have had less diffusion than first order upwinding. If such a scheme could be found and it is stable for small Courant numbers then it could be used here to give better results, since less numerical diffusion in the original scheme should mean less residual diffusion after the antidiffusive step.

4.4 van Leer's Schemes

In the discussion so far there has been no definition of what is meant by numerical results being smooth. A commonly used definition is that schemes should be positive definite. That is, if the initial condition is non-negative and there are appropriate boundary conditions,[1] then the numerical solution also should always be non-negative. The scheme should also conserve material and so the numerical solution should not be identically zero except in the two trivial cases of either, no material present initially and none entering the system, or all the material being advected out of the system. Thus, in general, the operator $\mathcal{L}$, corresponding to the

[1]Examples of "appropriate boundary conditions" are

1. Dirichlet boundary conditions with all values being non-negative, and
2. Neumann boundary conditions where the flux is into the region.

finite difference equation must satisfy the two conditions

1. $[\mathcal{L}(\underset{\sim}{\rho})]_j \geq 0$ for all $\underset{\sim}{\rho}$ such that $\rho_j^0 \geq 0$ for all j

2. $\mathcal{L}(\underset{\sim}{\rho}) = \underset{\sim}{0}$ if and only if $\underset{\sim}{\rho} = \underset{\sim}{0}$

The second part will obviously be satisfied by any conservative difference scheme. For if there is no gain or loss of material to the system except through the boundaries (which would be the same as for the true solution $\hat{\rho}$) then the numerical solution ρ cannot become identically zero unless the true solution does. These two constraints mean that the difference operator should be a positive definite operator. This definition is, however, not quite strict enough, as by this definition, Smolarkiewicz's scheme is positive definite but may still produce oscillations.

An alternative definition of smoothness is that the difference schemes should preserve monotonicity of the data. This was described in detail by van Leer (1973) who gave two different criteria for schemes to be monotonicity preserving (or "*monotonic*"). The first was that there should be no over-shoots or under-shoots by the scheme. The second was an alternative stipulation that ρ_j^{n+1} should lie between the two values ρ_j^n and ρ_{j-1}^n.

This latter definition forces a slight reduction in the height of a numerical peak. This can be seen by considering a peak that exactly coincides with a grid-point, x_p. One time-step later, the peak will no longer coincide with x_p and the maximum permissible value of the solution will equal $\max\{\rho_p,\ \rho_{p+1}\}$ where the maximum is calculated over the time-advanced solution. Under the restriction that $\rho_{j-1}^n \leq \rho_j^{n+1} \leq \rho_j^n$, this maximum value will remain as an upper bound for the solution throughout the computation.

Any conservative scheme that satisfies either of these criteria will be positive definite, for if a scheme cannot admit under-shoots, then it cannot take non-negative

data and generate negative values. Similarly, if a scheme satisfies the second condition, then ρ_j^{n+1} is greater than, or equal, to the minimum of ρ_j^n and ρ_{j-1}^n for all positive n and, as both of these are non-negative, then so is ρ_j^{n+1}.

Van Leer (1973) showed how this *no over-shoot/no under-shoot* criterion could be used for schemes based on interpolation of the value of $\rho(j\Delta x - w\Delta t,\ n\Delta t)$ to estimate ρ_j^{n+1}. As an example, the Lax-Wendroff scheme was considered. By examining the x-coordinate of turning point of the interpolating quadratic, Eq. (2.28), it was found that the Lax-Wendroff scheme gives over- or under-shoots when

$$\zeta_{j+\frac{1}{2}} = \left| \frac{2(\rho_j^n - \rho_{j-1}^n)}{\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n} \right| < 1 \quad . \tag{4.12}$$

To overcome this, van Leer suggested that if this condition holds then the quadratic should only be forced to pass through (x_{j-1}, t^n) and (x_j, t^n) and be tangent to the line $\rho = \rho_{\min}^{n+1}$ where $\rho_{\min}^{n+1}$ is some lower bound on ρ_j^{n+1}. Often $\rho_{\min}^{n+1}$ can be taken to be zero, as this means that there will be no negative values appearing. In any case, the physics of the problem being modelled dictates what value should be used here. This modified polynomial gives the difference scheme

$$\begin{aligned} \rho_j^{n+1} = \rho_j^n \quad &- \quad c_{j+\frac{1}{2}}\rho_j^n + c_{j-\frac{1}{2}}\rho_{j-1}^n \\ &- \quad \frac{\sigma_{j+\frac{1}{2}}}{2} c_{j+\frac{1}{2}}(1 - c_{j+\frac{1}{2}})(\rho_{j+1}^n - \rho_j^n) \\ &+ \quad \frac{\sigma_{j-\frac{1}{2}}}{2} c_{j-\frac{1}{2}}(1 - c_{j-\frac{1}{2}})(\rho_j^n - \rho_{j-1}^n) \end{aligned} \tag{4.13}$$

where $\sigma_{j+\frac{1}{2}} = \min(\zeta_{j+\frac{1}{2}}, 1)$. The terms involving σ correspond to negative diffusive fluxes, i.e. these are the fluxes that have been added to first order upwinding to remove some of the numerical diffusion. If $\zeta_{j+\frac{1}{2}} > 1$ then the scheme reverts to Lax-Wendroff but if $\zeta_{j+\frac{1}{2}} < 1$ then the "negative" diffusive fluxes are reduced. This is equivalent to adding diffusion to the Lax-Wendroff scheme in order to produce smooth results. Since the inequality Eq. (4.12) detects over- and undershoots, it was shown by van Leer (1973) that Eq. (4.13) also prevents overshoots.

This technique may be used with any one of the upwinding schemes, since they are all based on interpolation. With the higher order schemes, however, the criteria

for under- and over-shooting become very complicated since finding the turning points of the high order interpolating polynomials becomes increasingly more difficult.

One of the problems with this approach is that there is no guarantee that the amount of diffusion added to the results is in any way minimized. To overcome this problem van Leer (1974) demonstrated how to convert Fromm's scheme

$$\rho_j^{n+1} = \rho_j^n - f_{j+\frac{1}{2}}^F + f_{j-\frac{1}{2}}^F \quad , \tag{4.14}$$

where

$$f_{j+\frac{1}{2}}^F = c_{j+\frac{1}{2}}\rho_j^n + \frac{c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}})}{4}(\rho_{j+1}^n + \rho_{j-1}^n) \quad , \tag{4.15}$$

into a monotonic scheme with minimal diffusion added to the original scheme. The fluxes for the new scheme are given by

$$f_{j+\frac{1}{2}}^{vL2} = f_{j+\frac{1}{2}}^F - \frac{\sigma_{j+\frac{1}{2}}}{4}c_{j+\frac{1}{2}}(1-c_{j+\frac{1}{2}})(\rho_{j+1}^n - 2\rho_j^n + \rho_{j-1}^n) \tag{4.16}$$

where

$$\sigma_{j+\frac{1}{2}} = \frac{\left|\rho_{j+1}^n - \rho_j^n\right| - \left|\rho_j^n - \rho_{j-1}^n\right|}{\left|\rho_{j+1}^n - \rho_j^n\right| + \left|\rho_j^n - \rho_{j-1}^n\right|} \tag{4.17}$$

is a smoothness monitor, similar in nature to $\zeta_{j+\frac{1}{2}}$ in Eq. (4.13).

The performance of these two schemes is shown in Fig. 4.5(a,b) and Fig. 4.6(a,b) and Tables 4.1-4.3. Clearly Eq. (4.16) is the better of the two, with Eq. (4.13) introducing a significant phase lag and producing a pulse that is more skewed upstream. The common feature of these two schemes is the considerable flattening of sharp peaks.

While this approach is successful in converting the two suggested schemes (Lax-Wendroff and Fromm's) into monotonic difference schemes, it is not a feasible way of converting an arbitrary difference operator into a monotonic operator. Certainly for implicit schemes it is not at all clear how the technique should be used to ensure monotonicity. So in an attempt to better utilize the high order schemes developed in Chapter Three, other techniques need to be examined. Other schemes were also

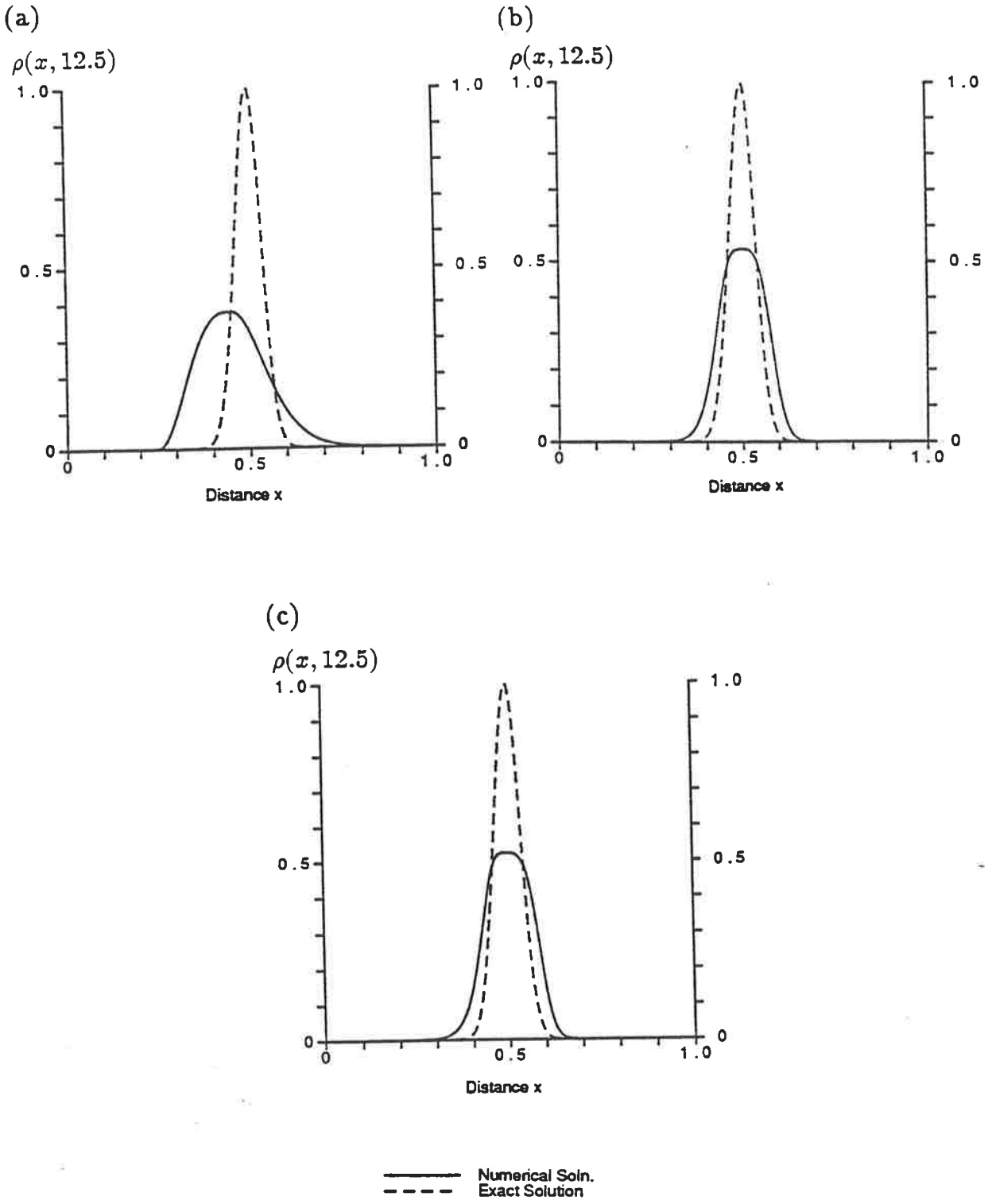

Figure 4.5: Illustration of the performance of the schemes (a) van Leer (I), Eq. (4.13), (b) van Leer(II), Eq. (4.16), and (c) the Self-adjusting hybrid scheme, Eq. (4.19). The test case is the advection of a Gaussian pulse with cyclic boundary conditions.

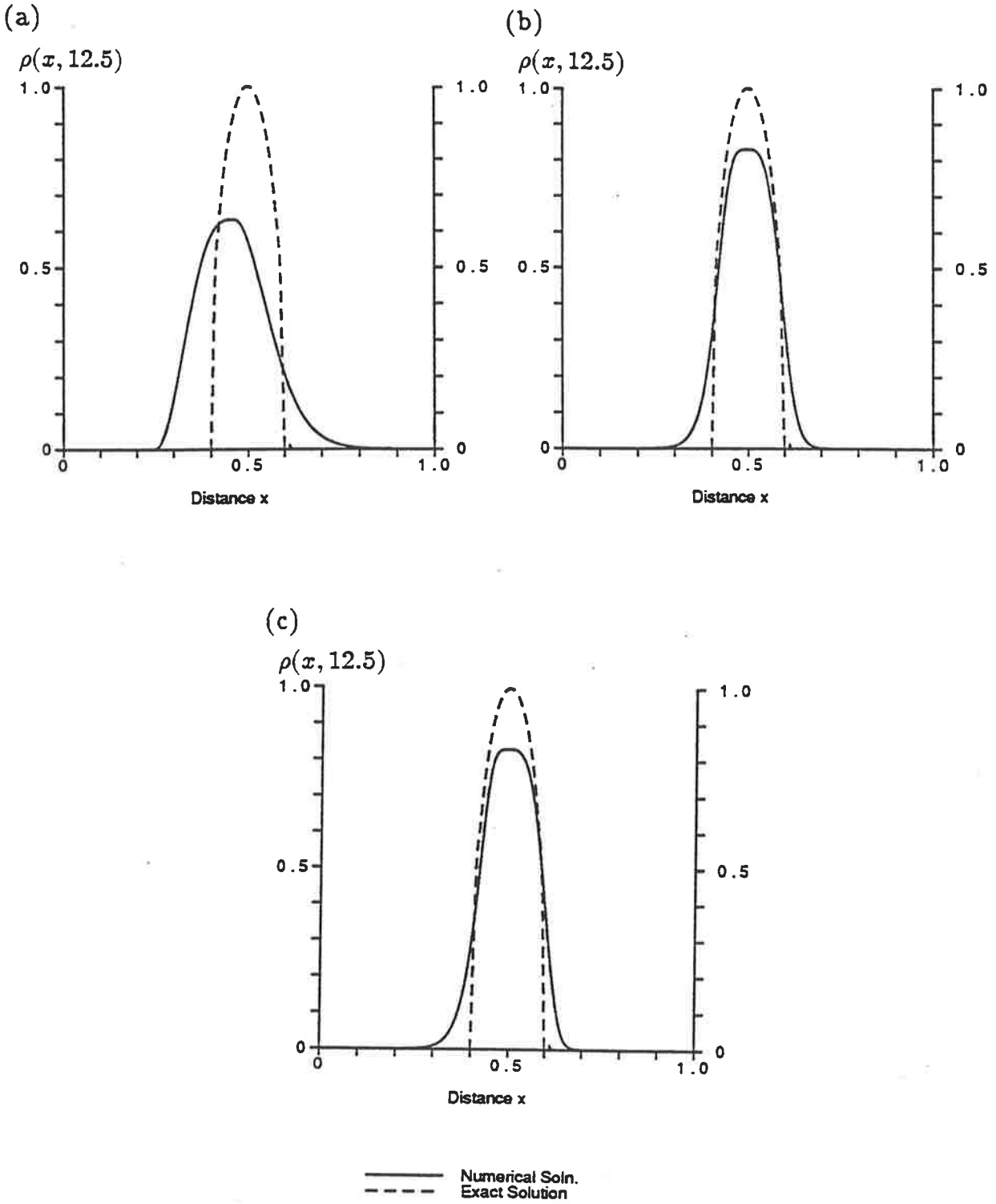

Figure 4.6: Illustration of the performance of the schemes (a) van Leer (I), Eq. (4.13), (b) van Leer(II), Eq. (4.16), and (c) the Self-adjusting hybrid scheme, Eq. (4.19). The test case is the advection of a semi-elliptical pulse with cyclic boundary conditions.

developed by van Leer (1977) but these all used slope limiters. The idea of using limiters leads to either TVD or FCT schemes discussed below and so these latter schemes of van Leer will not be discussed here.

4.5 Self-Adjusting Hybrid Scheme

It has been noted that sharp changes in the numerical solution cause the spurious oscillations of high order schemes. Harten (1978) suggested that by monitoring the smoothness of the solution, a choice could be made between a smooth (but diffuse) low order solution and a high order non-diffuse solution. In regions where the solution is sufficiently smooth, the high order solution will also give smooth results. In regions of dramatic change in the solution the low order solution is used. This is a generalization of van Leer's schemes where the smoothness monitors were ζ_j and σ_j. Harten suggested that the general form for the difference operator should be

$$\mathcal{L}\{\rho_j^n\} = (\theta_j \mathcal{L}^L + (1-\theta_j)\mathcal{L}^H)\{\rho_j^n\} \tag{4.18}$$

where $0 \leq \theta_j \leq 1$ is non-linear. The operators $\mathcal{L}^L$ and $\mathcal{L}^H$ refer to the smooth and high order schemes that are combined to form the hybrid scheme. Harten derived expressions for θ_j for the case where both high and low order schemes were explicit, but determining the equivalent expression for general implicit schemes is difficult. This process is, however, very effective. For example, using the Lax-Wendroff scheme, Eq. (2.29), as the high order scheme and first order upwinding, Eq. (2.27), as the low order scheme gives

$$\begin{aligned}
\rho_j^{n+1} &= \rho_j^n - f_{j+\frac{1}{2}} + f_{j-\frac{1}{2}} \\
f_{j+\frac{1}{2}} &= c_{j+\frac{1}{2}}\rho_j^n + \left[\frac{1}{2}c_{j+\frac{1}{2}}\left(1 - c_{j+\frac{1}{2}}\right) - \frac{\theta_{j+\frac{1}{2}}}{8}\right]\left(\rho_{j+1}^n - \rho_j^n\right) \\
\theta_{j+\frac{1}{2}} &= \max\left(\hat{\theta}_{j+\frac{1}{2}},\ \hat{\theta}_{j-\frac{1}{2}}\right) \\
\hat{\theta}_{j+\frac{1}{2}} &= \frac{\left|\rho_{j+2}^n - \rho_{j+1}^n\right| - \left|\rho_{j+1}^n - \rho_j^n\right|}{\left|\rho_{j+2}^n - \rho_{j+1}^n\right| + \left|\rho_{j+1}^n - \rho_j^n\right|}
\end{aligned} \tag{4.19}$$

As can be seen from Fig. 4.5(c) and Fig. 4.6(c), this scheme produces results very similar to those of van Leer (II), Eq. (4.16). This shows the advantage in attempting to minimize the added diffusion to obtain smooth results, however, such minimization is very difficult for general schemes.

4.6 Total Variation Diminishing (TVD) Schemes

Harten (1983, 1984), introduced the concept of using the Total Variation Diminishing (TVD) assumption,

$$\sum_j \left|\rho_{j+1}^{n+1} - \rho_j^{n+1}\right| = TV(\underline{\rho}^{n+1}) \leq TV(\underline{\rho}^n) = \sum_j \left|\rho_{j+1}^n - \rho_j^n\right| \quad . \tag{4.20}$$

to solve

$$\frac{\partial \rho}{\partial t} + \frac{\partial f(\rho)}{\partial x} = h(\rho) \quad . \tag{4.21}$$

Yee et. al. (1985) and Yee (1986, 1987) have extended this work to obtain a variety of schemes, both implicit and explicit, that are formally second order as well as positive definite. The non-linear components of the resulting schemes take the form of either a slope limiter or a flux limiter. The slope limiters impose constraints on the gradient of u; the flux limiters impose constraints on the gradients of the fluxes. As noted by Yee (1987), flux limiters are generally used in practice. An example TVD scheme for the constant velocity case, where $f(\rho) = u\rho$ and $h(\rho) = 0$, is the explicit scheme developed by Yee (1986), namely

$$\begin{aligned}
\rho_j^{n+1} &= \rho_j^n + \frac{\Delta t}{\Delta x}\left[C_-\left(\hat{a}_{j+\frac{1}{2}} + \hat{\gamma}_{j+\frac{1}{2}}\right)\left(\rho_{j+1}^n - \rho_j^n\right)\right. \\
&\qquad \left. - C_+\left(\hat{a}_{j-\frac{1}{2}} + \hat{\gamma}_{j-\frac{1}{2}}\right)\left(\rho_j^n - \rho_{j-1}^n\right)\right] \qquad (4.22) \\
C_-(z) &= \frac{1}{2}\left(\Psi(z) - z\right) \\
C_+(z) &= \frac{1}{2}\left(\Psi(z) + z\right) \\
\Psi(z) &= \begin{cases} |z| & , |z| < \epsilon \\ \left(z^2 + \epsilon^2\right)/(2\epsilon) & , \text{otherwise} \end{cases} \\
\hat{\gamma}_{j+\frac{1}{2}} &= \begin{cases} 0 & , \left|\rho_{j+1}^n - \rho_j^n\right| < \epsilon \\ \left(g_{j+1} - g_j\right) / \left(\rho_{j+1}^n - \rho_j^n\right) & , \text{otherwise} \end{cases}
\end{aligned}$$

$$
\begin{aligned}
g_j &= S\max\left\{0,\ \min\left\{\sigma\left(c_{j+\frac{1}{2}}\right)\left|\rho^n_{j+1}-\rho^n_j\right|,\right.\right. \\
&\qquad \left.\left. S\sigma\left(c_{j-\frac{1}{2}}\right)\left|\rho^n_j-\rho^n_{j-1}\right|\right\}\right\} \\
S &= \operatorname{sgn}\left(\rho^n_{j+1}-\rho^n_j\right) \\
\sigma(z) &= \frac{1}{2}\left(\Psi(z)-\frac{\Delta t}{\Delta x}z^2\right) \\
\hat{a}_{j+\frac{1}{2}} &= \begin{cases} c_{j+\frac{1}{2}} & , \left|\rho^n_{j+1}-\rho^n_j\right|<\epsilon \\ \left(c_{j+\frac{3}{2}}\rho^n_{j+1}-c_{j+\frac{1}{2}}\rho^n_j\right)/\left(\rho^n_{j+1}-\rho^n_j\right) & , \text{otherwise} \end{cases},
\end{aligned}
$$

where ϵ is a small number reflecting machine precision (taken to be 10^{-16} here).

TVD schemes have been successful in solving the problems for which they are designed, however, if the equation to be modelled is of the form

$$\frac{\partial \rho}{\partial t}+\frac{\partial(w\rho)}{\partial x}=0 \tag{4.23}$$

where w does not necessarily depend on ρ, then Eq. (4.23) must be rewritten in terms of $\underline{u}=[\rho,\ w\rho]^T$, where $\underline{u}$ satisfies Eq. (4.21). This gives a coupled pair of p.d.e's which substantially increases the CPU time required. The CPU times quoted in Table 4.3 are only for the constant velocity case and as such considerably underestimate the actual CPU time required for the general case. This is to be compared with the CPU times of the other schemes which are representative of the times required to solve Eq. (4.23) for general w.

The additional computation can be shown to be necessary, by considering the case of solving Eq. (4.23) when the velocity w is a linearly decreasing function of x. The exact solution in this case is given by

$$\rho(x,t)=\rho_0\left(\left(x+w_0\right)e^{-kt}-w_0\right)e^{-kt} \tag{4.24}$$

where $w=k(x+w_0)$ and $\rho_0(x)$ is the initial condition. If $k=-1$, $w_0=1$ and $\rho_0(x)$ is the square wave

$$\rho_0(x)=\begin{cases}1 & , \delta_1<x<\delta_2 \\ 0 & , \text{otherwise}\end{cases} \tag{4.25}$$

then

$$\sum_j\left|\rho^n_{j+1}-\rho^n_j\right|=e^{t^n}\sum_j\left|\rho_0\left(\left(x_{j+1}+1\right)e^{t^n}-1\right)-\rho_0\left(\left(x_j+1\right)e^{t^n}-1\right)\right| \tag{4.26}$$

but

$$|\rho_0(y_{j+1}) - \rho_0(y_j)| = \begin{cases} 1 & \text{, if the discontinuity is in } [y_j,\ y_{j+1}] \\ 0 & \text{, otherwise} \end{cases} \tag{4.27}$$

for and y_j. For suitable choices of δ_1 and δ_2, the term $|\rho_0(y_{j+1}) - \rho_0(y_j)|$ can only be non-zero for two distinct values of j, giving

$$\sum_j \left|\rho_{j+1}^n - \rho_j^n\right| = 2e^{t^n} \tag{4.28}$$

and so

$$\sum_j \left|\rho_{j+1}^{n+1} - \rho_j^{n+1}\right| > \sum_j \left|\rho_{j+1}^n - \rho_j^n\right| \quad . \tag{4.29}$$

That is, TVD schemes can only model equations of the form Eq. (4.21), and for general advection problems, the p.d.e. must be rewritten in the appropriate form. This gives a system of coupled p.d.e's describing the evolution of, for example, ρ and $w\rho$. The equations can still be made TVD, as it was shown by Yee (1987) that the definition of Total Variation must be modified when modelling systems of p.d.e's. With this new definition it can be shown that TVD schemes can be used to model general advective processes.

The results from using the explicit and implicit TVD schemes developed by Yee (1986) are shown in Figs. 4.7, 4.8 (a,b). The explicit scheme produces an almost symmetric pulse which also contains a large amount of residual diffusion. The implicit scheme produces an even more diffuse pulse which is also skewed upstream.

The major difficulty with these schemes is the amount of CPU time that is required, with other first order schemes producing more accurate results in less time. While these schemes do retain their level of performance in more complicated non-linear problems (when the performance of other schemes deteriorate), TVD schemes do not appear to be very efficient for modelling advective processes in which w is not only a function of ρ.

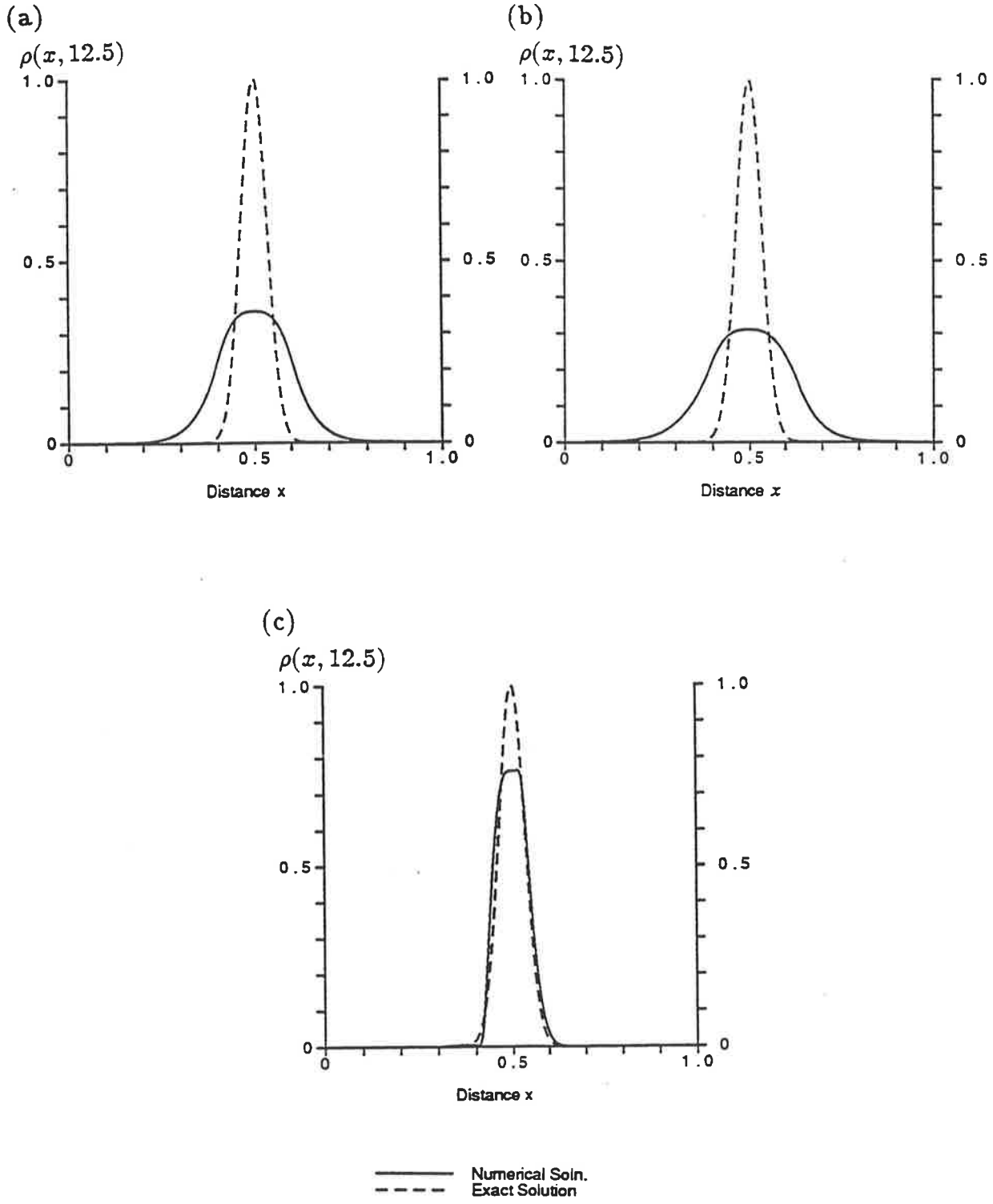

Figure 4.7: Illustration of the performance of the schemes (a) explicit TVD scheme (Yee, 1986), (b) implicit TVD scheme (Yee, 1986) and (c) Phoenical LPE SHASTA. The test case is the advection of a Gaussian pulse with cyclic boundary conditions.

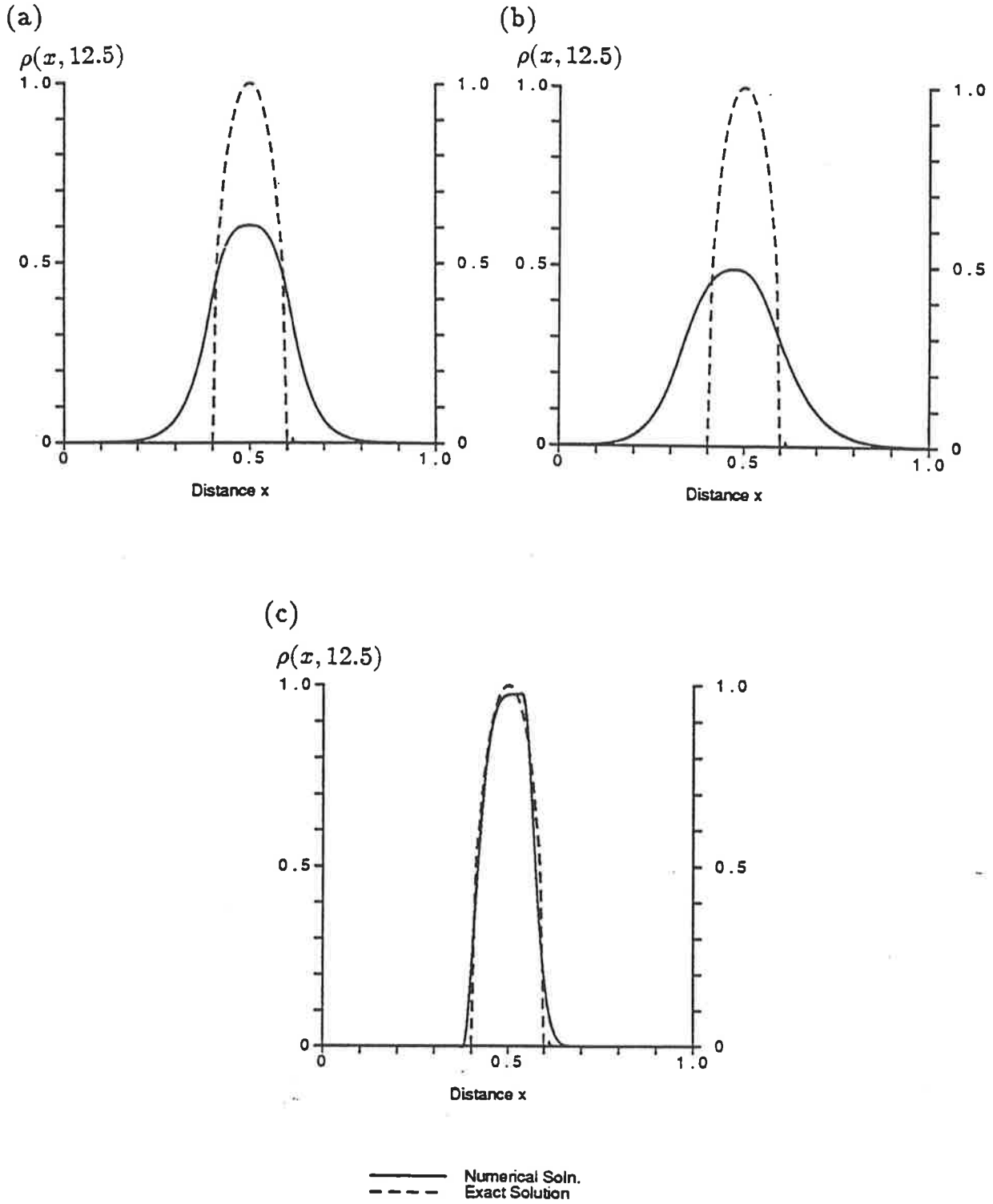

Figure 4.8: Illustration of the performance of the schemes (a) explicit TVD scheme (Yee, 1986), (b) implicit TVD scheme (Yee, 1986) and (c) Phoenical LPE SHASTA. The test case is the advection of a semi-elliptical pulse with cyclic boundary conditions.

4.7 Flux Corrected Transport

Flux Corrected Transport (FCT) was first described by Boris and Book (1973) and then Zalesak (1979) gave a more general interpretation of the technique. As with most of the techniques discussed in this chapter, FCT is also based on a non-linear weighting of a low order smooth solution and a high order non-diffusive solution. A general FCT algorithm is obtained by considering the fluxes into and out of each cell. The fluxes due to the low order scheme are taken as base values and then these are corrected so as to remove the diffusion from the low order scheme. The basic approach is to take low order fluxes and to convert them into high order fluxes, but if this causes an oscillation to appear that was not present in the low order solution then this flux is reduced.

In the algorithms developed by Boris and Book, the low order solution and the high order solution are related, in that one may be constructed from the other. This can be done in two ways. One is to look at the errors introduced by this scheme and then to adjust any free parameters (such as the diffusion present in the scheme) or to add extra terms to correct these errors. By this technique, a high order scheme may be developed from any low order scheme. An alternative method is to take an explicit finite difference technique and introduce sufficient diffusion to ensure that all the coefficients in the difference equation are positive. For if

$$\rho_j^{n+1} = \sum_i a_i \rho_{j+i}^n \tag{4.30}$$

where the a_i's are positive, then, if all the elements of $\underset{\sim}{\rho}^n$ are also positive, so are those of $\underset{\sim}{\rho}^{n+1}$, giving a sufficient condition for a scheme to be positive definite. This is also necessary because if one coefficient is negative then it is not difficult to construct an initial condition that will produce a negative value for ρ_j^{n+1}. For example, if $a_{-1} < 0$ then choosing $\underset{\sim}{\rho}^n = (0, \cdots, 0, 1, 0, \cdots, 0)^T$, where $\rho_{j-1}^n = 1$ for some fixed j will give $\rho_j^{n+1} < 0$. While this is a very artificial case, as long as the value of ρ_{j-1}^n is taken to be sufficiently larger than the surrounding values it is always possible to make $\rho_j^{n+1} < 0$.

The advantage of this approach is that the diffusion added to the high order solution is minimised (but only in a very simple sense). It is important for the low order solution to be as weakly diffusive as possible since the regions containing the most residual diffusion will correspond to the regions where the antidiffusive fluxes are set to zero. Setting these fluxes to zero is equivalent to using the low order solution for the time advanced solution. Hence, if the low order solution is overly diffusive then so will be the advanced solution.

Constructing high and low order solutions that are directly related has computational advantages as the antidiffusive fluxes can be then calculated directly. The antidiffusive fluxes calculated by this process usually have a simple form. This does limit, however, the choice of which methods may be used to produce the high order solution for not all high order schemes can be converted into positive definite schemes by the simple process of adding artificial diffusion, as will be seen in Chapter Five.

Zalesak (1979) described how any low order solution and any high order solution may be combined to give a smooth, high order solution. This approach involves considerations of fluxes into and out of each cell and then limiting them so they do not cause the final solution to be outside some specified interval. In the most general case, this approach is significantly more time consuming than Boris and Book's form, however, it does give a general technique for converting any oscillatory numerical solution into a smooth solution.

An outline of a general FCT algorithm then, is as follows. Given a smooth low order solution, $\{\rho_j^L\}$, and a nondiffusive high order solution, $\{\rho_j^H\}$, it is possible to derive the antidiffusive fluxes from either the low or high order fluxes. These new fluxes are then applied to $\{\rho_j^n\}$ and contain, in general, less diffusion than the low order fluxes but more than the high order fluxes, the balance being chosen so that the values of ρ_j^n do not exceed prescribed limits for any given $(x_j,\ t^n)$, but at the same time attempting to stay as close as possible to the high order fluxes. For example, consider the scheme Phoenical LPE SHASTA described in Boris and Book

(1976a). This was constructed by minimising the residual diffusion in a general three- point explicit finite difference scheme. The resulting low order solution is

$$\begin{aligned}\rho_j^L &= \rho_j^n - \frac{c_{j+\frac{1}{2}}}{2}(\rho_{j+1}^n + \rho_j^n) + \frac{c_{j-\frac{1}{2}}}{2}(\rho_j^n + \rho_{j-1}^n) \\ &\quad + \nu_{j+\frac{1}{2}}(\rho_{j+1}^n - \rho_j^n) - \nu_{j-\frac{1}{2}}(\rho_j^n - \rho_{j-1}^n) \end{aligned} \tag{4.31}$$

where $\nu = (1 + 2c^2)/6$, with antidiffusive fluxes, ϕ, of the form

$$\begin{aligned}\phi_{j+\frac{1}{2}} &= \frac{1}{6}(1 - c_{j+\frac{1}{2}}^2)(\rho_{j+1}^n - \rho_j^n + \delta_{j+\frac{3}{2}} - \delta_{j+\frac{1}{2}}) \\ \delta_{j+\frac{1}{2}} &= \rho_j^L - \rho_j^n - \nu_{j+\frac{1}{2}}(\rho_{j+1}^n - \rho_j^n) + \nu_{j-\frac{1}{2}}(\rho_j^n - \rho_{j-1}^n) \quad .\end{aligned} \tag{4.32}$$

The results of this scheme, presented in Fig. 4.7(c) and Fig. 4.8(c), compare favourably with other methods discussed in this chapter. The good performance is also evident from the error measures in Table 4.1.

An integral part of this entire approach is the flux limiter itself. This is a function that ensures that the modified fluxes do not introduce any new extrema into the low order solution. A suitable choice for such a limiter was the *"minmod"* function described by Boris and Book, namely

$$\begin{aligned}\Delta\rho_j^n &= \rho_{j+1}^n - \rho_j^n \\ S &= \text{sgn}\left(\Delta\rho_j^n\right) \\ f_{j+\frac{1}{2}} &= S\max\left\{0, \min\{S\Delta\rho_{j-1}^n, \left|\phi_{j+\frac{1}{2}}\right|, S\Delta\rho_{j+1}^n\}\right\} \quad .\end{aligned} \tag{4.33}$$

In Zalesak's formulation this is equivalent to taking

$$\rho_j^H = \rho_j^n - \left(\phi_{j+\frac{1}{2}} + f_{j+\frac{1}{2}}^L\right) + \left(\phi_{j-\frac{1}{2}} + f_{j-\frac{1}{2}}^L\right) \tag{4.34}$$

where $f_{j+1\frac{1}{2}}^L$ are the low order fluxes, as defined in Eq. (4.31) and $\phi_{j+\frac{1}{2}}$ are as in Eq. (4.32). The value of ρ_j^{n+1} is restricted to be in the interval $[\rho_j{}^{\min}, \rho_j{}^{\max}]$ where

$$\rho_j{}^{\max} = \max\{\rho_{j-1}^n, \rho_j^n, \rho_j^{n+1}, \rho_{j-1}^L, \rho_j^L, \rho_{j+1}^L\} \tag{4.35}$$

and similarly for $\rho_j{}^{\min}$. This gives limits on the values of the fluxes into and out of each grid-cell. In both formulations, the time advanced solution ρ_j^{n+1} is then

computed by applying the appropriate *limited* fluxes to ρ_j^L. It should be clear that, unless an extremum is advected exactly in an integral number of grid-points, the limiter will impose too tight a restriction on the corrected fluxes. This leads to the phenomenon of "clipping", where a sharp peak is rapidly converted to a plateau. This is quite evident in Fig. 4.7(c) where the peak has been replaced by a relatively flat region about five grid-points wide.

While Zalesak's approach allows for more general schemes to be used to produce the low and high order solutions, there are some points that need to be made. Firstly, by explicitly defining the interval of possible values for ρ_j^n, there is some scope for overcoming the problem of clipping by expanding this interval. Zalesak in fact suggested one such method, whereby the linear extrapolation of the low order solution was used to enhance any extrema there, thus expanding the window $[\rho_j^{\min}, \rho_j^{\max}]$. This does give slight improvement in the retention of peaks. Morrow (1985) has found, however, that this can lead to instabilities in some non-linear problems. For this reason Zalesak's peak reconstruction will not be used.

Hence, it seems most profitable to use Zalesak's approach to derive a FCT algorithm but Boris and Book's limiter for enacting it. If Zalesak's limiter is used without peak extrapolation, it is still more time consuming than Boris and Book's limiter, Eq. (4.34), but the results are identical. The idea of reconstructing the peak may yet provide a way of overcoming clipping but it cannot be done by simple polynomial extrapolation, because linear extrapolation is not satisfactory and higher order polynomials will allow under- and over-shoots.

4.8 Conclusion

This chapter has presented a brief review of the different approaches for obtaining smooth results in modelling advective processes. Rather than an exhaustive comparison of all possible smooth advective difference schemes, the intention was to provide a discussion on the different methodologies, with a view to adapting one

Table 4.1:Error measures for standard difference schemes for the test problem with a Gaussian pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx, by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
CIP	1.8×10^{-1}	7.1×10^{-1}	1.6×10^{-4}	10.76	-9.6×10^{-4}	0.1	0.29
Smolarkiewicz	1.4×10^{-1}	5.6×10^{-1}	-3.5×10^{-2}	5.38	2.6×10^{-3}	-1.0	0.44
- (Sc=1.06, IORD=1)	1.1×10^{-1}	4.6×10^{-1}	-6.5×10^{-2}	3.47	2.3×10^{-3}	-1.4	0.56
- (Sc=1, IORD=3)	1.3×10^{-1}	5.4×10^{-1}	4.6×10^{-1}	6.59	1.5×10^{-2}	-2.0	0.48
- (Sc=1.06, IORD=3)	9.8×10^{-2}	4.1×10^{-1}	5.5×10^{-2}	5.22	1.2×10^{-2}	-1.9	0.65
van Leer (I)	1.3×10^{-1}	5.4×10^{-1}	-3.4×10^{-2}	3.34	1.6×10^{-3}	0.1	0.46
van Leer (II)	1.1×10^{-1}	4.7×10^{-1}	-2.3×10^{-4}	1.64	-1.7×10^{-5}	-0.2	0.53
SAHS (Lax-Wendroff)	1.1×10^{-1}	4.7×10^{-1}	1.2×10^{-2}	1.69	-2.2×10^{-6}	-0.3	0.53
Yee - Explicit TVD	1.6×10^{-1}	6.4×10^{-1}	1.3×10^{-4}	5.68	2.7×10^{-5}	-0.1	0.36
Yee - Implicit TVD	1.8×10^{-1}	7.2×10^{-1}	-3.9×10^{-1}	10.43	5.2×10^{-3}	-2.8	0.29
Phoenical LPE SHASTA	4.9×10^{-2}	2.4×10^{-1}	5.2×10^{-3}	0.27	-1.4×10^{-4}	0.5	0.76

Table 4.2:Error measures for standard difference schemes for the test problem with a semi-elliptical pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
CIP	2.1×10^{-1}	5.1×10^{-1}	4.3×10^{-4}	5.57	-1.6×10^{-3}	0.1	0.49
Smolarkiewicz	1.4×10^{-1}	3.6×10^{-1}	-3.5×10^{-2}	2.62	4.7×10^{-3}	-0.9	0.74
- (Sc=1.06, IORD=1)	1.1×10^{-1}	3.1×10^{-1}	-6.6×10^{-2}	1.56	4.4×10^{-3}	-1.2	0.94
- (Sc=1, IORD=3)	1.3×10^{-1}	3.5×10^{-1}	6.2×10^{-1}	3.34	3.0×10^{-2}	-2.8	0.78
- (Sc=1.06, IORD=3)	1.1×10^{-1}	3.3×10^{-1}	6.9×10^{-1}	0.27	2.6×10^{-2}	-3.2	1.07
van Leer (I)	1.3×10^{-1}	4.8×10^{-1}	-2.0×10^{-2}	1.57	1.9×10^{-3}	0.6	0.74
van Leer (II)	8.4×10^{-2}	3.5×10^{-1}	-2.2×10^{-3}	0.72	2.5×10^{-4}	-0.3	0.83
SAHS (Lax-Wendroff)	8.5×10^{-2}	3.8×10^{-1}	2.3×10^{-2}	0.72	3.6×10^{-4}	-0.5	0.83
Yee - Explicit TVD	1.7×10^{-1}	4.2×10^{-1}	1.1×10^{-3}	2.76	1.2×10^{-4}	-0.1	0.60
Yee - Implicit TVD	2.2×10^{-1}	5.2×10^{-1}	-7.2×10^{-1}	5.20	8.6×10^{-3}	-2.8	0.40
Phoenical LPE SHASTA	4.4×10^{-2}	2.2×10^{-1}	-3.9×10^{-3}	0.37	-3.3×10^{-3}	2.5	0.98

Table 4.3:Comparisons of improvement in R.M.S. error versus increase in CPU time due to increasing resolution. The test problems use either a Gaussian pulse or a semi-circle as the initial condition along with cyclic boundary conditions. The CPU Times are for the calculations with $J = 100$ and 800.

	CPU	Gaussian Pulse				Semi-elliptical Pulse				CPU
Scheme	$J = 100$		$J = 200$	$J = 400$	$J = 800$	$J = 100$	$J = 200$	$J = 400$	$J = 800$	
First Order Upwinding	2	2.2×10^{-1}	2.0×10^{-1}	1.5×10^{-1}	1.1×10^{-2}	2.9×10^{-1}	2.6×10^{-1}	2.2×10^{-1}	1.7×10^{-1}	110
CIP	45	1.8×10^{-1}	1.5×10^{-1}	1.2×10^{-1}	8.2×10^{-2}	2.1×10^{-1}	1.6×10^{-1}	1.2×10^{-1}	8.0×10^{-2}	2900
Smolarkiewicz	12	1.4×10^{-1}	8.3×10^{-2}	3.4×10^{-2}	1.1×10^{-2}	1.4×10^{-1}	8.7×10^{-2}	5.6×10^{-2}	3.5×10^{-2}	770
" (Sc=1.06, IORD=1)	12	1.1×10^{-1}	4.4×10^{-2}	5.2×10^{-2}	4.2×10^{-2}	1.1×10^{-1}	8.6×10^{-2}	6.9×10^{-2}	7.1×10^{-2}	780
" (Sc=1, IORD=3)	29	1.3×10^{-1}	8.1×10^{-2}	4.0×10^{-2}	1.5×10^{-2}	1.3×10^{-1}	8.3×10^{-2}	5.2×10^{-2}	3.3×10^{-2}	1870
" (Sc=1.06, IORD=3)	29	9.8×10^{-2}	7.0×10^{-2}	1.2×10^{-1}	9.8×10^{-2}	1.1×10^{-1}	1.0×10^{-1}	8.0×10^{-2}	7.7×10^{-2}	1900
van Leer (I)	7	1.3×10^{-1}	7.9×10^{-2}	3.7×10^{-2}	1.6×10^{-2}	1.3×10^{-1}	7.2×10^{-2}	4.3×10^{-2}	2.7×10^{-2}	420
van Leer (II)	9	1.1×10^{-1}	4.8×10^{-2}	1.6×10^{-2}	5.9×10^{-3}	8.4×10^{-2}	4.1×10^{-2}	2.6×10^{-2}	1.6×10^{-2}	550
SAHS (Lax-Wendroff)	9	1.1×10^{-1}	4.9×10^{-2}	1.8×10^{-2}	5.5×10^{-3}	8.5×10^{-2}	4.2×10^{-2}	2.6×10^{-2}	1.6×10^{-2}	550
Yee - Explicit TVD	53	1.6×10^{-1}	1.1×10^{-1}	5.3×10^{-2}	2.0×10^{-2}	1.7×10^{-1}	8.7×10^{-2}	4.7×10^{-2}	3.0×10^{-2}	3400
Yee - Implicit TVD	58	1.8×10^{-1}	1.5×10^{-1}	1.3×10^{-1}	1.1×10^{-1}	2.2×10^{-1}	1.6×10^{-1}	1.2×10^{-1}	1.0×10^{-1}	3700
Phoenical LPE SHASTA	12	4.9×10^{-2}	2.0×10^{-2}	7.3×10^{-3}	4.3×10^{-3}	4.4×10^{-2}	3.5×10^{-2}	2.1×10^{-2}	1.3×10^{-2}	725

of them for use with high order implicit difference schemes. Other versions of each of the local adjustment schemes can be, or have been derived elsewhere, but the versions discussed here are all built around comparable high order schemes, and so may be treated as representative of each class of algorithms. As such, the relative performance of these particular algorithms gives an indication of the relative merits of the different approaches.

It is clear from the discussions on linear filtering and the CIP scheme, that using linear schemes introduces too much residual diffusion into the numerical solution. To overcome this, some form of non-linear calculations must be introduced into the numerical schemes. These non-linearities appear even in the constant velocity case. The approach of Smolarkiewicz leads to other problems, the two main ones are: the production of over-shooting oscillations, and the consistent deformation of any pulse onto something resembling a Gaussian distribution. The suggested improvements to this scheme do increase the height of the pulse, but unfortunately this can lead to a significant over estimation of the peak height in even simple problems, such as the semi-elliptical test case. The iteration of the antidiffusive step introduces two additional errors, namely the lag in the position and the large upstream skew of the pulse.

The schemes of van Leer (1973, 1974) use a smoothness monitor to decide whether diffusion should be added. While these schemes are fast in comparison to some of the others discussed in this chapter, both methods have a tendency to to convert sharp pulses (such as the Gaussian) to rather broad, squat pulses. If higher order schemes were used, this could be overcome but the necessary conditions for higher order schemes to be positive definite rapidly become very difficult. The Self-adjusting hybrid scheme can be made to be a general approach for explicit schemes although the extension to implicit schemes is also difficult.

The schemes based on the TVD assumption have been shown by Yee and others to be very accurate in modelling shocks and similar phenomenon. For modelling

advection in a prescribed velocity field, however, they require significantly more computation. Even for the case of constant coefficients, these schemes require more time than other schemes. If the velocity becomes a function of space and time, the TVD schemes will be even more expensive relative to the other schemes. This high cost in CPU time is not rewarded in increased accuracy in such problems.

Phoenical LPE SHASTA was the most accurate of all the schemes discussed in this chapter, although it does slightly skew the pulse. This can be seen in Table 4.1 and Table 4.2 where the RMS and Maximum errors and the relative error of the second moment are the best, or close to the best, for both types of initial condition, but the third order moments are not as good as those of other schemes. This scheme was also amongst the most efficient, along with the second of van Leer's schemes, Eq. (4.16), and the Self-adjusting hybrid scheme, Eq. (4.19).

The general approach of flux corrected transport has the advantage over the other approaches in that it is very easy to incorporate high order methods, especially if Zalesak's approach is used. Chapter Three described the success of the high order implicit schemes in providing very efficient schemes by either being exceptionally accurate or, reasonably accurate and unconditionally stable. The combination of these implicit schemes with one of the techniques for obtaining smooth results that are discussed in this chapter should lead to an accurate, robust positive definite scheme. Of the smoothing techniques suggested, the TVD and FCT approaches provide the most straightforward extension to implicit schemes. The problem with the TVD approach is that when applied to the problem of advection with specified velocity, it is very time consuming. This leaves FCT as the obvious choice for producing an efficient, implicit and high order positive definite scheme.

There are some problems with many of the FCT algorithms described by Boris and Book (1976a). The most obvious is that many of the algorithms resulted in numerical waveforms that tended to have a positive curvature along the trailing edge of a pulse and then a sudden change to negative curvature on the leading edge.

This is due mainly to the low order schemes being highly diffusive, and the "high order" schemes being only second order. Using third or higher order schemes with a certain amount of care, overcomes the problem of the sudden change in the sign of the curvature.

From Table 4.3 it can be seen that of all the schemes described, the Self-adjusting hybrid scheme, Eq. (4.19), provides the greatest improvement in accuracy with increased resolution, that is the ratio of the RMS error for $J = 100$ to that for $J = 800$ is no worse than for any other scheme. This is true for both initial conditions, and so it appears that this approach is very good at preserving order. That this scheme is better than first order follows directly from the derivation in Harten (1978).

It should be noted that the low order scheme is very important. This may be one of the reasons why the FCT algorithms based on the low order scheme SHASTA (Boris and Book, 1973) do not show the same improvement in accuracy with increased resolution, despite being more accurate overall. The low order scheme, Eq. (4.31), is even more diffusive than first order upwinding. This is best seen by comparing the amplitude response for Eq. (4.31), shown in Fig. 4.9(b), with that for first order upwinding in Fig. 2.1(b). A further problem with the low order scheme for Phoenical LPE SHASTA is that is it only stable for $c \leq 1/2$. The poor quality of the low order scheme will be seen in Chapter Five to have a significant impact on the overall accuracy of the algorithm.

One of the important features of a comparison between different techniques for producing positive definite algorithms is that all the schemes produce very smooth waveforms. There are no sharp signs of discontinuous changes in the derivatives. It was shown in Chapter Three that the high order implicit schemes advected such profiles (e.g. the Gaussian pulse) with exceptional accuracy. Thus, once the waveform has been slightly smoothed during the first few time-steps, it will be advected with very little change by the implicit schemes. This means that greater weight can be given to the high order solution which should help reduce the amount of residual

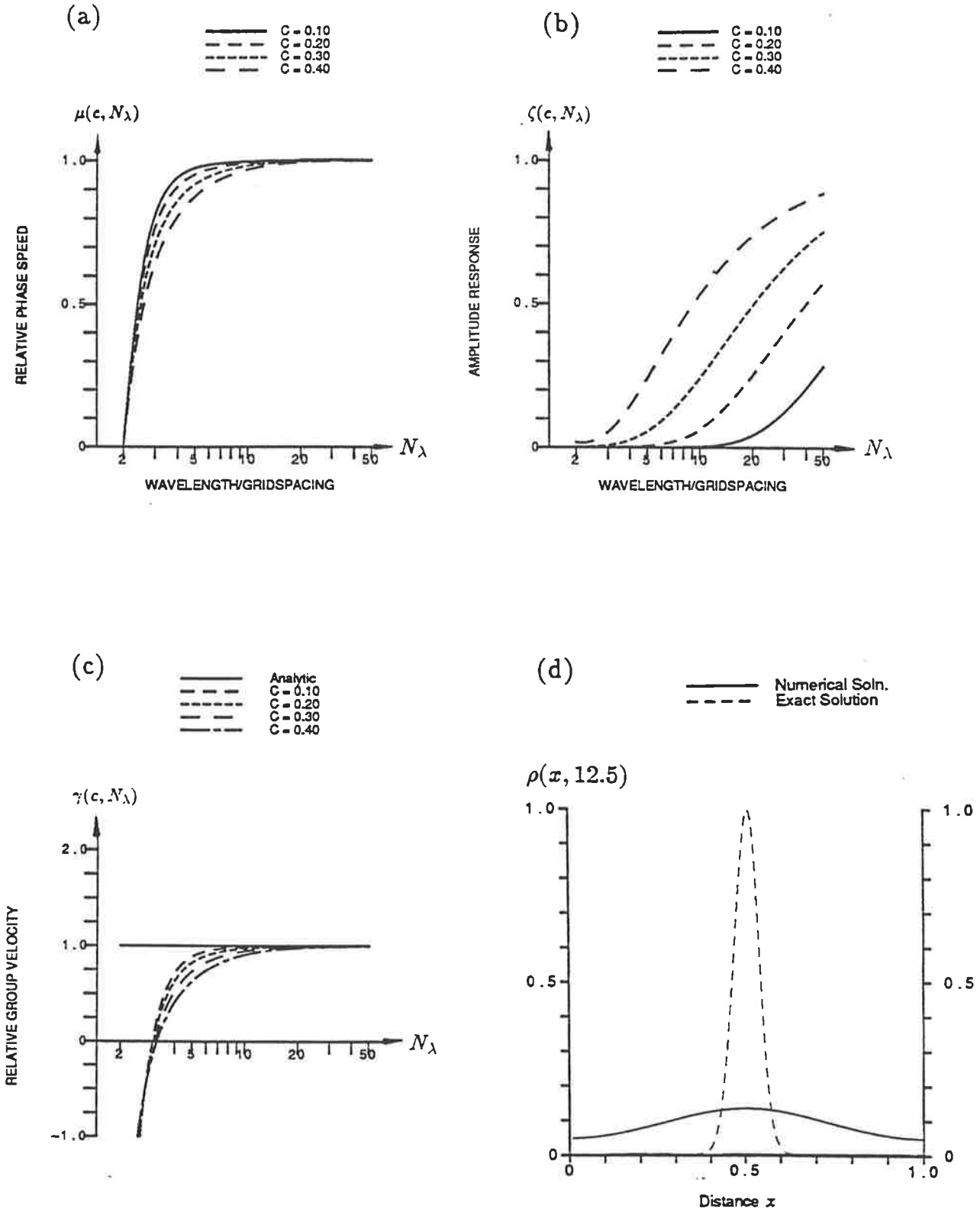

Figure 4.9: Illustration of the performance of the low order scheme for Phoenical LPE SHASTA, Eq. (4.31). The diagrams show (a) the relative phase speed, μ, (b) the amplitude response, ζ, (c) the relative group velocity, γ, and (d) the results of the numerical test case with a Gaussian pulse as the initial condition and cyclic boundary conditions.

diffusion in the numerical solution. The next chapter deals with how high order implicit schemes can be incorporated into an FCT algorithm.

Chapter 5

Smooth Implicit Finite Difference Solutions

This chapter deals with the use of implicit finite difference schemes to obtain approximate solutions to the advection equation, Eq. (2.1) that are free from numerical oscillations. In Chapter Three, it was shown how implicit schemes can be used to obtain results that are either of comparable accuracy to high order explicit schemes, but without the restriction on the time-step, or considerably more accurate than any other schemes. These schemes were also seen to perform to their maximum capabilities when there were no abrupt changes in the derivative of the initial condition.

Chapter Four gave a brief review of techniques for converting a variety of finite difference schemes that produced oscillatory results, into difference schemes that produced smooth results. Of all the techniques discussed, flux corrected transport was the most direct approach for converting arbitrary finite difference schemes into non-oscillatory schemes. It is also worth noting that all the smoothing techniques discussed in Chapter Four produce waveforms with no sudden changes in the first derivative. This is a direct consequence of their construction, since all these techniques relax back to low order, diffusive schemes in the presence of sudden changes in the derivatives of the solution. The change from sharp to smooth profiles occurs relatively quickly. Once this transformation has occurred, the high order implicit schemes can be expected to advect this smooth profile very accurately, without producing the large oscillations that appear when with sharp profiles are advected. A

less oscillatory numerical solution means greater weight can be given to the high order fluxes, reducing the impact of the low order fluxes and the flux limiter and thereby reducing any continued distortion of the peak by clipping.

A further advantage of using implicit schemes arises in the modelling of problems that exhibit a fine structure and include some diffusion. A fine grid is required in the vicinity of this structure if the detail is to be accurately modelled. If diffusion is present in the problem, then stability restrictions depend on the parameter $s = \alpha\Delta t/(\Delta x)^2$ (α being the coefficient of diffusion), in much the same manner as stability for advective processes depend on the Courant number. If an explicit scheme is used to model diffusion, then the stability restrictions within the fine grid can severely limit the time-step that can be taken. This can result in a considerable increase of CPU time for very little gain in accuracy, as noted by Morrow (1987). In such problems, it is more efficient to use implicit differencing of the diffusion terms, but if this done, there is very little overhead required to use implicit differencing for the advective terms. So for a variety of reasons, there is much to be gained by deriving smooth implicit difference schemes.

5.1 Boris and Book's REVFCT

Boris and Book (1976a) described some FCT algorithms which were implicit in certain aspects. Generally it was only the antidiffusion that was employed implicitly, with the advective part still being explicit as in the Implicit LPE SHASTA algorithm. This approach was also used by Patnaik et al., (1987) in their "barely implicit" FCT, a code that models everything, except advection, implicitly. This is different to the problem being examined here which is to develop an FCT algorithm where the advection is treated implicitly and possibly the antidiffusion as well.

Boris and Book observed that it is indeed preferable to incorporate the antidiffusion implicitly, for then the scheme can automatically be made to have "Zero Residual Damping"(ZRD) by keeping the differencing centred. Zero residual damping

means that if the numerical solution is diffused and then immediately antidiffused (bypassing the flux limiter) there is no damping of any components. An algorithm that meets all these conditions is REVFCT, as described in Boris and Book (1976a). This was derived by taking the Crank Nicolson form of the advection equation with diffusion simultaneously added and subtracted, that is

$$\begin{aligned} \rho_j^H &+ \left[c_{j+\frac{1}{2}}\left(\rho_{j+1}^H + \rho_j^H\right) - c_{j-\frac{1}{2}}\left(\rho_j^H + \rho_{j-1}^H\right)\right]/4 \\ &+ \nu_{j+\frac{1}{2}}\left(\rho_{j+1}^H - \rho_j^H\right) - \nu_{j-\frac{1}{2}}\left(\rho_j^H - \rho_{j-1}^H\right) \\ = \rho_j^n &- \left[c_{j+\frac{1}{2}}\left(\rho_{j+1}^n + \rho_j^n\right) - c_{j-\frac{1}{2}}\left(\rho_j^n + \rho_{j-1}^n\right)\right]/4 \\ &+ \nu_{j+\frac{1}{2}}\left(\rho_{j+1}^n - \rho_j^n\right) - \nu_{j-\frac{1}{2}}\left(\rho_j^n - \rho_{j-1}^n\right) \quad . \end{aligned} \tag{5.1}$$

It was noted that by putting $\nu = (2+c^2)/12$ the fourth order method Eq. (3.14) was obtained. Since the fluxes involving ν on the left hand side of the above equation correspond to "negative" diffusion and those on the right correspond to "positive" diffusion, Boris and Book reasoned that an appropriate low order scheme would be

$$\rho_j^L = \rho_j^H + \nu_{j+\frac{1}{2}}\left(\rho_{j+1}^H - \rho_j^H\right) - \nu_{j-\frac{1}{2}}\left(\rho_j^H - \rho_{j-1}^H\right) \tag{5.2}$$

giving the antidiffusive fluxes

$$\phi_{j+\frac{1}{2}} = \nu_{j+\frac{1}{2}}\left(\rho_{j+1}^H - \rho_j^H\right) \quad . \tag{5.3}$$

While this approach gives a very simple form for the antidiffusive fluxes, it does not yield a smooth difference scheme. This was partially acknowledged by Boris and Book [1],

> "Because both the transport and the diffusion are implicit, the transport causes numerical precursors which cross the mesh in one cycle. These extend far beyond the reach of the relatively local flux limiter and hence cannot be controlled by it."

[1] page 114, Boris and Book (1976a)

and [2]

> "This objection to implicit treatments of the convective term would seem to be quite general even though only a specific case was tested"

One of the aims of this chapter is to show that such claims are incorrect, and that it was just their particular approach which caused problems. It will be shown that although the flux limiter only directly uses information from a few gridvolumes, the fluxes for implicit schemes are themselves defined recursively. Either the fluxes are calculated by a marching scheme or are calculated explicitly from expressions that involve more than one value of the high order solution. This means a certain amount of information is carried downstream from every point, or equivalently, any point obtains some information from all the upstream points. So despite the limiter being derived by considering only local fluxes, the fluxes themselves contain enough information from upstream points to allow the flux limiter to cope with any numerical precursors. Another reason for not believing that the flux limiter is incapable of handling numerical precursors is that if this were the case then the same problem should appear with high order explicit schemes. Consider fifth order upwinding, Eq. (2.39). A sharp gradient at any gridpoint will cause this scheme to produce an oscillation up to three gridpoints away, since ρ_{j-3}^n is used in the calculation of ρ_j^{n+1}. As higher and higher upwind schemes are used then these oscillations will occur over a wider number of points. It is easily verified that such behaviour does not occur, and it appears that the flux limiter is capable of controlling these fluctuations that occur over many gridpoints. So it is not unreasonable to believe that there is a possibility that the flux limiter is capable of controlling the fluctuations in the limiting case, where the oscillations spread throughout the domain after one time-step, as occur in higher order schemes.

If it is not the limiter that is at fault then there must be a problem with some

[2] page 118, Boris and Book (1976b)

other part of the REVFCT algorithm. One such problem is that the low order scheme does not guarantee smooth results, which violates the basic assumptions of any FCT algorithm. This can be verified by using $\rho_j^{n+1} = \rho_j^L$ for three time-steps with a square wave as the initial condition. The minimum value for the numerical solution in such a test is -5.3×10^{-4} and the sum of the negative values is -1.1×10^{-3}. After a few more time-steps, enough diffusion is added to the solution (recall that only the low order solution is being considered here) for the oscillations to disappear. In a non-linear problem, however, these oscillations may remain, causing the problems noted by Boris and Book. This simple example also contradicts the assertion on page 114 of Boris and Book (1976a) that

> "... REVFCT is great for passive advection...".

REVFCT is quite clearly not ideal for linear advection problems as it is not a smooth difference operator, and so it is not surprising that it is unsuitable for non-linear problems. Part of the reason for the appearance of these oscillations is connected with the discussion on linear filters in Chapter Four. Equation (5.2) is similar to the second order approximation to the diffusion equation Eq. (4.2) with $s = \nu$. This type of process, applied to the high order solution given by Eq. (5.1) has been shown not to necessarily provide oscillation free results. The next two sections describe corrections to the errors in REVFCT and demonstrate that the flux limiter can be used with implicit schemes.

5.2 Conditions for Implicit Equations to be Positive Definite

If implicit schemes are to be used, then it is necessary to know when smooth results can be guaranteed. For a linear scheme, this is equivalent to it being positive definite, since adding or subtracting a constant from the initial condition results in the new numerical solution changing by the same constant. Hence, if a linear scheme produces over-shoots such as those of Smolarkiewicz's scheme, then the same over-

shoots should appear as a spurious sign change with a modified initial condition. This does not appear with Smolarkiewicz's scheme as it is non-linear. Sufficient conditions under which the solution of the set of linear algebraic equations

$$\sum_k a_{jk}\rho_k^{n+1} = \sum_k b_{jk}\rho_k^n \tag{5.4}$$

will produce positive values, given $\rho_j^n \geq 0$ for all j, are :

1. b_{jk}, a_{jj} are non-negative, and
2. $a_{jk} \leq 0$ for all $k \neq j$ and $\sum_{k,k\neq j} |a_{jk}| < a_{jj}$.

This can be shown as follows:

Suppose that b_{jk} , $a_{jj} > 0$, $-a_{jj} < a_{jk} \leq 0$ for $k \neq j$, and $\rho_j^n > 0$ for all j. Eq. (5.4) can be rewritten in matrix form

$$(\mathbf{I} - \mathbf{A})\,\underset{\sim}{\rho}^{n+1} = \underset{\sim}{d} \tag{5.5}$$

where the elements of $\mathbf{A}$ are given by

$$a_{ji} = \begin{cases} -a_{ji}/a_{jj} & j \neq i \\ 0 & j = i \end{cases} \tag{5.6}$$

$$d_j = \left(\sum_k b_{jk}\rho_k^n\right)/a_{jj} \quad . \tag{5.7}$$

Under such conditions, each of the elements of $\mathbf{A}$ and $\underset{\sim}{d}$ will be non-negative. So Eq. (5.5) has solution

$$\underset{\sim}{\rho}^{n+1} = \sum_{k=0}^{\infty} \mathbf{A}^k \underset{\sim}{d} \quad . \tag{5.8}$$

Since each element of $\mathbf{A}$ is non-negative, then so is each element of $\mathbf{A}^k$, and by the non-negativity of d_j, $\rho_j^{n+1}$is non-negative for all j. The convergence of the series $\sum_{k=0}^{\infty} \mathbf{A}^k$ follows directly from

$$\max \sum_{k=1}^{J} |a_{jk}| = \|\mathbf{A}\|_\infty < 1 \quad , \tag{5.9}$$

which is ensured by the construction of $\mathbf{A}$ and the second assumption on the diagonal dominance of $\mathbf{A}$

It has now been shown that if an implicit difference equation is written in the form

$$(\mathbf{I}-\mathbf{A})\,\underset{\sim}{\varrho}^{n+1} = \mathbf{B}\underset{\sim}{\varrho}^{n} \tag{5.10}$$

and

1. a_{ij}, b_{ij} are non-negative for all i, j

2. $\|\mathbf{A}\|_\infty < 1$,

then smooth results are guaranteed. Less restrictive conditions could be found by only requiring that all elements of $(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$ are nonnegative. Such conditions are unlikely to be of much use in developing conditions for the coefficients of a finite difference equation to be positive definite since, in general, the relationship between the elements of $(\mathbf{I}-\mathbf{A})^{-1}$ and the coefficients in the difference equation is very complicated. Applying the non-linear constraint

$$\sum_k \left[(\mathbf{I}-\mathbf{A})^{-1}\right]_{ik} b_{kj} \geq 0 \quad \textit{for all } i \tag{5.11}$$

to the design of a finite difference scheme would be too cumbersome. The advantage of the two restrictions given above is that they are simply related to constraints on the coefficients of the finite difference equation, and as such, may be directly incorporated in the construction of any smooth finite difference scheme.

It should be noted that these conditions have been shown to be sufficient for an implicit scheme to be positive definite. The necessary conditions under which Eq. (5.4) is positive definite will not be too dissimilar. In practice, there are more likely to be many negative elements of $\mathbf{A}$ than just one or two, since each row of $\mathbf{A}$ comes from the same difference equation. So that if one element, a_{ij} for example, is negative then there are likely to be several other negative elements nearby. (This need not necessarily be the case as near stagnation points, or other special locations, it may be possible to have just one or two negative values, which will not propagate into the rest of the solution domain. In general however, any

ripples will usually grow, within bounds, as they are advected about the domain). If there is a systematic pattern of negative elements in the matrices $\mathbf{A}$, however, then negative values may also appear in the powers of $\mathbf{A}$, and hence in $(\mathbf{I}-\mathbf{A})^{-1}$. If this were to occur, it is possible to admit negative values of ρ_j^{n+1}. A similar argument holds for the matrix $\mathbf{B}$, for if any elements, b_{ij}, were negative then there exist vectors $\underset{\sim}{x}$ such that $\mathbf{B}\underset{\sim}{x}$ contains negative values, and so even if all the powers of $\mathbf{A}$ contained only positive values, it would still be possible for negative values of ρ_j^{n+1} to appear.

It is important to distinguish between a matrix being positive definite and a finite difference scheme being positive definite. For an arbitrary matrix $\mathbf{K}$ to be positive definite, then $\underset{\sim}{x}^T\,\mathbf{K}\,\underset{\sim}{x}$ must be positive for all $\underset{\sim}{x}$, but does not imply anything about the signs of individual elements of $\mathbf{K}\underset{\sim}{x}$. If a finite difference scheme is positive definite then for all $\underset{\sim}{x}$ such that $x_i \geq 0$ for all i, the finite difference operator $\mathcal{L}$ must be such that each element of $\mathcal{L}\underset{\sim}{x}$ is non-negative. In this respect, referring to finite difference schemes as positive definite is somewhat misleading, since for a linear finite difference scheme, the corresponding matrices may be positive definite, without the finite difference scheme being positive definite. This can be seen in the following example.

The behaviour of a linear finite difference scheme is dictated by the nature of the matrix $(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$. If all the elements of $\mathbf{B}$ are positive then the finite difference scheme will be positive definite if each element of $(\mathbf{I}-\mathbf{A})^{-1}\underset{\sim}{x}$ is non-negative for all $\underset{\sim}{x}$ such that $x_i \geq 0$ for all i. The following examples show that neither $(\mathbf{I}-\mathbf{A})^{-1}$ or $\mathbf{A}$ being positive definite will ensure a positive definite finite difference scheme. Consider the matrix

$$\mathbf{K} = \begin{bmatrix} \epsilon & 0 \\ -\epsilon/2 & \epsilon \end{bmatrix} , \tag{5.12}$$

for some positive number ϵ. In this case

$$\underset{\sim}{x}^T\mathbf{K}\underset{\sim}{x} = \epsilon(x_1^2 + x_2^2 - \frac{1}{2}x_1x_2) \tag{5.13}$$

which is positive since one of x_1^2 or x_2^2 is greater than or equal to $|x_1 x_2|$. It is also clear that if $x_1 > 2x_2$ then $[\mathbf{K}\underline{x}]_2$ is negative. Thus, the fact that any matrix, such as $(\mathbf{I}-\mathbf{A})^{-1}$ is positive definite, is insufficient to ensure that the corresponding finite difference scheme is positive definite. Furthermore, the fact that the matrix $\mathbf{A}$ is positive definite is insufficient to ensure that all elements $(\mathbf{I}-\mathbf{A})^{-1}\underline{x}$ are non-negative when all the elements of $\underline{x}$ are non-negative. Consider the case where

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix} \quad . \tag{5.14}$$

Now

$$\underline{x}^T \mathbf{A} \underline{x} = 2(x_1^2 + x_2^2 + \frac{1}{2} x_1 x_2) \tag{5.15}$$

which will be positive since $x_1^2 + x_2^2 > |x_1 x_2|$, however, the first element of $(\mathbf{I}-\mathbf{A})^{-1}\underline{x}$ is $-x_1$, and so the corresponding finite difference scheme will not be positive definite. It is also worth noting that the second constraint ensures that the matrix $(\mathbf{I}-\mathbf{A})$ is diagonally dominant, and so inversion of this matrix by elimination schemes, such as the Thomas Algorithm will be stable.

Having determined sufficient conditions for implicit finite difference schemes to produce smooth results it is now possible to correct REVFCT. This is will be done in the following section

5.3 Corrections to REVFCT

The first point to note about calculating high and low order solutions for implicit schemes in the manner of Eq. (5.2), is that it makes a difference whether the low order solution is calculated from the high order solution, as in REVFCT, or vice versa. This is best seen by writing Eq. (5.1) and Eq. (5.2) in the form

$$\begin{aligned} (\mathbf{I} + \mathbf{A} + \mathbf{D})\, \underline{\rho}^H &= \underline{d} \\ \underline{\rho}^L &= (\mathbf{I} + \mathbf{D})\, \underline{\rho}^H \end{aligned} \tag{5.16}$$

where $\mathbf{A}$ contains the coefficients involving c and $\mathbf{D}$ contains the coefficients involving ν. Eq. (5.16) is equivalent to solving

$$\begin{aligned} \left[\mathbf{I}+\mathbf{A}\left(\mathbf{I}+\mathbf{D}\right)^{-1}\right]\underset{\sim}{\rho}^{L} &= \underset{\sim}{d} \\ \underset{\sim}{\rho}^{H} &= \left(\mathbf{I}+\mathbf{D}\right)^{-1}\underset{\sim}{\rho}^{L} \quad . \end{aligned} \tag{5.17}$$

This will generally have different solutions to

$$\begin{aligned} \left(\mathbf{I}+\mathbf{A}\right)\underset{\sim}{\rho}^{L} &= \underset{\sim}{d} \\ \underset{\sim}{\rho}^{H} &= \left(\mathbf{I}-\mathbf{D}\right)\underset{\sim}{\rho}^{L} \end{aligned} \tag{5.18}$$

which correspond to calculating the low order solution first and then calculating the high order solution. The latter approach also changes the equation for the antidiffusive fluxes, with ρ_j^L replacing ρ_j^H in Eq. (5.3).

One of the errors in REVFCT results from this. The low order scheme was intended to be the Crank Nicolson scheme with diffusion added at time-level t^n and it was argued by Boris and Book that this diffusion was enough to give smooth results. Even if this were so, by using Eq. (5.2) to calculate the low order solution (and hence the antidiffusive fluxes), the calculated low order solution need no longer be smooth.

Another error is that the additional diffusion is not strong enough to guarantee smooth results. That the low order scheme

$$\begin{aligned} \rho_j^L &+ \left[c_{j+\frac{1}{2}}\left(\rho_{j+1}^L+\rho_j^L\right)-c_{j-\frac{1}{2}}\left(\rho_j^L+\rho_{j-1}^L\right)\right]/4 \\ = \rho_j^n &- \left[c_{j+\frac{1}{2}}\left(\rho_{j+1}^n+\rho_j^n\right)-c_{j-\frac{1}{2}}\left(\rho_j^n+\rho_{j-1}^n\right)\right]/4 \\ &+ \nu_{j+\frac{1}{2}}\left(\rho_{j+1}^n+\rho_j^n\right)-\nu_{j-\frac{1}{2}}\left(\rho_j^n-\rho_{j-1}^n\right) \quad . \end{aligned} \tag{5.19}$$

is not positive definite as can be verified by using Eq. (5.19) to advect a square wave for three time-steps with $c_{j+\frac{1}{2}} = 0.4$ for all j. It must be stressed that although Eq. (5.19) was intended to be the low order scheme for REVFCT, the method of calculating the low order solution, using Eq. (5.2) meant that a slightly different scheme was actually being used, as discussed in the previous paragraphs. Using the

scheme Eq. (5.19) yields a numerical solution with a minimum value of -5.70×10^{-4} and the sum of the negative values is -5.72×10^{-4}. It can be directly compared against the results from using low order solution Eq. (5.2) which were -5.3×10^{-3} and -1.1×10^{-3}, respectively. Such a comparison shows that calculating the low order solution from the high order solution (as in REVFCT) gives a less diffuse solution than if the low order solution is calculated directly (as above). The result is consistent with the statement that calculations of the type Eq. (5.2) are not necessarily positive definite, even if calculations similar to Eq. (5.19) are positive definite.

It is also interesting to note that Eq. (5.19) violates the constraints developed in the previous section, as the coefficient of ρ_{j+1}^{n+1} is positive. While this does not necessarily mean that the scheme is not positive definite, it demonstrates what happens when one coefficient of the difference equation consistently violates the constraint . In this case, every element of the super-diagonal of the matrix **A** is negative, since the coefficient of ρ_{j+1}^{n+1} is positive for all j.

So far it has been shown that there are difficulties with REVFCT other than the alleged inability of the flux limiter to control the numerical precursors of the implicit solution. It remains to be shown that when these new difficulties are overcome, the scheme becomes positive definite. While the non-linear test cases used by Boris and Book (1976a) are not reproduced here, as the support software was not available, it has been shown that REVFCT is not ideal for simple advection problems. It is possible that the errors discussed above are the cause of the difficulties in the non-linear models, rather than any problems with the flux limiter. It will now be shown how the REVFCT algorithm may be converted into a positive definite scheme.

Assuming for the moment that the only problem with REVFCT is that the low order solution is not positive definite then, using the two constraints mentioned above, it should be possible to correct the algorithm. In order to show that the errors in REVFCT are due to the calculation of the low order solution, and hence the

antidiffusive fluxes, an algorithm will be developed which closely follows REVFCT, to the point of only calculating the antidiffusive fluxes, rather than both the high and low order solutions. This does mean, however, that the antidiffusive fluxes must be calculated from the low order solution and not the high order solution.

Boris and Book (1976a) derived REVFCT by noting that Eq. (5.1) is the Crank Nicolson scheme, Eq. (3.8), with diffusion added at time-level t^n and subtracted at time-level t^{n+1}. One possible method of correcting REVFCT is to calculate the necessary diffusion that should be added to the Crank Nicolson scheme in order to obtain a positive definite low order solution. There are several ways of achieving this, depending on which coefficients need altering in order to satisfy the appropriate constraints. It should also be noted that the diffusion is being used as a device to alter the coefficients and not to accurately model any physical process. For this reason the order of the approximation is not of vital importance. The approximation should still represent diffusion in some sense, however, as diffusion is the physical equivalent of smoothing.

Using the Crank Nicolson scheme Eq. (3.8), it can be seen that the coefficients of ρ_{j+1}^{n+1} and ρ_{j+1}^{n} are both of the wrong sign if the constraints developed in the previous section are to be satisfied. Hence, diffusion must be added at both time-levels, and an appropriate approximation to the diffusive term is

$$\begin{aligned} &\frac{\partial}{\partial x}\left(\alpha\frac{\partial \rho}{\partial x}\right) \approx \qquad (5.20)\\ &\frac{1}{(\Delta x)^2}\left[(\theta\alpha)_{j+\frac{1}{2}}\left(\rho_{j+1}^{n+1}-\rho_{j}^{n+1}\right)-(\theta\alpha)_{j-\frac{1}{2}}\left(\rho_{j}^{n+1}-\rho_{j-1}^{n+1}\right)\right]\\ &+\ \frac{1}{(\Delta x)^2}\left[(\alpha-\theta\alpha)_{j+\frac{1}{2}}\left(\rho_{j+1}^{n}-\rho_{j}^{n}\right)-(\alpha-\theta\alpha)_{j-\frac{1}{2}}\left(\rho_{j}^{n}-\rho_{j-1}^{n}\right)\right] \quad , \end{aligned}$$

for some arbitrary weight function, θ, defined on the grid, with $0 \leq \theta \leq 1$ everywhere. For $\theta \neq 1/2$ this approximation is not centred at $(x_j,\ t^{n+\frac{1}{2}})$ and so will lead to errors of $O\{\Delta t\}$. This does not matter, since at the moment, the main task is to obtain a positive definite implicit scheme, rather than a highly accurate scheme. The inequalitites that guarantee smooth results require that the coefficients of ρ_{j-1}^{n+1}

and ρ_{j+1}^{n+1} must be negative, giving the two constraints

$$\begin{aligned} \frac{c_{j+\frac{1}{2}}}{4} - (\theta s)_{j+\frac{1}{2}} &\leq 0 \\ -\frac{c_{j+\frac{1}{2}}}{4} - (s - \theta s)_{j+\frac{1}{2}} &\leq 0 \end{aligned} \tag{5.21}$$

where $s_{j+\frac{1}{2}} = \alpha_{j+\frac{1}{2}} \Delta t/(\Delta x)^2$. These constraints imply that

$$\frac{c_{j+\frac{1}{2}}}{4\theta_{j+\frac{1}{2}}} \leq s_{j+\frac{1}{2}} \leq \frac{c_{j+\frac{1}{2}}}{4(1-\theta_{j+\frac{1}{2}})} \tag{5.22}$$

which can only be true if $1 - \theta_{j+\frac{1}{2}} \leq \theta_{j+\frac{1}{2}}$, or

$$\frac{1}{2} \leq \theta_{j+\frac{1}{2}} \leq 1 \quad . \tag{5.23}$$

In order to ensure that the diffusion in the low order scheme is minimized (which in turns helps minimize the residual diffusion), it is appropriate to choose

$$s_{j+\frac{1}{2}} = \frac{c_{j+\frac{1}{2}}}{4\theta_{j+\frac{1}{2}}}. \tag{5.24}$$

For the coefficient of ρ_{j+1}^{n} to be positive the additional constraint

$$-\frac{c_{j+\frac{1}{2}}}{2} + \frac{c_{j+\frac{1}{2}}}{4\theta_{j+\frac{1}{2}}} \geq 0 \tag{5.25}$$

must hold. This implies that $\theta_{j+\frac{1}{2}} \leq \frac{1}{2}$, and using Eq. (5.23) it is clear that the optimal choice of parameters is

$$\theta_{j+\frac{1}{2}} = \frac{1}{2} \quad \text{and} \quad s_{j+\frac{1}{2}} = \frac{c_{j+\frac{1}{2}}}{2} \leq 1 \quad , \tag{5.26}$$

everywhere. This gives the difference equation

$$-\frac{c_{j-\frac{1}{2}}}{2}\rho_{j-1}^{n+1} + \left(1 + \frac{c_{j+\frac{1}{2}}}{2}\right)\rho_j^{n+1} = \frac{c_{j-\frac{1}{2}}}{2}\rho_{j-1}^{n} + \left(1 - \frac{c_{j+\frac{1}{2}}}{2}\right)\rho_j^{n} \tag{5.27}$$

This scheme produces smooth results if $0 \leq c_{j+\frac{1}{2}} \leq 2$ for all j. To retain consistency with REVFCT, the antidiffusive fluxes are calculated explicitly using

$$\phi_{j+\frac{1}{2}} = \frac{c_{j+\frac{1}{2}}}{4}\left(\rho_{j+1}^{L} - \rho_j^{L} + \rho_{j+1}^{n} - \rho_j^{n}\right) \quad . \tag{5.28}$$

The results for this scheme are shown in Fig. 5.1(c,d) The numerical solution in the figures is now shown by crosses in the diagrams. The reason for this change

is that, in previous chapters, it was the distortion of the pulse that was being examined, and while this distortion was due to linear damping or dispersion it was best illustrated by depicting the numerical solution as a line. This was particularly true of the dispersive numerical solutions, where a continuous representation gave a clearer view of the wavelength of the oscillations. In this chapter, however, most of the distortion will be due to clipping, which is best illustrated by using a discrete representation of the numerical solution.

This scheme is, however, quite diffusive. This is mainly due to the heavy diffusion in the low order scheme, which is apparent in Fig. 5.1(b) and from the modified equation

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u\Delta x}{2}\frac{\partial^2 \rho}{\partial x^2} + O\{(\Delta x)^3\} \quad . \tag{5.29}$$

The error measures for this scheme are given under CN2POS in Table 5.1 and Table 5.2. The distortion with the sudden change in curvature near the peak is due to fluxes on the upstream side of the pulse being more heavily limited than on the downstream side. This is in turn due to the slow propagation of the short wavelength Fourier components by the high order scheme introducing oscillations that trail a region of sharp gradients.

An alternative correction to Boris and Book's REVFCT is to use the high order scheme of REVFCT (the fourth order CTCS scheme) and determine how much diffusion should be added to make it positive definite, since the amount added originally was insufficient. With diffusion incorporated, the fourth order difference equation becomes

$$\begin{aligned} \rho_j^{n+1} &+ \frac{c_{j+\frac{1}{2}}}{4}\left(\rho_{j+1}^{n+1} + \rho_j^{n+1}\right) - \frac{c_{j-\frac{1}{2}}}{4}\left(\rho_j^{n+1} + \rho_{j-1}^{n+1}\right) \\ &+ \left(\frac{1}{6} + \frac{c_{j+\frac{1}{2}}^{\,2}}{12} - (\theta s)_{j+\frac{1}{2}}\right)\left(\rho_{j+1}^{n+1} - \rho_j^{n+1}\right) \\ &- \left(\frac{1}{6} + \frac{c_{j-\frac{1}{2}}^{\,2}}{12} - (\theta s)_{j-\frac{1}{2}}\right)\left(\rho_j^{n+1} - \rho_{j-1}^{n+1}\right) \\ = \rho_j^n &- \frac{c_{j+\frac{1}{2}}}{4}\left(\rho_{j+1}^n + \rho_j^n\right) + \frac{c_{j-\frac{1}{2}}}{4}\left(\rho_j^n + \rho_{j-1}^n\right) \end{aligned} \tag{5.30}$$

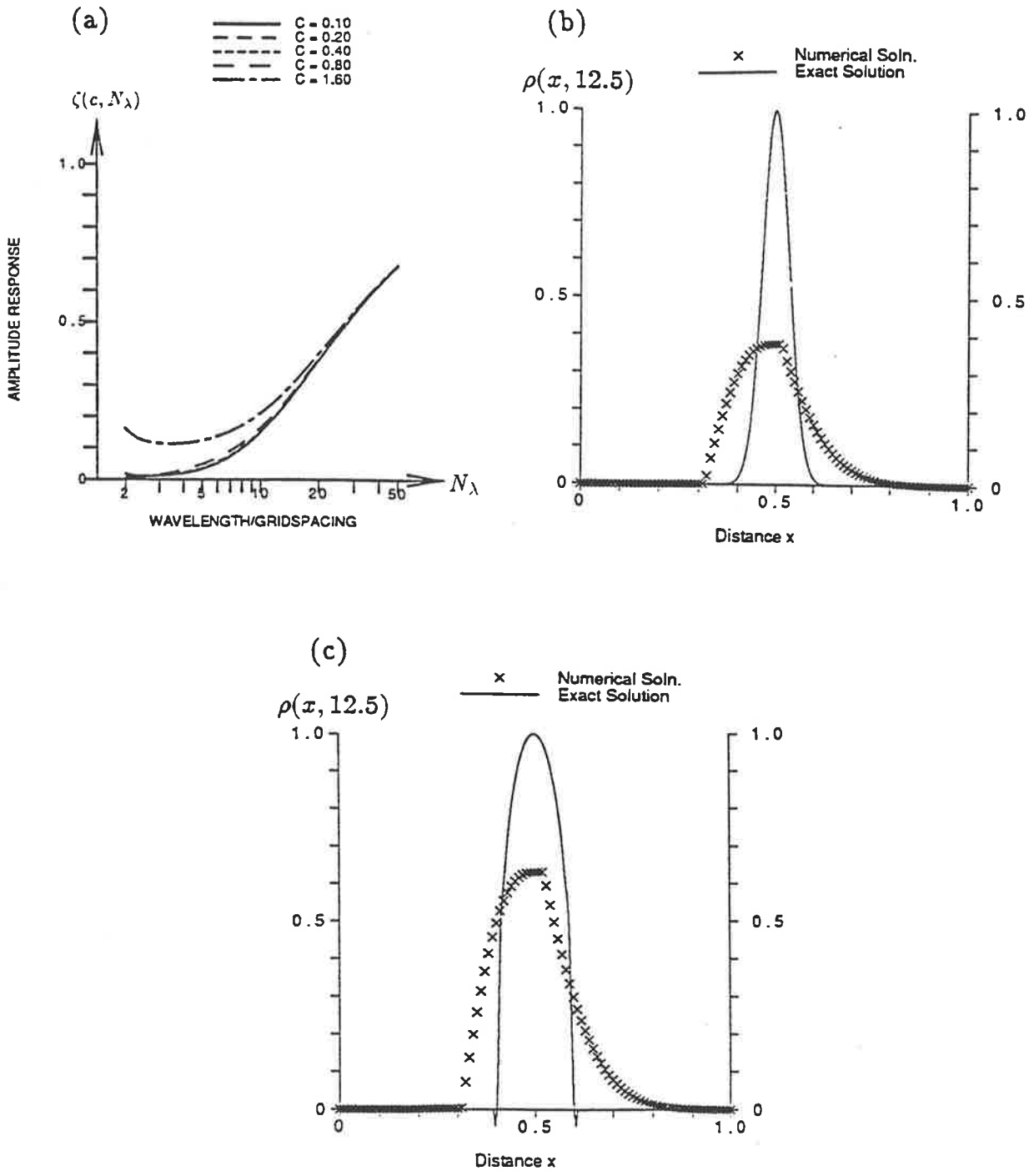

Figure 5.1: Illustration of the performance of CN2POS using Eq. (5.27) and Eq. (5.28) as the low order schemes and antidiffusive fluxes. The diagrams show (a) the amplitude response ζ of the low order scheme, and the results of numerical test cases with cyclic boundary conditions and initial conditions that are (b) a Gaussian pulse and (c) a semi-ellipse.

Table 5.1:Error measures for some positive definite difference schemes for the test problem with a Gaussian pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
REVFCT	0.049	0.23	3.8×10^{-2}	0.14	3.0×10^{-4}	-0.5	0.77
CN2POS	0.159	0.49	-8.6×10^{-2}	2.95	4.5×10^{-4}	1.5	0.63
REVPOS	0.182	0.55	-4.7×10^{-2}	3.38	2.2×10^{-3}	-0.5	0.45
CN4FCT	0.028	0.13	9.3×10^{-3}	0.004	-8.0×10^{-5}	0.5	0.87
CN2FCT	0.204	0.56	-6.9×10^{-2}	4.11	1.4×10^{-3}	0.5	0.44
NS4FCT	0.019	0.10	-3.1×10^{-4}	0.038	-1.7×10^{-4}	0.5	0.90
NS5FCT	0.017	0.09	6.9×10^{-4}	0.017	-1.5×10^{-4}	0.5	0.91
Fifth order upwinding + FCT	0.034	0.17	-1.9×10^{-3}	0.058	9.2×10^{-5}	-0.3	0.83
van Leer (II)	0.109	0.47	-2.3×10^{-4}	1.64	-1.7×10^{-5}	-0.2	0.53
Phoenical LPE SHASTA	0.049	0.24	5.2×10^{-3}	0.27	-1.4×10^{-4}	0.5	0.76
Implicit LPE FCT	0.048	0.23	6.8×10^{-3}	0.25	2.2×10^{-4}	-0.5	0.78

Table 5.2: Error measures for some positive definite difference schemes for the test problem with a semi-elliptical pulse as the initial condition and cyclic boundary conditions. The Peak Shift is given as a fraction of the grid spacing, Δx by which the numerical peak leads the true peak. All other measures are given in absolute terms.

Scheme	RMS. Error	Maximum |Error|	Rel. Error of 1^{st} Moment	Rel. Error of 2^{nd} Moment	Error of 3^{rd} Moment	Peak Shift	Peak Height
REVFCT	0.038	0.18	-5.1×10^{-3}	0.06	-1.7×10^{-5}	0.0	0.97
CN2POS	0.174	0.49	-8.6×10^{-2}	2.94	4.6×10^{-4}	1.5	0.63
CN4POS	0.222	0.48	-3.4×10^{-2}	1.54	1.8×10^{-3}	0.5	0.73
CN4FCT	0.039	0.18	-1.2×10^{-2}	0.04	-6.5×10^{-4}	0.5	0.98
CN2FCT	0.146	0.50	-4.7×10^{-2}	1.99	-2.8×10^{-5}	1.5	0.72
NS4FCT	0.030	0.16	-4.1×10^{-3}	0.02	5.3×10^{-4}	-0.5	0.99
NS5FCT	0.024	0.14	-3.3×10^{-3}	0.10	-1.8×10^{-3}	1.5	0.99
Fifth order upwinding + FCT	0.039	0.18	-5.4×10^{-4}	0.15	1.8×10^{-3}	-1.5	0.98
van Leer (II)	0.084	0.35	-2.2×10^{-3}	0.72	2.5×10^{-4}	-0.3	0.83
Phoenical LPE SHASTA	0.044	0.22	-3.9×10^{-1}	0.37	-3.3×10^{-3}	2.5	0.98
Implicit LPE SHASTA	0.046	0.22	-3.6×10^{-4}	0.13	6.3×10^{-4}	-0.5	0.99

$$+ \left(\frac{1}{6} + \frac{c_{j+\frac{1}{2}}{}^2}{12} + (s - \theta s)_{j+\frac{1}{2}}\right)\left(\rho^n_{j+1} - \rho^n_j\right)$$
$$- \left(\frac{1}{6} + \frac{c_{j-\frac{1}{2}}{}^2}{12} + (s - \theta s)_{j-\frac{1}{2}}\right)\left(\rho^n_j - \rho^n_{j-1}\right) \quad ,$$

where, again, θ is some weight function defined on the grid and contained within the interval [0, 1]. Firstly consider the case where $0 < c_{j+\frac{1}{2}} \leq 1$, so the system of equations is to be solved by the Thomas algorithm. The coefficient of ρ^{n+1}_{j+1} will be negative provided that

$$\frac{(1 + c_{j+\frac{1}{2}})(2 + c_{j+\frac{1}{2}})}{12} - (\theta s)_{j+\frac{1}{2}} \leq 0 \tag{5.31}$$

and so the diffusion of the low order scheme is minimized if

$$s_{j+\frac{1}{2}} = \frac{(1 + c_{j+\frac{1}{2}})(2 + c_{j+\frac{1}{2}})}{12} \quad , \tag{5.32}$$

and $\theta_{j+\frac{1}{2}} = 1$. The resulting low order scheme is

$$\begin{aligned} -\frac{c_{j-\frac{1}{2}}}{2}\rho^L_{j-1} \;+\; & \left(1 + \frac{c_{j+\frac{1}{2}}}{2}\right)\rho^L_j \\ = \rho^n_j \;-\; & \frac{c_{j+\frac{1}{2}}}{4}\left(\rho^n_{j+1} + \rho^n_j\right) + \frac{c_{j-\frac{1}{2}}}{4}\left(\rho^n_j + \rho^n_{j-1}\right) \\ +\; & \left(\frac{1}{6} + \frac{c^2_{j+\frac{1}{2}}}{12}\right)\left(\rho^n_{j+1} - \rho^n_j\right) - \left(\frac{1}{6} + \frac{c^2_{j-\frac{1}{2}}}{12}\right)\left(\rho^n_j - \rho^n_{j-1}\right) \quad , \end{aligned} \tag{5.33}$$

and the corresponding antidiffusive fluxes are

$$\phi_{j+\frac{1}{2}} = \frac{1}{12}\left(c_{j+\frac{1}{2}} + 2\right)\left(c_{j+\frac{1}{2}} + 1\right)\left(\rho^L_{j+1} + \rho^L_j\right) \quad . \tag{5.34}$$

This scheme will be denoted REVPOS in the following diagrams and tables.

The diffusion in the low order part of REVFCT was $(c_{j+\frac{1}{2}})^2/12$ which is clearly less than that required here. This is consistent with the deduction that the low order fluxes of REVFCT do not guarantee smooth results. Unfortunately, the heavy diffusion required to obtain the low order solution significantly degrades the overall results as shown in Fig. 5.2(b,c).

The degradation relative to REVFCT can also be seen by comparing the modified equivalent equations for the two low order schemes. The low order scheme for

REVFCT, calculated directly using Eq. (5.19), rather than from Eq. (5.2), has the modified equivalent equation

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u\Delta x}{2}\left(\frac{c}{6} + \frac{1}{3c}\right)\frac{\partial^2 \rho}{\partial x^2} + O\{(\Delta x)^2\} \tag{5.35}$$

whereas Eq. (5.33) has a modified equivalent equation

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u\Delta x}{2}\left(\frac{1}{2} + \frac{c}{6} + \frac{1}{3c}\right)\frac{\partial^2 \rho}{\partial x^2} + O\{(\Delta x)^2\} \quad . \tag{5.36}$$

Both of these low order schemes cause considerable damping of the solution, due to the terms involving c^{-1} (as $0 < c \leq 1$) and so neither is particularly suitable as a low order scheme. At the moment this does not matter, as the present aim is to show that with appropriate corrections, REVFCT can be made positive definite in order to show that the numerical precursors are not the only explanation of the failure of REVFCT in non-linear problems. The development of an accurate scheme would have been a bonus.

Another option for correcting the REVFCT algorithm is to add the diffusion to the marching form of Eq. (5.1). With the equation in this form, diffusion again must be added at time t^n and t^{n+1} since the coefficients of ρ_{j-1}^{n+1}, ρ_j^{n+1}, ρ_j^n and ρ_{j+1}^n all have the incorrect sign. There are two possible approximations that may be used, either Eq. (5.20) or

$$\frac{\partial}{\partial x}\left(\alpha\frac{\partial \rho}{\partial x}\right) \approx \alpha_{j+\frac{1}{2}}\left(\frac{\rho_{j+1}^n - \rho_j^n}{\Delta x^2}\right) - \alpha_{j-\frac{1}{2}}\left(\frac{\rho_j^{n+1} - \rho_{j-1}^{n+1}}{\Delta x^2}\right) \quad . \tag{5.37}$$

Neither of these are appropriate. It does not matter that Eq. (5.37) is not a particularly good approximation, since the emphasis is still on obtaining a positive definite implicit scheme, rather than a highly accurate scheme. In the case of Eq. (5.20) the constraints on the coefficients of ρ_{j-1}^L, ρ_j^L, ρ_{j+1}^n and ρ_j^n require that

$$\begin{aligned} \nu_{j+\frac{1}{2}} - \frac{c_{j+\frac{1}{2}}}{4} - s_{j+\frac{1}{2}} &< 0 \\ \nu_{j+\frac{1}{2}} - \frac{c_{j+\frac{1}{2}}}{4} + s_{j+\frac{1}{2}} &> 0 \end{aligned} \tag{5.38}$$

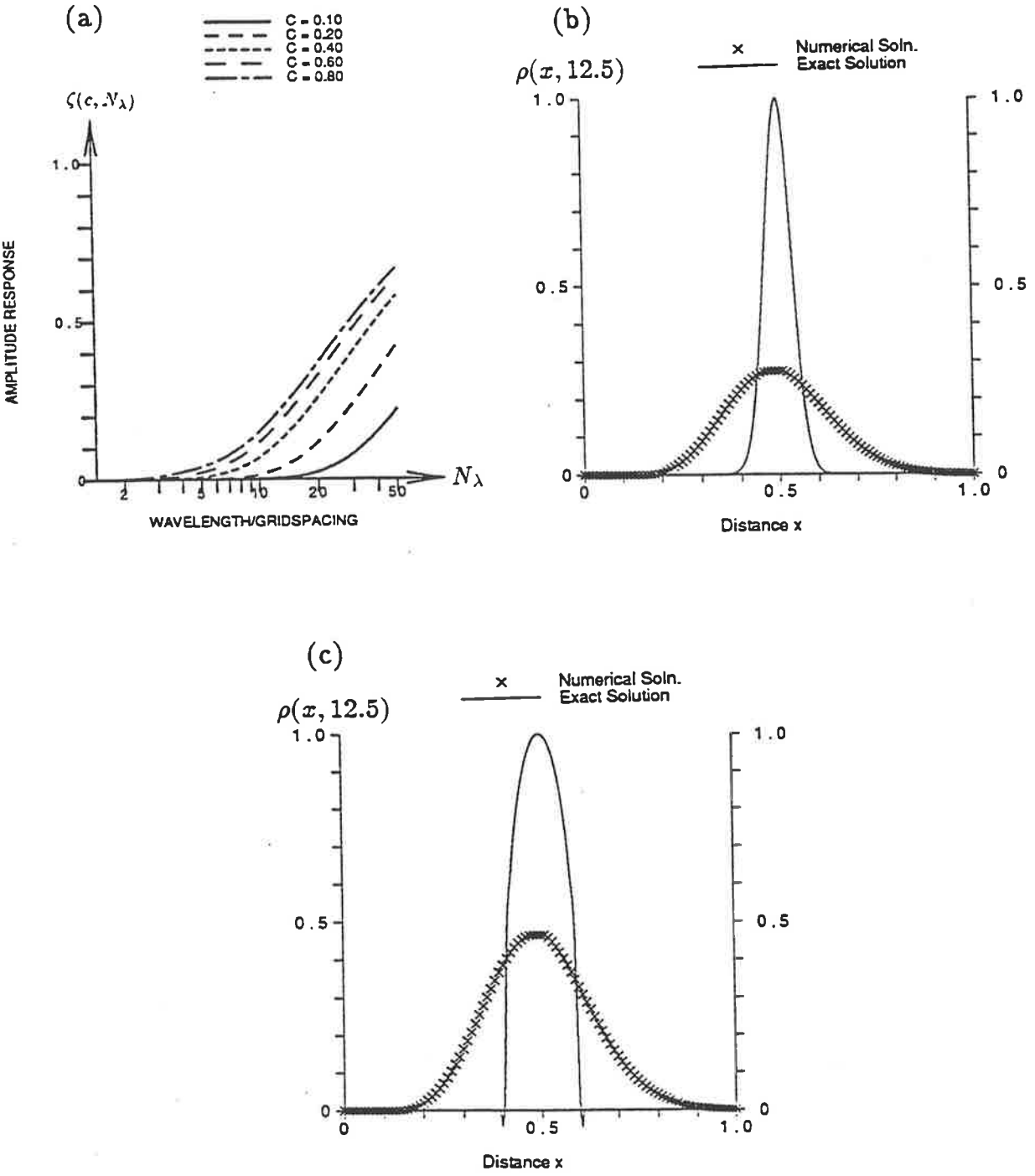

Figure 5.2: Illustration of the performance of REVPOS using Eq. (5.33) and Eq. (5.34) as the low order schemes and antidiffusive fluxes. The diagrams show (a) the amplitude response ζ of the low order scheme, and the results of numerical test cases with cyclic boundary conditions and initial conditions that are (b) a Gaussian pulse and (c) a semi-ellipse.

$$\begin{aligned} \frac{2}{3} - \frac{c_{j+\frac{1}{2}}^2}{6} + 2s_{j+\frac{1}{2}} &< 0 \\ \frac{2}{3} - \frac{c_{j+\frac{1}{2}}^2}{6} - 2s_{j+\frac{1}{2}} &> 0 \quad , \end{aligned}$$

where $\nu_{j+\frac{1}{2}} = (2 + c_{j+\frac{1}{2}}^2)/12$. The first two inequalities require

$$s_{j+\frac{1}{2}} > \left| \nu_{j+\frac{1}{2}} - \frac{c_{j+\frac{1}{2}}}{4} \right| \quad (\geq 0) \tag{5.39}$$

and the last two require

$$2s_{j+\frac{1}{2}} < -\left| \frac{2}{3} - \frac{c_{j+\frac{1}{2}}^2}{6} \right| \quad (\leq 0) \quad . \tag{5.40}$$

It is impossible to satisfy both (5.39) and (5.40). If Eq. (5.37) is used then $s_{j+\frac{1}{2}}$ is replaced by $\frac{1}{2}s_{j+\frac{1}{2}}$ in (5.39) and (5.40), and a similar contradiction is obtained.

The remaining possibility is to select a form for $s_{j+\frac{1}{2}}$ that will eliminate the coefficient of ρ_{j+1}^{n+1}. In this case it is the coefficient of ρ_j^{n+1} that must be positive and this leads to Eq. (5.33) as before.

The two FCT algorithms developed here, CN2POS and REVPOS, show that with appropriate corrections REVFCT can be made positive definite. These new algorithms follow the original calculations of REVFCT very closely in that an implicit solution is calculated, from which the antidiffusive fluxes are evaluated directly. These fluxes are then limited and the final solution is obtained in exactly the same manner as for REVFCT. Only two modifications are required to the REVFCT algorithm to make it positive definite. Firstly, the low order scheme had to be made positive definite. This involved determining constraints on the coefficients of the difference equation to guarantee a positive definite scheme. Secondly, the antidiffusive fluxes must be calculated from the low order solution to retain the positive definite nature of this solution. It is unfortunate that these modifications result in low order schemes that are so diffusive as to seriously degrade the results of REVFCT. However, it does show that implicit schemes can be used in FCT algorithms and the flux limiter is capable of controlling the overall solution.

5.4 A High Order Implicit FCT Algorithm

Given that the major source of error in the two schemes CN2POS and REVPOS was due to the very strongly diffusive low order solution, there is still scope for developing an improved solution by using a less diffusive low order solution. The reason for this improvement is that it is predominantly via the low order scheme that diffusion enters the final solution, and so the less diffusive the low order scheme, the less diffusion will be present in the final, corrected solution. The modified equivalent equation approach discussed in Chapter Two is ideally suited to obtaining the least diffusive linear finite difference scheme, since the problem is then one of minimizing the coefficient of $\partial^2\rho/\partial x^2$, namely η_2, with respect to the coefficients of a general finite difference equation. The minimization must be performed over a class of implicit schemes since Godunov (1959) demonstrated that first order upwinding was the least diffusive positive definite, explicit, linear finite difference scheme. As a result the constraints on the minimization become the conditions for an implicit scheme to be positive definite, described in Section 5.2.

The most general three point implicit linear finite difference scheme (for modelling constant velocity advection) is

$$\begin{aligned}\rho_j^{n+1} - \rho_j^n \;+\; \frac{u\Delta t}{\Delta x}\Big[\theta_1(\rho_{j+1}^{n+1} - \rho_j^{n+1}) + \theta_0(\rho_j^{n+1} - \rho_{j-1}^{n+1}) \\ +\zeta_1(\rho_{j+1}^n - \rho_j^n) + \zeta_0(\rho_j^n - \rho_{j-1}^n)\Big] = 0 \qquad (5.41)\end{aligned}$$

where

$$\theta_1 + \theta_0 + \zeta_1 + \zeta_0 = 1 \quad , \qquad (5.42)$$

for consistency. The corresponding difference equation is

$$\begin{aligned}-c\theta_0\rho_{j-1}^{n+1} \;+\; & (1 - c\theta_1 + c\theta_0)\rho_j^{n+1} + c\theta_1\rho_{j+1}^{n+1} \\ = \; & c(1 - \theta_0 - \theta_1 - \zeta_1)\rho_{j-1}^n \qquad (5.43) \\ & +(1 + 2c\zeta_1 - c + c\theta_0 + c\theta_1)\rho_j^n - c\zeta_1\rho_{j+1}^n\end{aligned}$$

and the modified equivalent equation is

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} = \frac{u\Delta x}{2}(1 - c - 2(1-c)\theta_1 + 2c\theta_0 - \zeta_1)\frac{\partial^2 \rho}{\partial x^2} + O\{(\Delta x)^3\} \quad . \tag{5.44}$$

For convenience, implicit schemes will be consistently referred to as either "matrix inversion", implying use of the Thomas algorithm, or "marching", implying that the solution is obtained by a recurrence relation. Strictly speaking both schemes are forms of direct matrix inversion, but for the want of better terms, these two will be used. Consider the case where Eq. (5.43) with $j = 1, \ldots, J$, is to be solved by matrix inversion. If the scheme is to be positive definite, the inequalities

$$\begin{aligned} \theta_0 &\geq 0 \\ \theta_1 &\leq 0 \\ \theta_0 + \theta_1 + \zeta_1 &\leq 1 \\ 2\zeta_1 + \theta_0 + \theta_1 &\leq -1 + 1/c \quad (c \neq 0) \\ \zeta_1 &\leq 0 \end{aligned} \tag{5.45}$$

must hold. The first two constraints ensure that the coefficient of ρ_j^{n+1}, namely $(1 - c\theta_1 + c\theta_0)$ is positive. The solution is chosen to minimize the numerical diffusion, namely $|1 - c - D|$ where

$$D = 2(1-c)\theta_1 - 2c\theta_0 + \zeta_1 \quad . \tag{5.46}$$

If $c \leq 1$ then, $D \leq 0$ and the minimum of $|1 - c - D|$ occurs when D is a maximum. The largest D can be is zero, which occurs when $\theta_1 = \theta_0 = \zeta_1 = 0$, which produces the first order upwinding scheme, Eq. (2.27). Furthermore, if $\zeta_1 = \theta_1 = 0$, so that only methods involving ρ_{j-1}^{n+1}, ρ_j^{n+1}, ρ_{j-1}^n and ρ_j^n are used, then again D attains a maximum value of zero when $\theta_0 = 0$.

The discussion so far has concentrated on three point schemes that give rise to systems of equations that must be solved by matrix inversion techniques. One of the reasons for this is that any implicit scheme involving three or more unknowns, that

is solved by a marching method (these methods are sometimes referred to as semi-implicit), contains at least one oscillatory mode within the solution. For marching schemes, the second condition to guarantee positive definite results, given in Section 5.2, can be relaxed to be : $a_{jk} \leq 0$, $|a_{jk}| < a_{jj}$for all $k \neq j$. This follows directly from the conditions for explicit schemes to provide positive definite results. For schemes that satisfy these conditions none of these oscillatory modes will dominate in linear models as the solution must be positive definite. In problems involving non-linear feedbacks, however, it is conceivable that such fluctuations could produce problems. For example, in gas dynamics, during the calculation of the charge density, ρ, the velocity may be taken to be prescribed, providing a linear system of equations for the density. Later in the modelling process these densities are in turn used to calculate the particle velocities, and this type of non-linear feedback may excite any oscillatory modes in the solution, possibly leading to the appearance of non-physical extrema. A detailed description of these oscillatory modes follows.

Firstly, it must be noted that the behaviour of any linear system of equations can be determined by treating them as a set of recurrence relations. The decision as to whether the solution is to be calculated by matrix inversion or by use of the recurrence relation is governed only by the stability of the solution scheme. When analyzing the behaviour of the solution of the system of equations it is appropriate to assume that the calculations are performed to infinite precision, in which case the method of calculation is immaterial, since both techniques must give the same results by the uniqueness of solutions to such systems of linear algebraic equations. This means that in order to analyze the behaviour of the solution to a linear system of algebraic equations, it is sufficient to analyze the solution to the corresponding recurrence relations.

Consider the system of equations of the form

$$\sum_{k=0}^{K} a_k \rho_{j-k}^{n+1} = d_j \tag{5.47}$$

representing a difference approximation to the constant velocity form of the advec-

tion equation. The summation is taken over all points involved in the recurrence relation and without loss of generality a_0 and a_K may be assumed to be non-zero. The solution to this equation is given by

$$\rho_j^{n+1} = \sum_{i=1}^{K} A_i \lambda_i^j + P_j \tag{5.48}$$

where P_j is determined by the vector function $\{d_j\}$, the A_i are determined by the boundary conditions and the λ_i are the roots of the characteristic polynomial

$$\sum_{k=0}^{K} a_{K-k} \lambda^k \quad . \tag{5.49}$$

By equating the coefficients of this polynomial with those of

$$a_0 \prod_{k=1}^{K} (\lambda - \lambda_k) \tag{5.50}$$

it is clear that all of the eigenvalues, λ_k, are positive if, and only if, the coefficients, a_{K-k} of the recurrence relation are alternating in sign. By comparing this result with the criteria for implicit, marching schemes to be positive definite, described above it can be seen that for such a scheme involving three or more unknowns at least two coefficients must be of the same sign (both negative), indicating the existence of at least one negative root of Eq. (5.49). It should be noted that this argument also applies to any implicit scheme of four or more points that is solved by direct (or elimination) methods.

It must be stressed that these negative roots do not invalidate the conditions for smooth results, since the presence of a negative root does not automatically imply the presence of any new extrema in the time advanced solution. If a difference scheme obeys the stated conditions then the dominant root will be positive and so the scheme will be positive definite, provided that the boundary conditions are appropriate. The smaller amplitude oscillatory mode does, however, mean that there is not a steady, exponential decay from a sudden disturbance and it is thus possible that such fluctuations could be amplified in problems involving any non-linear feedbacks.

A further problem is the requirement for the calculation of additional boundary conditions, e.g. the calculation of ρ_1^{n+1} in a three point marching scheme. If care is not taken in the calculation of this value, this too may produce a new extremum. This is demonstrated by the following example. Consider material arriving at a previously empty region, i.e. $\rho_j^n \equiv 0$ for all j and $\rho_0^{n+1} = 1$. If ρ_1^{n+1} is taken to be zero, as would be the case if it were calculated by a purely explicit scheme and a three point scheme is used from there on, clearly $\rho_2^{n+1} > 0$, implying that a new minimum has appeared in the numerical solution. This simple case may be corrected by matching fluxes in and out of the first grid cell, but it is still possible that in more complicated problems a numerical extremum may be generated.

A final problem is that due to the expense in solving penta-diagonal systems of linear algebraic equations it does not seem profitable to devise a five point implicit low order scheme for use in conjunction with a three point implicit high order scheme. For example, if a five point implicit high order scheme (such as NS4 or NS5) is used, then the computational cost of inverting two pentadiagonal matrices would produce a scheme that is likely to be extremely expensive in terms of CPU time. Furthermore, it is unlikely that such a scheme would be of sufficient accuracy to justify the expense of CPU time. For these reasons, it is considered unnecessary to further investigate other implicit smooth difference schemes for use within an FCT algorithm, and it appears that first order upwinding is a natural choice for a low order scheme within such an algorithm. The use of such a low order scheme, however, means that Boris and Book's approach to FCT must be abandoned in favour of Zalesak's.

Zalesak's formulation of FCT does not require adding or subtracting diffusion from a scheme in order to determine the antidiffusive fluxes, allowing complete freedom in the choice of high and low order schemes. The antidiffusive fluxes are simply defined to be the difference between the high and the low order fluxes. The source of either of these fluxes is completely up to the user. For example, the high

order fluxes may be chosen to be those of REVFCT Eq. (5.1),

$$\begin{aligned}\phi^H_{j+\frac{1}{2}} &= -\frac{c_{j+\frac{1}{2}}}{4}\left(\rho^{n+1}_{j+1}+\rho^{n+1}_{j}+\rho^{n}_{j+1}+\rho^{n}_{j}\right)\\ &\quad -\left(\frac{1}{6}+\frac{c^2_{j+\frac{1}{2}}}{12}\right)\left(\rho^{n+1}_{j+1}-\rho^{n+1}_{j}-\rho^{n}_{j+1}+\rho^{n}_{j}\right)\end{aligned} \tag{5.51}$$

which are the same as for the fourth order CTCS scheme, Eq. (3.14). First order upwinding could be used to give the low order fluxes,

$$\phi^L_{j+\frac{1}{2}} = c_{j+\frac{1}{2}}\rho^n_j \tag{5.52}$$

giving the antidiffusive fluxes

$$\begin{aligned}\phi_{j+\frac{1}{2}} &= -\frac{c_{j+\frac{1}{2}}}{4}\left(\rho^{n+1}_{j+1}+\rho^{n+1}_{j}+\rho^{n}_{j+1}-3\rho^{n}_{j}\right)\\ &\quad -\left(\frac{1}{6}+\frac{c^2_{j+\frac{1}{2}}}{12}\right)\left(\rho^{n+1}_{j+1}-\rho^{n+1}_{j}-\rho^{n}_{j+1}+\rho^{n}_{j}\right)\end{aligned} \tag{5.53}$$

These fluxes, along with Boris and Book's limiter, produce a positive definite, implicit difference scheme (denoted CN4FCT) which is also slightly more accurate than that of REVFCT. Diagrams illustrating the performance of CN4FCT and REVFCT are shown in Fig. 5.3. The difference in accuracy is not very obvious in these figures but can be detected in Tables 5.1 and 5.2.

There is one problem with the scheme as described above, namely that it is still not valid for $c \geq 1$. Although the high order difference equation is von Neumann stable for $c \geq 1$, neither the low order scheme or the matrix inversion used to evaluate ρ^H_j is valid.

To overcome the restriction on the time-step, multiple upwind steps can be taken during each time-step of the high order solution, provided the high order solution remains stable. The simplest method for calculating the low order fluxes is to use

$$f_{j+\frac{1}{2}} = \rho^L_j - \rho^n_j - f_{j-\frac{1}{2}} \quad . \tag{5.54}$$

The only restrictions on the individual time-steps are that each low order calculation must be stable, and the sum of these time-steps equals the time-step of the high

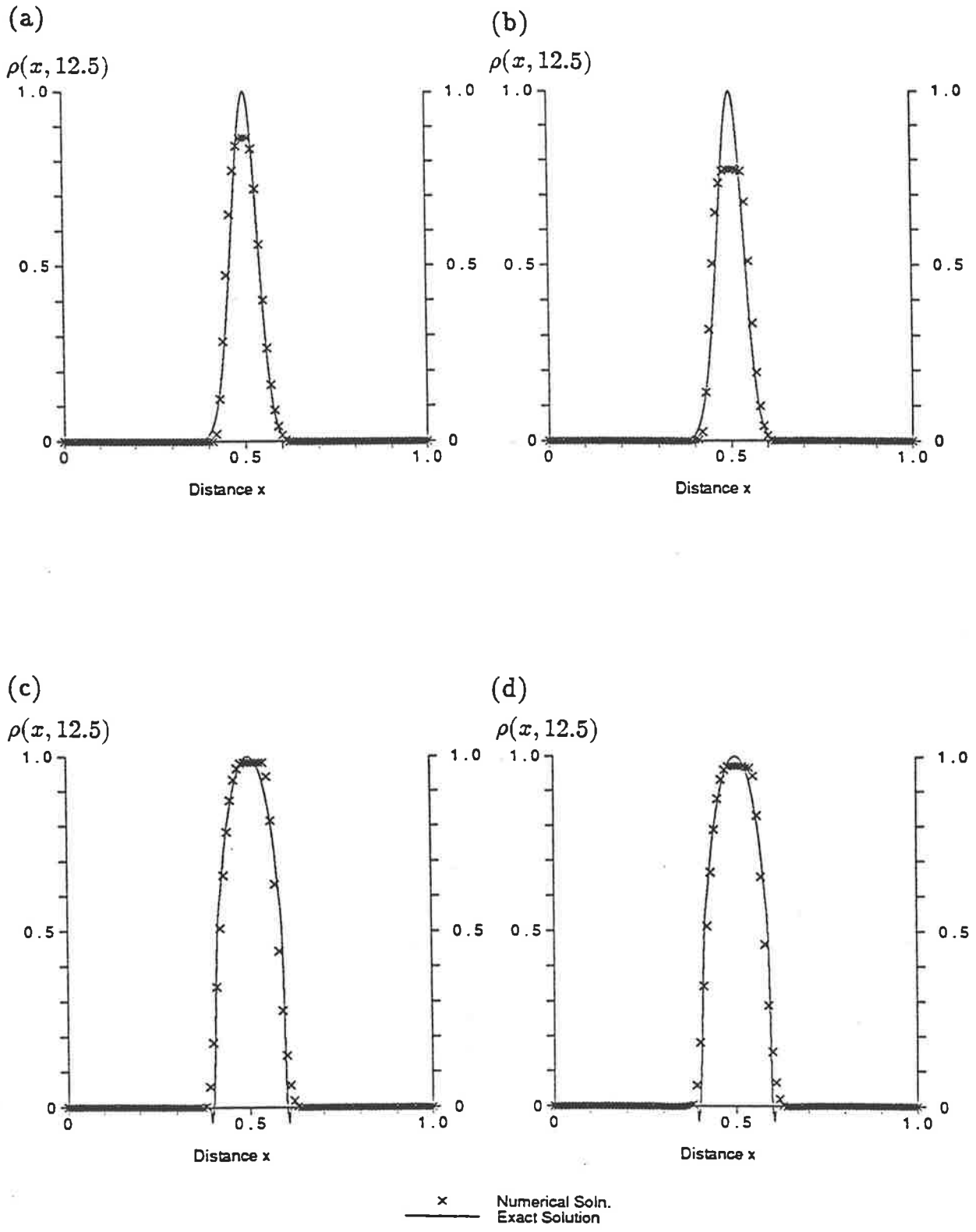

Figure 5.3: Comparisons of CN4FCT and REVFCT. The first two diagrams show the results of using (a) CN4FCT and (b) REVFCT with a Gaussian pulse as an initial condition. The other two diagrams show (c) CN4FCT and (d) REVFCT with a semi-elliptical initial condition.

order calculation. In the test cases provided, the time-steps were chosen so that the maximum value of $c_{j+\frac{1}{2}}$ for the i^{th} low order time-step did not exceed 0.8, with the last time step being adjusted to bring the low order solution in to line with high order solution. The value of 0.8 was chosen because in large, complicated models it is often the case that the time-step is chosen so that the maximum Courant number is slightly less than that required for stability (Leslie et. al., 1985). While any value less than one would avoid instability in the linear case, it is necessary to err on the side of caution when dealing with non-linear problems. Since the aim is to produce a general scheme, it was decided to imitate practical non-linear modelling as close as possible.

Eq. (5.54) can be used to calculate the fluxes from any numerical scheme defined on a grid, not just finite difference schemes. If the high order solution, ρ_j^H is used in Eq. (5.54), instead of the ρ_j^L, then this equation can also be used to calculate the high order fluxes. The antidiffusive fluxes can also be calculated by the same equation, since

$$\begin{aligned} \rho_j^H - \rho_j^L &= -f_{j+\frac{1}{2}}^H + f_{j+\frac{1}{2}}^L + f_{j+\frac{1}{2}}^H - f_{j-\frac{1}{2}}^L \\ &= \phi_{j+\frac{1}{2}} - \phi_{j-\frac{1}{2}} \end{aligned} \tag{5.55}$$

or

$$\phi_{j+\frac{1}{2}} = \rho_j^L - \rho_j^H + \phi_{j-\frac{1}{2}} \quad . \tag{5.56}$$

This approach can also be used to overcome the problem of solving for ρ_j^H when $c \geq 1$ and the Thomas algorithm is unstable. In this case the marching form of Eq. (5.1) must be used, but the fluxes must be calculated *a posteriori*. If this is done explicitly using

$$\begin{aligned} f_{j+\frac{1}{2}}^H &= \frac{c_{j+\frac{1}{2}}}{4}\left(\rho_{j+1}^H + \rho_j^H - \rho_{j+1}^n - \rho_j^n\right) \\ &\quad -\nu_{j+\frac{1}{2}}\left(\rho_{j+1}^H - \rho_j^H - \rho_{j+1}^n + \rho_j^n\right) \end{aligned} \tag{5.57}$$

then there is a significant overhead relative to the original calculation of ρ_j^H. Alternatively, using Eq. (5.56) to calculate the antidiffusive fluxes requires only two extra

additions per grid-point per time-step. This method can also be used when $c < 1$, the results being identical to those obtained by the explicit calculations of $\phi_{j+\frac{1}{2}}$ in both cases.

This approach for calculating the high order fluxes allows the marching method to be merged with the inversion method when the Courant number changes from greater than one to less than one. In a region of decreasing velocity, the marching scheme can be used to provide an upstream boundary condition (at the point where the Courant number decreases from greater than one to less than one) for the inversion method. In the case of an increasing velocity field, the low order solution can be used to give a downstream boundary condition (at the point where the Courant number changes from less than one to greater than one) for the inversion technique. The marching method can then be used from that point.

5.5 Extensions to Implicit FCT

Described above is a finite difference scheme that guarantees smooth results for all values of c. The approach adopted was quite general, any other finite difference scheme could have been used as the high order scheme. In fact any numerical scheme may have been used since it was shown how to construct the antidiffusive fluxes using only the grid-point values from two schemes, one high order and the other low order, along with the flux at one point (usually at a boundary). Similarly, any numerical scheme that produces smooth results may have equally well been used as a low order scheme. There are three components to the algorithm CN4FCT, the high order solution, Eq. (3.14), the low order solution, Eq. (2.27), and the flux limiter, Eq. (4.34). The importance of each of these components will be examined in this section.

The fourth order CTCS scheme, Eq. (3.14) was used to provide the high order solution in the above algorithm, because CN4FCT was originally a corrected form of REVFCT. In Chapter Three it was shown that the fourth order scheme is an accur-

ate, unconditionally stable and efficient algorithm for modelling passive advection. Of course, other high order schemes could have been chosen, but it will be demonstrated here that this particular choice of high order scheme is especially suited for use in a general purpose, positive definite algorithm for modelling advection. This will be shown by comparing the results of CN4FCT with similar algorithms using other high order schemes such as the Crank-Nicolson scheme, Eq. (3.8) and the two five point implicit schemes NS4, Eq. (3.50) and NS5, Eq. (3.60). Comparisons will also be made with algorithms using explicit high order schemes such as the Lax-Wendroff scheme, Eq. (2.29) and fifth order upwinding, Eq. (2.39) to demonstrate the efficiency of using implicit high order schemes in FCT algorithms. These five algorithms will be denoted CN2FCT, NS4FCT, NS5FCT, LWFCT and UW5FCT respectively. Further comparisons will be made with other positive definite algorithms such as van Leer's second scheme (VLII), Eq. (4.16), Phoenical LPE SHASTA and Implicit LPE SHASTA. Van Leer's scheme (VLII) was found to be one of the most efficient positive definite algorithms discussed in Chapter Four, with the other two schemes being the most accurate for a given resolution. The stability restriction of $0 < c \leq 1/2$ for these last two schemes, however, requires that the time-step must be half of that for comparable schemes and so the efficiency of these schemes is reduced. Illustrations of the results from the numerical tests of these last three schemes were presented in Chapter Four, and the results of the schemes CN2FCT, LWFCT, UW5FCT, NS4FCT and NS5FCT are shown in Figs. 5.4 - 5.5. Quantitative measures of the accuracy of all these numerical tests are presented in Tables 5.1 and 5.2. Table 5.3 compares the effect of improved resolution on the performance of these schemes. These tables also include the algorithms discussed earlier in this chapter, namely REVFCT, CN2POS and REVPOS.

The initial conditions used are Gaussian and semi-elliptical pulses. The first conclusion to be made from these comparisons is that the high order scheme in an FCT algorithm should ideally be better than second order. This can be seen

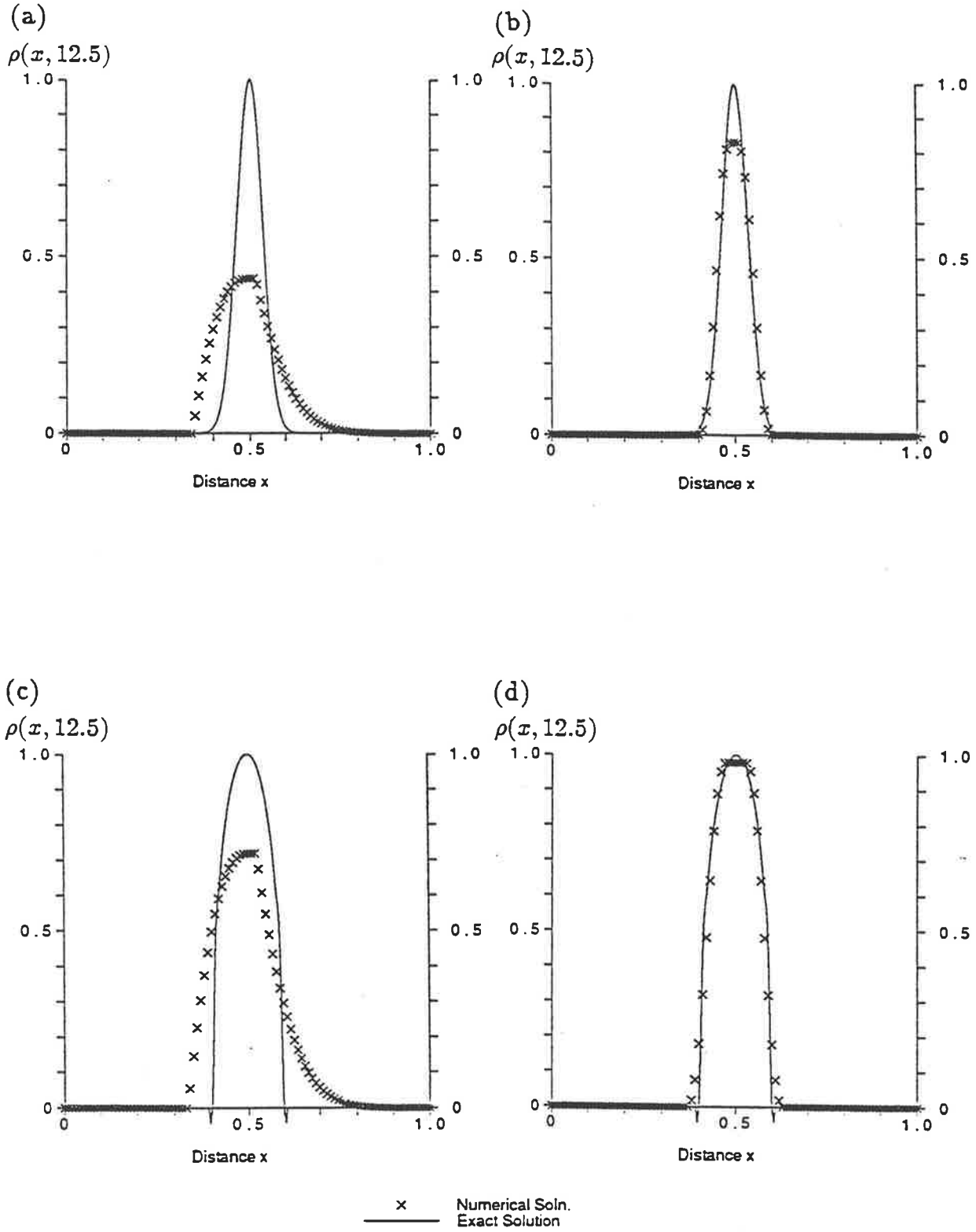

Figure 5.4: Illustration of the performance of CN2FCT and UW5FCT which use 2^{nd} order Crank Nicolson, Eq. (3.8) and 5^{th} order upwinding, Eq. (2.39) as high order schemes. The low order scheme is 1^{st} order upwinding, Eq. (2.27) in both cases. The initial conditions are Gaussian and semi-elliptical pulses. The diagrams (a,c) show the results of CN2FCT, and (b,d) show UW5FCT.

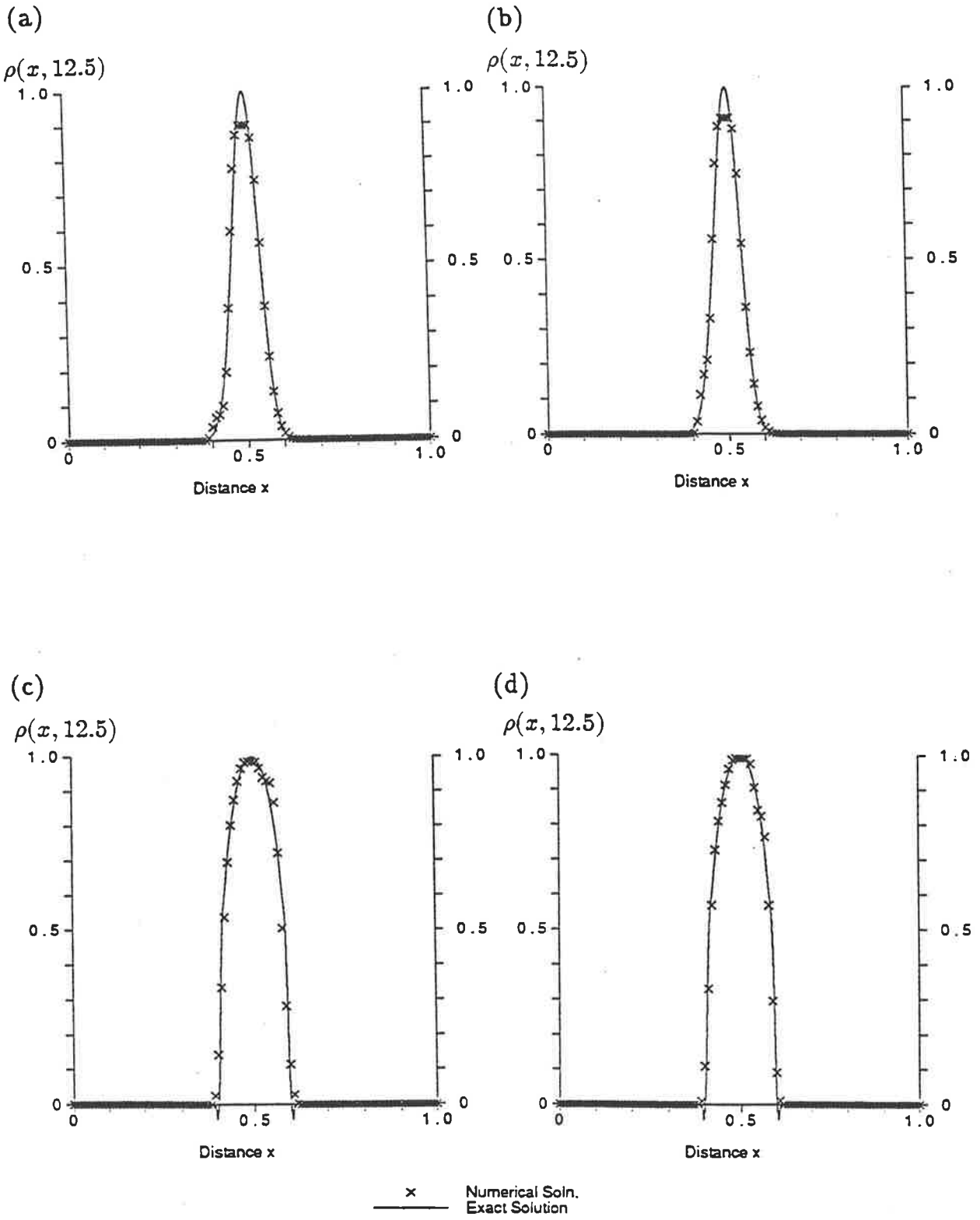

Figure 5.5: Illustration of the performance of NS4FCT and NS5FCT which use Eq. (3.50) and Eq. (3.60) as high order schemes. The low order scheme is 1^{st} order upwinding, Eq. (2.27) in both cases. The initial conditions are Gaussian and semi-elliptical pulses. The diagrams (a,c) show the results of NS4FCT, and (b,d) show NS5FCT.

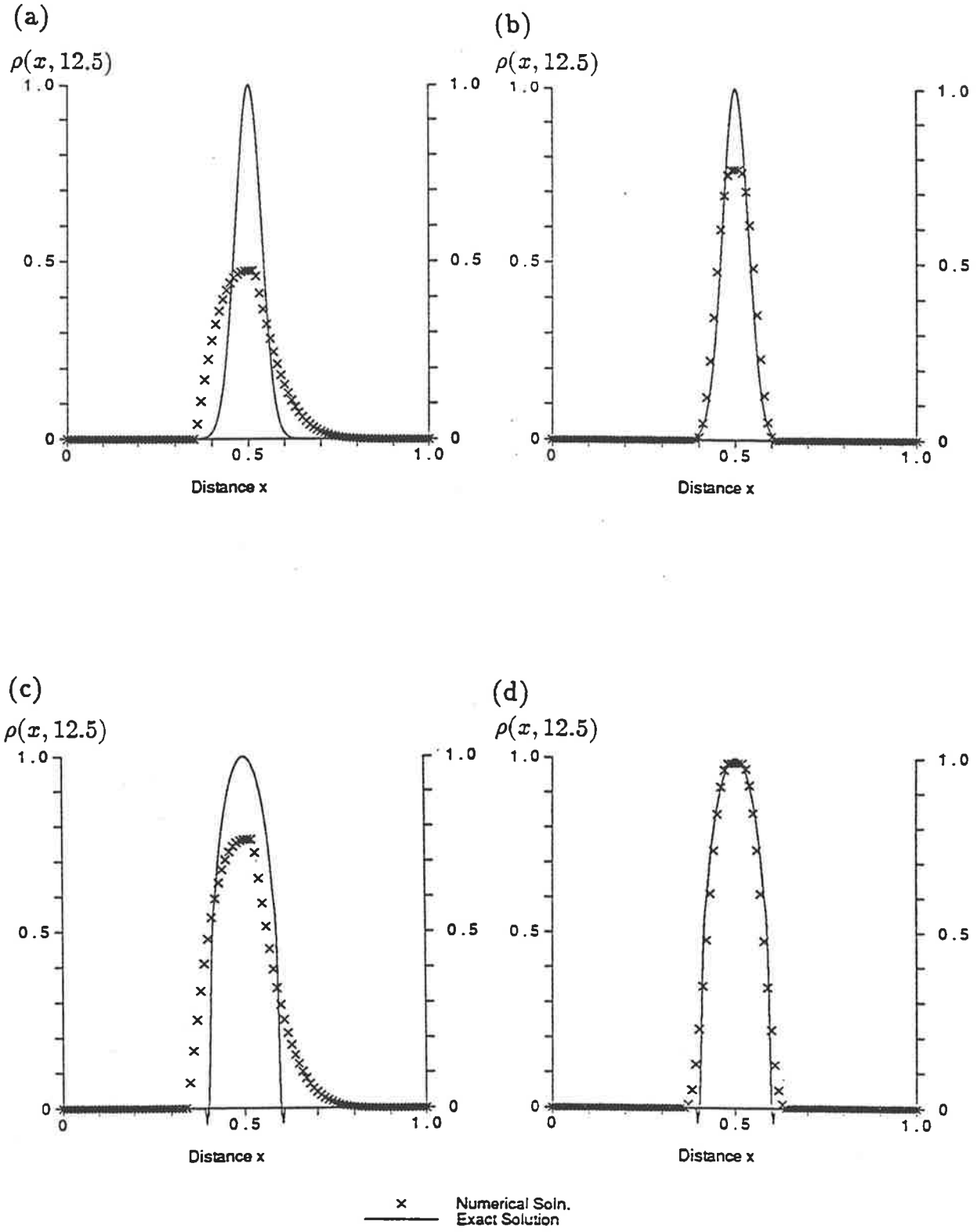

Figure 5.6: Illustration of the performance of Lax Wendroff scheme, Eq. (2.29) with FCT, and Implicit LPE SHASTA. The initial conditions are Gaussian and semi-elliptical pulses. The diagrams (a,c) show the results of Lax-Wendroff with FCT, and (b,d) show Implicit LPE SHASTA.

Table 5.3:Comparisons of improvement in R.M.S. error versus increase in CPU time due to increasing resolution. The test problems use either a Gaussian pulse or a semi-circle as the initial condition along with cyclic boundary conditions. The CPU Times are for the calculations with $J = 100$ and 800.

	CPU	Gaussian Pulse				Semi-elliptical Pulse				CPU
Scheme	$J = 100$		$J = 200$	$J = 400$	$J = 800$	$J = 100$	$J = 200$	$J = 400$	$J = 800$	
REVFCT	35	4.9×10^{-2}	1.5×10^{-2}	3.5×10^{-3}	7.9×10^{-4}	3.8×10^{-2}	2.2×10^{-2}	1.3×10^{-2}	7.6×10^{-3}	2220
CN2POS	27	1.6×10^{-1}	1.1×10^{-2}	5.3×10^{-2}	2.6×10^{-2}	1.7×10^{-1}	9.8×10^{-2}	5.5×10^{-2}	3.3×10^{-2}	1720
REVPOS	30	1.8×10^{-1}	1.5×10^{-1}	1.2×10^{-1}	8.4×10^{-2}	2.2×10^{-1}	1.7×10^{-1}	1.2×10^{-1}	8.2×10^{-2}	1920
CN4FCT	33	2.8×10^{-2}	6.1×10^{-3}	1.4×10^{-3}	3.2×10^{-4}	3.9×10^{-2}	2.2×10^{-2}	1.3×10^{-2}	7.6×10^{-3}	2140
CN2FCT	31	1.4×10^{-1}	8.8×10^{-2}	4.1×10^{-2}	2.0×10^{-2}	1.5×10^{-1}	8.0×10^{-2}	4.6×10^{-2}	3.0×10^{-2}	2000
NS4FCT	64	1.9×10^{-2}	4.7×10^{-3}	1.1×10^{-3}	2.4×10^{-4}	3.0×10^{-2}	1.7×10^{-2}	9.3×10^{-3}	5.1×10^{-3}	4070
NS5FCT	68	1.7×10^{-2}	4.3×10^{-3}	9.6×10^{-4}	2.0×10^{-4}	2.4×10^{-2}	1.3×10^{-2}	7.2×10^{-3}	3.9×10^{-3}	4330
UW5FCT	30	3.4×10^{-2}	7.1×10^{-3}	1.6×10^{-3}	3.6×10^{-4}	3.9×10^{-2}	2.2×10^{-2}	1.2×10^{-2}	6.7×10^{-3}	1950
van Leer (II)	8.5	1.1×10^{-1}	4.8×10^{-2}	1.6×10^{-2}	4.9×10^{-3}	8.5×10^{-2}	4.1×10^{-2}	2.6×10^{-2}	1.6×10^{-2}	550
Phoenical LPE SHASTA	24	4.9×10^{-2}	2.0×10^{-2}	7.3×10^{-3}	4.3×10^{-3}	4.4×10^{-2}	3.5×10^{-2}	2.1×10^{-2}	1.3×10^{-2}	1510
Implicit LPE FCT	35	4.8×10^{-2}	1.2×10^{-2}	2.1×10^{-3}	4.5×10^{-4}	4.6×10^{-2}	2.8×10^{-2}	1.6×10^{-2}	9.4×10^{-3}	2260
SAHS (Lax-Wendroff)	8.5	1.1×10^{-1}	4.9×10^{-2}	1.8×10^{-2}	5.5×10^{-3}	8.5×10^{-2}	4.2×10^{-2}	2.6×10^{-2}	1.6×10^{-2}	550

by comparing the results for CN2FCT and LWFCT with other schemes. The only difference between these two algorithms and CN4FCT, UW5FCT, NS4FCT and NS5FCT is the choice of the high order scheme used, with exactly the same low order scheme and flux limiter being used. The poor performance of the algorithms using second order schemes is directly attributable to the low quality of the high order solution. Using CN4FCT with half the resolution produces superior results and requires less CPU time. Using either the Crank-Nicolson or Lax-Wendroff schemes also introduces significant distortion to the shape of the pulse, with a sudden change in curvature at the peak. In the initial discussion of these second order schemes in Chapters Two and Three it was seen that the phase lag of the shorter wavelengths results in the leading edge of the pulse being smooth and the oscillations occurring on the trailing side of the pulse and upstream from there. As a result the high order fluxes are used to calculate most of the leading edge, but the low order fluxes are used to calculate most of the trailing edge. This causes most of the residual diffusion to appear on the upstream side of the pulse resulting in the deformation seen in Fig. 5.4(a,c) and Fig. 5.6(a,c). As such the deformation is typical of these two methods, and is easily verified by using a variety of initial conditions. The deformation is also reflected in the substantial degradation of the second and third moments of the numerical solution, relative to that of CN4FCT. This is particularly noticeable with a Gaussian pulse as the initial condition.

A second point that can be drawn from these comparisons is that using even higher order schemes is still very efficient in respect of the accuracy attained in a given amount of CPU time, although the improvement is not as marked as was seen in Chapter Three, due to the errors associated with the limiter. The errors due to the flux limiter are predominantly associated with the five or seven gridpoints near the numerical peak, irrespective of the grid-spacing. The error associated with the limiter diminishes only slowly as the mesh is refined, and so the increase in accuracy of a particular scheme that can be obtained by using a finer mesh is less than was

seen in Chapters Two and Three. This means that given the choice of using a higher order method on a coarse mesh or a low order method on a fine mesh, it is preferable to use the former, assuming that both configurations require approximately the same CPU time. An additional reason for the sustained efficiency of the algorithms involving high order implicit schemes is that the flux correction process requires approximately the same amount of computation for all the algorithms. Since the flux correction takes a significant fraction of the total time, the additional time spent inverting a penta-diagonal matrix, as a fraction of the total CPU time, is much less than was seen in Chapter Three. It should be noted however, that the matrix inversions associated with the schemes NS4 and NS5 become unstable as the Courant number, c, increases beyond one, but the fourth order CTCS scheme is unconditionally stable. In addition, the accuracy of the three point scheme scheme does not deteriorate significantly until c is well in excess of two. As a result, the times for CN4FCT can almost be halved since the high order solution and flux limiting process is only calculated half as often and so this scheme becomes the most efficient of all.

Comparing the results for the Gaussian initial condition with those for the semi-ellipse indicates that the use of higher order implicit schemes in conjunction with FCT is most advantageous with irregular profiles. The results for the Gaussian pulse test case indicate that CN4FCT and NS5FCT are of a similar efficiency for a given Courant number. The additional CPU time spent using CN4FCT on a high resolution grid as compared to using NS5FCT on a coarse grid, being compensated by the gain in accuracy. However, for the semi-ellipse case, the results for NS5FCT on the coarse grid are quite close to those of CN4FCT on the finer grid. At first this may appear somewhat surprising given the results in Chapter Three where NS5 exhibited only moderate improvement over fourth order CTCS in the case of the semi-ellipse test but substantial improvement for the Gaussian test case when both were tested on the same grid. The improvement of NS5FCT in the semi-ellipse test

is due to the limiter, which has greatest impact on a sharp peak such as Gaussian pulse and less impact on a flatter peak such as that of the semi-ellipse. Furthermore, higher order schemes produce higher frequency oscillations, as noted in Chapter Three. The shorter the wavelength of this numerical noise the more efficiently it is eliminated by the flux limiter, leaving less residual diffusion in the final solution.

It is also interesting to note the performance of the UW5FCT algorithm relative to CN4FCT. In general it would appear that there is little to separate the two algorithms. With smooth initial conditions, CN4FCT is slightly more accurate, but UW5FCT shows some improvement with irregular profiles, such as the semi-ellipse pulse. The explanation for this is that the short wavelengths are damped by fifth order upwinding. In the case where no flux limiting procedure is used, as was the case in Chapters Two and Three, such damping is an obvious advantage, particularly for profiles with discontinuous derivatives. The damping, however, is still not strong enough to eliminate all the oscillations and so the flux limiter is still required, resulting in two sources of numerical diffusion, the high and low order solution. These two factors appear to more than outweigh any advantages due to fifth order upwinding propagating the Fourier components with better phase speed than the fourth order CTCS scheme. Note that, although using fifth order upwinding as a high order scheme produces a slightly faster algorithm in that less calculation is required per time-step the scheme is still restricted to $0 < c \leq 1$. There is no such restriction when the fourth order CTCS scheme is used and as was seen before, the time-step can be almost tripled before the solution begins to significantly deteriorate, resulting in a substantially more efficient algorithm.

The comparisons with other high order schemes illustrate how well the fourth order CTCS scheme is suited for use within an FCT algorithm, in that it provides results of comparable accuracy to other high order schemes relatively efficiently. The comparisons with the use of second order schemes such as the Crank Nicolson and the Lax Wendroff schemes illustrate that even with the same time-step, using the

fourth order CTCS scheme is more efficient.

The final comparison to be made is between CN4FCT and other positive definite schemes discussed in Chapter Four. The schemes included in this comparison were van Leer's extension of Fromm's scheme (VLII), Harten's Self Adjusting Hybrid scheme (SAHS) applied to the Lax Wendroff scheme, Implicit LPE SHASTA and Phoenical LPE SHASTA. The first two algorithms were the most efficient of the schemes discussed in Chapter Four and the last two were the most accurate for a given resolution (ignoring the CPU time necessary for the calculations). To attain the same accuracy as CN4FCT, the first two schemes require more than twice the spatial resolution corresponding to a more than four-fold increase in CPU time relative to the calculations on the original (coarse) grid. This also demonstrates the high efficiency of CN4FCT since the two schemes, VLII and SAHS require more than one-quarter of the time of CN4FCT when used on the same grid.

Comparisons with Implicit LPE SHASTA and Phoenical LPE SHASTA demonstrates the accuracy of the CN4FCT algorithm. This is also important since all computers have a finite amount of memory available for the storage of arrays and sometimes this can be the main restriction and so it is important for an algorithm to be accurate for a given resolution. It is also important for an algorithm to obtain a solution reasonably quickly since CPU time is often limited. As seen from the comparisons with alternative schemes, the use of the fourth order CTCS within an FCT algorithm leads to a very accurate and efficient positive definite numerical scheme.

The remainder of this section will be concerned with the errors due to clipping, which is due to the interaction of the flux limiter and the low order solution. The limiter cannot distinguish between the part of the high order solution corresponding to the "true" peak and an overshoot, especially when the true peak does not lie exactly on a gridpoint. As a result the part of the high order solution coresponding to the true peak is treated as an overshoot and the low order solution is invoked.

Due to the diffusive nature of the low order calculations and the fact that the phase speed of this low order scheme will generally differ from the high scheme, this eventually leads to a region of constant values - the plateau that is characteristic of FCT schemes

There are three ways of reducing the effect of clipping. Firstly, the adjustments to the high order fluxes that are needed to ensure a smooth solution can be reduced. As was seen in Chapters Two and Three, high order finite difference schemes generally produce less severe oscillations than lower order schemes and so the corrections to the high order fluxes of fourth or higher order schemes are much less than those of second order schemes. This was seen above where the errors due to the limiter in the FCT algorithms using higher order schemes (CN4FCT, NS5FCT, etc.) are much less than those of the algorithms involving lower order schemes (CN2FCT, LWFCT, etc.).

The next alternative for reducing clipping errors is to improve the limiter, so that it gives less weight to the low order solution when choosing the range of admissable values for the final solution. Sweby (1984) compared a variety of limiters for TVD schemes, and it was noted that TVD limiters and FCT limiters have some inherent differences and cannot be arbitrarily interchanged. A further limiter has been discussed by Leonard (1990) - the ULTIMATE limiter, which uses values at half gridpoints as well as values at the standard gridpoints. The extra information artificially enhances the accuracy of this limiter in much the same way as Holly and Preissmann's scheme provides substantial improvements over third order upwinding when used on the same grid. When comparable amounts of information were used, the two third order schemes produced similar results. A similar argument applies to the ULTIMATE limiter, as the clipping is still spread over about seven values, but since values at half gridpoints are used, the clipping appears to be spread over only three or four gridpoints. So that when compared on the same grid the ULTIMATE limiter provides greatly improved results, but a fairer test is to use a standard lim-

iter on a double grid and then compare every second gridpoint. This is not to say that there is not scope for improvement on Boris and Book's limiter, but it is not a straightforward task. The other advantage of Boris and Book's limiter is its ease of use and there is no problem with non-linear instabilities such as is the case with Zalesak's peak extrapolating limiter.

The third approach to reducing clipping is to improve the low order scheme, as this will also reduce the necessary adjustments to the high order fluxes. That clipping can be avoided, even when using Boris and Book's limiter, can be seen in Fig. 5.7(a,c, where Smolarkiewicz's scheme was used as a low order scheme and fourth order CTCS was the high order scheme) This method is not proposed as a useful scheme, due to the possibility of overshoots in the low order scheme, but more as a demonstration of the potential for reducing clipping by using a low order scheme that maintains the peak of a pulse. The results from the semi-ellipse test show how inappropriate this scheme is for many profiles. Nevertheless, the results from the Gaussian pulse test show that when a definite peak (as opposed to a broad plateau) appears in the low order scheme, the peak will also appear in the flux-corrected solution. In order to improve the low order scheme it is necessary to use a non-linear scheme, for it has been shown that first order upwinding is the most appropriate linear, low order scheme. Thus it is not such a simple matter to improve the low order scheme since the accuracy of the low order scheme in the vicinity of extrema is most important, and many non-linear schemes simply revert to first order upwinding in such areas. This can be seen in the demonstrations of the use of van Leer's second scheme (VLII) in Fig. 5.7(b,d), which only provides marginal improvement over using first order upwinding, and very little improvement in the vicinity of the peak. If Harten's SAHS is used instead of VLII, the results are indistinguishable. In fact, by using these non-linear schemes it is possible to actually degrade the solution relative to using first order upwinding, since the clipping that would normally appear when these schemes are used by themselves can add to the

clipping of the limiter, leading to an even broader plateau and larger RMS errors.

5.6 Conclusion

This chapter described the development of an accurate and robust positive definite algorithm, CN4FCT. In order to derive such a scheme it was necessary to demonstrate that Boris and Book's objections to using implicit schemes are not necessarily valid. These objections were based on the failure of the algorithm REVFCT to produce positive definite results and it was asserted that the flux-limiter was incapable of handling an implicit high order solution. It has been shown in this chapter that this need not be the case, because there is an alternative reason, namely that the low order solution as calculated within REVFCT was not positive definite. Such a problem equally well explains any difficulties in a non-linear model and it was demonstrated that this could lead to small unphysical oscillations even in the constant coefficient case.

To correct the algorithm REVFCT, it was necessary to derive a set of constraints on the coefficients of implicit finite difference schemes which guarantee positive definite results. This led to the schemes CN2POS and REVPOS, which always produce positive definite results (provided the calculation of the low order solution is stable). The only differences between REVFCT and REVPOS were in the calculation of the low order solution, with the underlying derivations of the two algorithms being otherwise identical. This indicates that it was indeed the calculation of the low order solution that was at fault, rather than any inherent weaknesses due either to the flux-limiter or to the general FCT approach.

The apparent failure of the two positive definite algorithms, REVPOS and CN2POS to model accurately passive advection is due to the excessive numerical diffusion in the low order scheme which is a direct consequence of ensuring positive definite solutions. It must be kept in mind that the primary objective in developing these two schemes was to arrive at an algorithm that closely related REVFCT, while

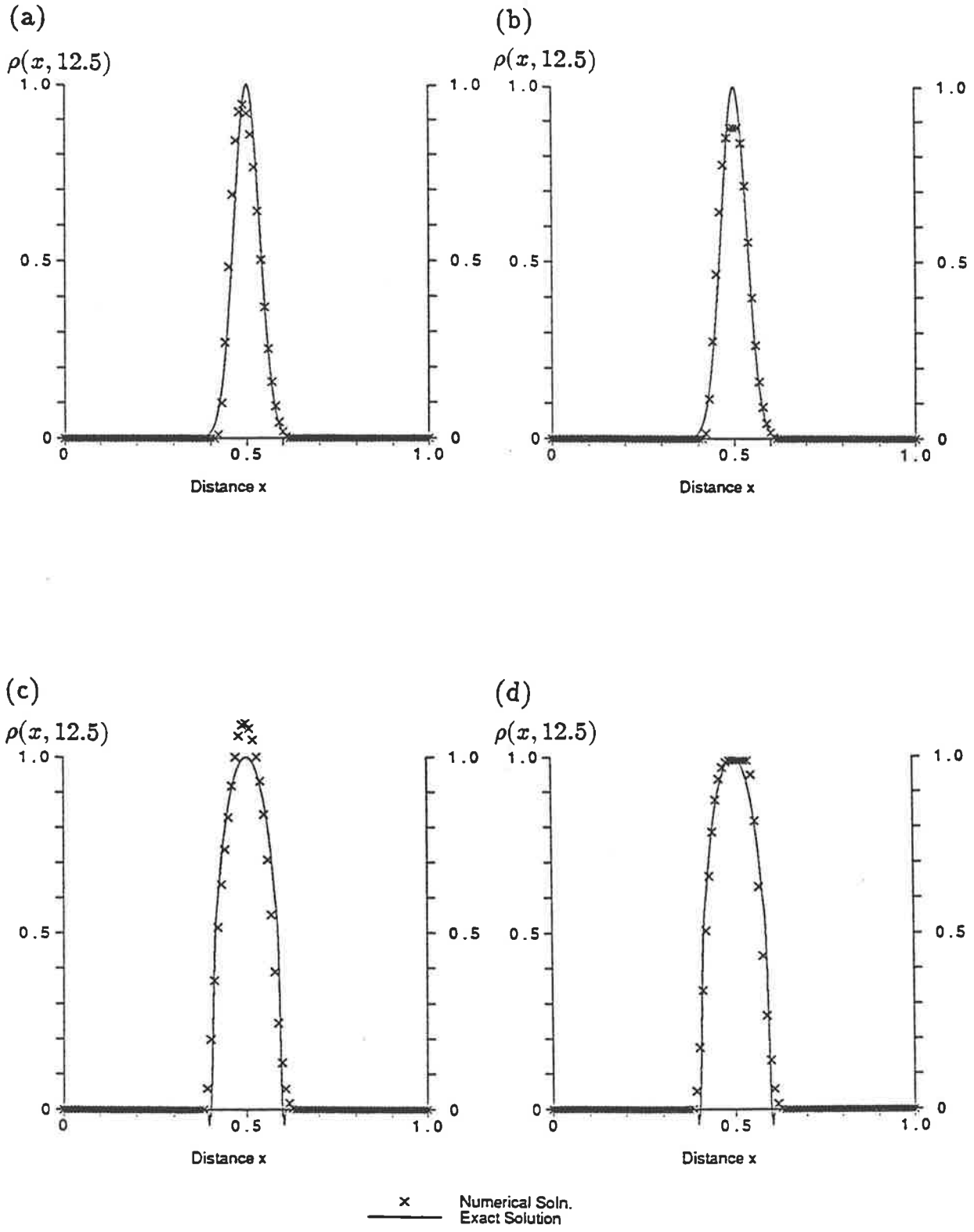

Figure 5.7: Illustration of the performance of CN4FCT, with (a,c) Smolarkiwiecz and (b,d) van Leer (II) schemes providing low order solutions. The initial conditions are Gaussian and semi-elliptical pulses.

maintaining positivity. Because of the high residual diffusion in these schemes, Boris and Book's formulation of FCT was abandoned in favour of Zalesak's approach which allows a greater choice of possible low order solutions. It was shown in Chapters Two and Three that the modified equivalent equation approach provided a convenient measure of a scheme's overall performance. By constructing the modified equivalent equation for a general three point implicit scheme it was shown that first order upwinding was the most suitable choice for a low order scheme. This method has the further advantage of requiring very little CPU time to obtain its solution. Other stencils, involving more than three points, could be used for the low order scheme but the presence of a small, negative root to the characteristic equation means that there is some chance of the non-linear precursors described by Boris and Book reappearing. In addition these schemes would generally require more CPU time and it is questionable how much would be gained, even in the linear case.

It is important to appreciate that CN4FCT is a genuine use of an implicit technique within an FCT framework. A semi-implicit method was discussed by Grandjouan (1990), that used values at half time levels, so that the time advanced solution was calculated explicitly, producing a predictor-corrector algorithm. The resulting scheme only guaranteed positive definite results for $c \leq 1/2$, as compared to CN4FCT which always produces positive definite results. The use of the modified equation approach to calculate the "best" diffusive and anti-diffusive fluxes is appealing, however account must be taken of the constraints on the coefficients of the low order part of calculation, as described in this chapter. Furthermore, when higher order schemes are used, the flux-limiter (and hence the diffusive fluxes) only have a significant effect near the peak of a pulse. In this region, the flux limiter will always set the anti-diffusive fluxes to zero, producing the plateau that is characteristic of the flux-limiter. In the case of NS4FCT and NS5FCT, this plateau is only five or seven grid spacings wide, which is as narrow as can be hoped for, given the particular flux-limiter. This means that the errors in the final solution are dominated

by the errors in the diffusive fluxes, which are determined by how much numerical diffusion is contained in the low order solution. There may well be some benefit to be obtained from using the approach of Grandjouan, but it does not address the major source of errors, namely the flux-limiter or the quality of the low order fluxes.

The importance of each of the components of an FCT algorithm were demonstrated. If a scheme of only second order is used, then the distortion of a pulse can be quite severe. By using higher order schemes, the distortion is considerably reduced. It was also noted that the amount of diffusion present in the low order scheme had considerable impact on the accuracy of the final algorithm. The flux limiter itself is more difficult to analyze, in that its performance is tightly bound in with the accuracy of the low order scheme and to improve the general performance of a limiter is no simple matter. To a certain extent the combined effect of the limiter and the low order solution can be seen by using non-linear low order schemes. Although these schemes are not proposed as realistic algorithms, it does show the difficulty in developing an improvement to the limiter that will have a significant impact on the solution, since by necessity the limiter must modify the result towards the low order solution near the peak.

The results from CN4FCT were shown to compare favourably with alternative positive definite algorithms. Not only were the numerical solutions quite accurate, but they were obtained relatively quickly. This efficiency, coupled with the unconditional stability of the scheme, means that CN4FCT is particularly suited to advecting arbitrary waveforms. It now remains to be shown that these results carry over to more general problems. This will be discussed in the next chapter.

Chapter 6

Further Tests of Implicit FCT

In the previous chapters, the discussion and comparisons have concentrated on the constant velocity case of the advection equation. This was done for reasons of clarity, as the shortcomings of any methods become immediately obvious when such a simple test case has been used. These comparisons, however, have only limited value when considering schemes for practical applications. If the velocity, w, varies with either space or time, the waveform is then distorted by the advection. In Chapter Three it was shown how high order implicit schemes are very accurate when advecting a pulse in constant velocity fields. It remains to be shown that in a varying velocity field, these same schemes will now advect a pulse in the correct manner.

As mentioned in Chapter One, further problems occur when there are non-linear interactions between the velocity field and the material being advected. One of the major difficulties when solving non-linear problems is how to overcome spurious oscillations, as these can lead to severe problems with a numerical model. Such problems range from physically meaningless results to the model run aborting due to an illegal operation. There is only one practical way of demonstrating that a method is capable of handling all the complexities of a non-linear problem, and that is to apply the method to such a problem. For this reason a brief discussion is presented on the results of Morrow (1991), using CN4FCT to model gas discharges. The discussion will concentrate on the numerical aspects of the problem, with only a brief description of the physical processes.

Due to the complexities of modelling gas discharges, the model results can not be verified against an analytic solution, although, the general properties of the solution are known from observation. The observations however, do not resolve all of the small scale structure, but in conjunction with some conceptual models, these observations can be used to infer some of this structure. It is also possible to verify the behaviour of the numerical model in certain simplified cases. From these cases and the knowledge gained in previous chapters about the performance of different numerical schemes it will be shown that CN4FCT is particularly suited to modelling gas discharges.

6.1 A Mathematical Model of Gas Discharges

A mathematical model describing the distribution of charge and electric current in negative corona was described by Morrow (1985). The field is set up between a negatively charged sphere and a positively charged plane. The problem is essentially two dimensional, with the primary variation of the solution being in the axial (x), and radial (r) directions. Observations indicate negligible variation in the angular (θ) direction.

The variation in the radial direction is predominantly due to the finite radius of the discharge, there being little radial development of the discharge. This variation can be calculated by solving Poisson's equation in two dimensions as discussed by Davies and Evans (1967). The problem then becomes one of determining the axial component of the electric field, $E(x)$, which can be found via the method of disks (Davies et al., 1964). This involves representing the discharge as a cylinder, with uniform radial distribution and variable axial distribution of charge. The electric field has two components, the first being due to the space charge in the gap, namely

$$E(x) = \frac{1}{2\epsilon_0}\left[-\int_0^x \rho(x-x')\left\{1+\frac{x'}{\sqrt{x'^2+R^2}}\right\}dx' + \int_0^{d-x} \rho(x+x')\left\{1-\frac{x'}{\sqrt{x'^2+R^2}}\right\}dx'\right] \quad (6.1)$$

where $\rho(x)$ is the net charge density, R is the cylinder radius and d is the distance across the gap from the sphere to the plane. To this, must be added the other component obtained from solving Laplace's equation. The boundary conditions being introduced via the use of image charges beyond the electrode surface. This approach means that the problem is reduced to one of solving for the one-dimensional motion of the charged particles, in order to obtain the spatial distribution of the net charge density.

The calculation of the spatial distribution of the charged particles is complicated by the fact that not only do these particles interact with the electric field, but also with other particles. The interactions between the different species of charged particles are as follows:

1. Neutral particles may be ionized, increasing the number of free electrons and positive ions.

2. Free electrons may recombine with neutral particles to form negative ions.

3. Any oppositely charged particles may recombine to form neutral particles.

In addition, the high velocities of the electrons relative to the ion velocities, means that electron diffusion is also an important process.

Since charge must be conserved, the continuity equations may be used to describe the motion of the three species of charged particle. Denoting the number of electrons, positive ions and negative ions by N_e, N_p and N_n respectively, the resulting equations are

$$\begin{aligned}
\frac{\partial N_e}{\partial t} &= \alpha_i |W_e| N_e - \alpha_a |W_e| N_e - \alpha_r N_p N_e - \frac{\partial (W_e N_e)}{\partial x} + \frac{\partial^2 (D_e N_e)}{\partial x^2} \\
\frac{\partial N_p}{\partial t} &= \alpha_i |W_e| N_e - \alpha_r N_e N_p - \alpha_r N_n N_p - \frac{\partial (W_p N_p)}{\partial x} \\
\frac{\partial N_n}{\partial t} &= \alpha_a |W_e| N_e - \alpha_r N_p N_n - \frac{\partial (W_n N_n)}{\partial x}
\end{aligned} \tag{6.2}$$

where W_e, W_p and W_n are the drift velocities for the three types of charged particle, and α_i, α_a, α_r and D_e are functions representing the ionization, attachment, recombination and electron diffusion.

The boundary conditions for this problem are also influenced by a variety of processes. At the anode, the electrons are totally absorbed and their number density approaches zero (Braglia and Lowke, 1979), giving effectively

$$N_e(d,t) = 0 \quad , \qquad for\ all\ t \tag{6.3}$$

but at the cathode, secondary electrons may be produced by either photons or ions, giving

$$N_e(0,t) = N_e^p(0,t) + N_e^i(0,t) \qquad . \tag{6.4}$$

The number of secondary electrons released by photons, N_e^p, can be calculated from the electron drift velocity, W_e, the number of electrons, N_e and the radial extent of the charge. The number of secondary electrons released by ions is

$$N_e^i(0,t) = \frac{e_i N_i(0,t)\,|W_i(0,t)|}{W_e(0,t)} \tag{6.5}$$

where e_i is the ion-secondary-emission coefficient N_i is the number of ions and W_i is the drift velocity of those ions. The boundary condition at the cathode for the negative ions is

$$N_n(0,t) = 0 \quad , \qquad for\ all\ t. \tag{6.6}$$

The nonlinearity in this system of equations comes from the dependence of most of the coefficients in Eq. (6.2) on the ratio of the electric field, E to the neutral gas number density, N. Only the recombination coefficient, α_r, is taken to be a constant. The remaining coefficients all involve exponential functions of E/N. Examples of the precise form of these functions is given in Morrow (1985, Appendix A).

Morrow (1985) also discussed the important spatial features of the numerical solution. From a computational stand point, the main features of interest are, the presence of a boundary layer in the electron density, and nearby, a low field region where charge accumulates. These regions are shown in Fig. 6.1(a-c), from Morrow (1985). In order for the numerical results to be physically reasonable, the numerical algorithm must satisfy the constraints,

1. the number of each species of charged particle must be non-negative,

2. there should be no overshooting of the electric field in the low field region, and

3. there should be no spurious accumulation of charged particles.

The boundary layer or "cathode sheath" corresponds to a region of rapid decrease in field strength, and hence a rapid change in electron drift velocities. This feature poses some difficulty for numerical models and can cause the numerical solution to produce non-physical results. It should also be noted that these features are all interconnected by the non-linearity of the solution. For example, spurious accumulations of charge will introduce spurious disturbances in the electric field and vice versa. This feedback loop means that any numerical model of gas discharges must not produce any numerical oscillations.

Given that the cathode sheath and the low field region are of such importance to the entire process, as well as causing many of the numerical difficulties, it is worth discussing these regions in a little more detail. The electrons leave the cathode at high velocity, and must then slow down to a small fraction of their initial velocity as they pass from the cathode sheath to the low field region. The high velocity region, near the cathode, must contain relatively little structure in the charge density, because the electrons flow across this area with such high velocity that there is no mechanism for any detailed structure to form. On the other hand, the low field region is characterized by the continual arrival of electrons coupled with the relatively slow motion of any electrons already there. Such an area is inevitably going to exhibit considerable structure. This structure is important, as it is from here that the discharge emanates.

This transition from a region of high velocities and weaker gradients (less structure) to one of low velocities and strong gradients (greater structure) is typical of problems involving particles moving across decreasing velocity fields. The large variation in velocity and gradients can be partly catered for by using a finer mesh in both the boundary layer and low field regions. By choosing a suitable grid, the range

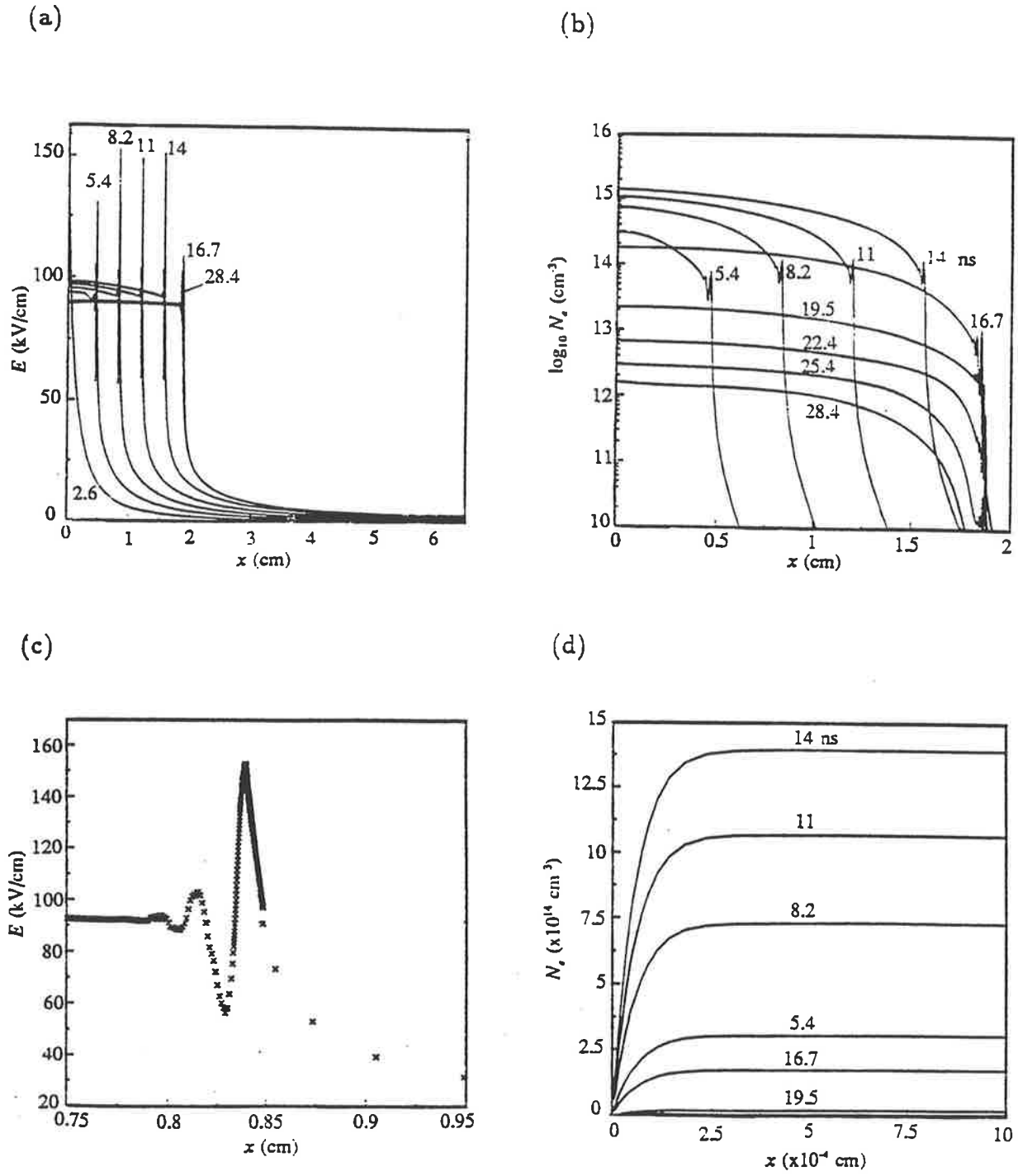

Figure 6.1: Illustration of spatial distribution of charge in a coronal discharge. The cathode is at $x = 6.5$ cm and the anode at $x = 0$. The diagrams show (a) the field across the entire domain, (b) the electron number density over the domain, (c) detail of the electric field in the region of high gradients and (d) the electron distribution near the anode. Note that the sharp gradients and oscillations are well resolved, and that the oscillations are physical rather than numerical.

of Courant numbers involved may be reduced, and any detailed structure may be suitably resolved. The extension of FCT schemes to a variable mesh as described by Morrow and Cram (1983) is readily extended to implicit FCT schemes and will not be discussed any further.

Any mesh refinement has significant impact on the modelling of the diffusion term, $\partial^2(D_e N_e)/\partial x^2$. The stability of a finite difference approximation to this term is governed by the parameter $s = D_e \Delta t/(\Delta x)^2$, in much the same way as the Courant number governs the stability of the approximations to the advective terms. It has been found that if a suitably fine mesh is used to resolve the velocity and density gradients then the stability (and hence time-step) is determined by the value of s, rather than the Courant number. This means that either very small time-steps must be taken or implicit differencing of the diffusion term is necessary. The overhead of using an implicit scheme turns out to be far less than the CPU time required by an explicit differencing of diffusion. Since there is negligible cost in also including advection implicitly, it seemed profitable to investigate the performance of smooth, fully implicit schemes for modelling advection. In Chapter Three it was seen that the accuracy of implicit schemes is predominantly dictated by the sharpness of the density gradients, with only weak dependence on Courant number for values of c less than about three. It should be noted that this range of values for the Courant number can be extended if the high velocity regions also correspond to the regions of high Courant number since the gradients in these regions are generally weaker. As such, it was considered that smooth implicit schemes showed some promise for application to this type of problem. The next sections will demonstrate that this is the case, by investigating the performance of a variety of schemes in spatially varying velocity fields.

6.2 Decreasing Velocity Fields

The performance of different numerical methods in modelling the motion of electrons in the vicinity of a cathode, was given by Morrow (1981). This comparison, was based on the advection of a square wave in a velocity field that decreased linearly with distance, x. While these tests demonstrated the usefulness of using FCT schemes to model advection in decreasing velocity fields, the relative performance of the different schemes was not quantified and only qualitative comparisons were provided. By using quantitative error measures, this section will provide a more detailed description of the relative accuracy of a selection of smooth difference operators.

The test used here consists of advecting a semi-elliptical profile of unit height and half-width of 0.1. The pulse is initially centred about $x = 0.2$ and specified on a uniform grid, with $\Delta x = 0.01$. The velocity field linearly decreases to approximately 1% of its value at the upstream boundary. That is, the advection equation is modelled, with $w(x) = 1.01 - x$ for x in the interval $[0, 1]$, and an initial condition of

$$\rho_0(x) = \begin{cases} 10\sqrt{(0.1)^2 - (x - 0.2)^2} & 0.1 \leq x \leq 0.3 \\ 0 & \text{otherwise} . \end{cases} \tag{6.7}$$

The exact solution to this problem is

$$\rho(x,t) = \rho_0 \left((x - 1.01)e^t + 1.01 \right) e^t \quad . \tag{6.8}$$

The values, $\rho(0,t)$ are specified, and the downstream boundary is free. If a downstream boundary is required by an algorithm, to close the system of equations, then outgoing fluxes are calculated from the low order scheme. For the case $c \leq 0.9$ and $J = 100$, 60 timesteps were used, with the number of timesteps in the other cases varying according to changes in the maximum value of c and the grid spacing, so that the final time is constant for all tests. A semi-elliptical profile was used in preference to either a Gaussian pulse or square wave for two reasons. Firstly, the wave form becomes progressively thinner and taller as it is advected, and in the case

of a Gaussian pulse, the numerical solution at a time $t = T$ can become so thin that the bulk of the waveform is only a few grid-points wide. This means that the error measures for the different schemes are dominated by just a few values and as such are not particularly meaningful. This is highlighted by the fact that as the time interval of integration is varied, the relative accuracy of the different schemes also varies, and by either increasing or decreasing the length of time of the integration (provided T is large enough) it is possible to show any one of a number of schemes in a favourable light. This is not surprising, as it was shown in Chapters Two and Three that no finite difference schemes accurately model the advection of short wavelength Fourier components. Hence, if the bulk of detail in an exact solution is of this scale then all schemes will perform poorly, and intercomparisons will be of little use. By using a semi-elliptical profile the final solution is better resolved and so the error measures provide more representative comparisons between the different schemes. An alternative is to use a Gaussian pulse that is initially very broad, but this has the disadvantage that for most of the numerical integration, the waveform is extremely well resolved, and all schemes perform well and little information is gained by comparing different schemes.

The second reason for using a semi-elliptical pulse is that there is still some detail that needs to be captured, besides the sharp gradients. A square wave also contains the sharp gradients, but there is no other detail to model. It is important that schemes not only capture sharp gradients, but that any nearby, more gradual changes are also well modelled. For these reasons it is not surprising that when results for many different runs are examined, in which the width and shape of the profile and the time of integration are all varied, the comparisons between the different schemes are generally similar to those provided here. That is, the advection of a semi-elliptical test provides the most representative basis for the comparison of different schemes

The RMS errors for selected schemes are presented in Table 6.1. The maximum

Table 6.1 : Comparisons of improvement in R.M.S. error ($\times 100$) due to increasing resolution. The test problem involves a semi-elliptical pulse as the initial condition in a linearly decreasing velocity field. The CPU Times are relative to First Order Upwinding with $c \leq 0.9$.

Scheme	max c	$J = 100$	$J = 200$	$J = 400$	$J = 800$
1^{st} order Upwind	0.45	33.5	26.07	18.58	14.53
	0.90	27.9	21.76	15.12	11.88
Lax-Wendroff + FCT	0.45	17.1	12.86	7.45	5.57
	0.90	16.1	12.05	6.94	5.18
van Leer (II)	0.45	16.4	11.55	6.17	4.45
	0.90	14.7	10.24	5.29	3.81
Phoenical LPE SHASTA	0.45	13.0	8.80	4.99	3.57
Implicit LPE SHASTA	0.45	12.1	7.95	4.34	3.07
UW5FCT	0.45	13.6	9.21	5.11	3.62
	0.90	12.6	8.60	4.72	3.38
CN4FCT	0.45	12.1	7.67	4.63	3.19
	0.90	12.5	7.65	4.67	3.16
	3.60	13.3	8.15	4.87	3.36
	5.40	15.7	11.18	6.59	4.56

value of c is given by $w_{-\Delta x/2}\Delta t/\Delta x$ and may be considered to be a scaled time-step since all calculations were performed on the same grid and with the same velocity field. Again a range of grid spacings were used, to show the effects of changing resolution on the results. This is also useful for checking the validity of the error measures, as alternative error measures could have been used instead of RMS error. These other error measures, however, provide little extra information, with the relative overall performance of the schemes being unaffected. Furthermore, some of the other error measures seem to occasionally give anomalous values, in that it is possible for any particular scheme to have either a very poor or a very good value for a particular resolution. In this case it is evident that such a value is not truly representative of the overall performance of the scheme. These anomalous values vary according to which scheme, resolution and error measure is used. This problem with anomalous values does not seem to appear in the RMS error values given in Table 6.1.

It was shown in Morrow (1981) that Phoenical LPE SHASTA was suitable for

modelling the motion of electrons near a cathode, whereas the schemes such as first order upwinding and LWFCT produced too much distortion of the initial pulse. The error measures for these schemes have been included as benchmarks for acceptable and unacceptable errors respectively. As can be seen from Table 6.1, the "high order" FCT schemes such as Implicit LPE SHASTA and CN4FCT provide very accurate results, better in fact than Phoenical LPE SHASTA. The scheme, UW5FCT, is probably best considered to be comparable to Phoenical LPE SHASTA, in that UW5FCT can use a larger time-step and is then more accurate the Phoenical LPE SHASTA, but if the same time-step is used then it is less accurate. The scheme van Leer (II) is definitely better than schemes such as LWFCT but not as accurate as the other schemes.

The advantage of using CN4FCT is that high accuracy may be retained over a wider range of time-steps. In comparison with the SHASTA FCT schemes, the time-step may be increased five or six times, without seriously degrading the accuracy of the results. Furthermore, since the electron diffusion is modelled implicitly, there is minimal overhead in also using an implicit scheme to model electron advection.

It is also worth noting that CN4FCT is still more accurate than UW5FCT. This can be explained by the fact that although fifth order upwinding approximates the phase speed of the Fourier components more accurately than the fourth order CTCS scheme, this only succeeds in reducing the error due to dispersion, i.e the errors associated with the spurious oscillations. These errors are virtually removed by the flux correction process anyway. Fifth order upwinding does, however, have an effect on the amplitude of these components and this is not taken into account by the flux correction process. It seems therefore, that the advantage in using a less dispersive high order scheme is only slight in comparison with the use of a less diffusive high order scheme, since flux correction reduces dispersive errors but exacerbates diffusive errors.

6.3 Increasing Velocity Fields

The case where the velocity, $w(x)$, is an increasing function, is also of some interest. The exact solution corresponding to an initial condition of $\rho_0(x)$, is given by

$$\rho(x,t) = \rho_0(xe^{-t})e^{-t} \tag{6.9}$$

The analytic solution now describes a waveform that becomes broader and flatter. As such, this case is of less value for comparing the general performance of numerical schemes, since any strong gradients in the initial conditions gradually weaken as the waveform is advected which enhances the performance of most schemes. That is not to say, however, that motion in increasing velocity fields is of no interest, for it was noted above that an important part of modelling gas discharges is accurately capturing the absorption particles at outflow boundaries. In order to verify that schemes such as CN4FCT are capable of faithfully reproducing this, comparisons are made with Phoenical LPE, and Implicit LPE SHASTA.

The results of modelling a Gaussian pulse of variance 1/400, starting at $x = 0.6$, for 120 time steps ($c_{\max} = 0.45$) for CN4FCT, Phoenical and Implicit LPE SHASTA schemes are shown in Fig. 6.2(a-c). The number of time-steps were chosen so that at the final time step, the rear edge of the pulse was passing through the boundary. By examining the pulse at a variety of times, it was found that during this particular phase of the solution, any errors associated with the boundary condition were most pronounced. The boundary values are handled in the same manner as for the decreasing velocity case. A Gaussian pulse is used rather than a semi-ellipse pulse (as in the decreasing velocity experiments) since now that the profile becomes broader and flatter, it is a more rigourous test to start with a thin pulse, such as a Gaussian, rather than a a wider pulse such as a semi-ellipse.

The difference between the diagram for CN4FCT with $c_{\max} = 0.9$, using 60 time steps and that for 120 time steps with $c_{\max} = 0.45$ is beyond the resolution of the graphs and so the larger time step case is not presented. The RMS errors for these examples are 4.4×10^{-3} for CN4FCT, 4.9×10^{-3} for Implicit LPE SHASTA and

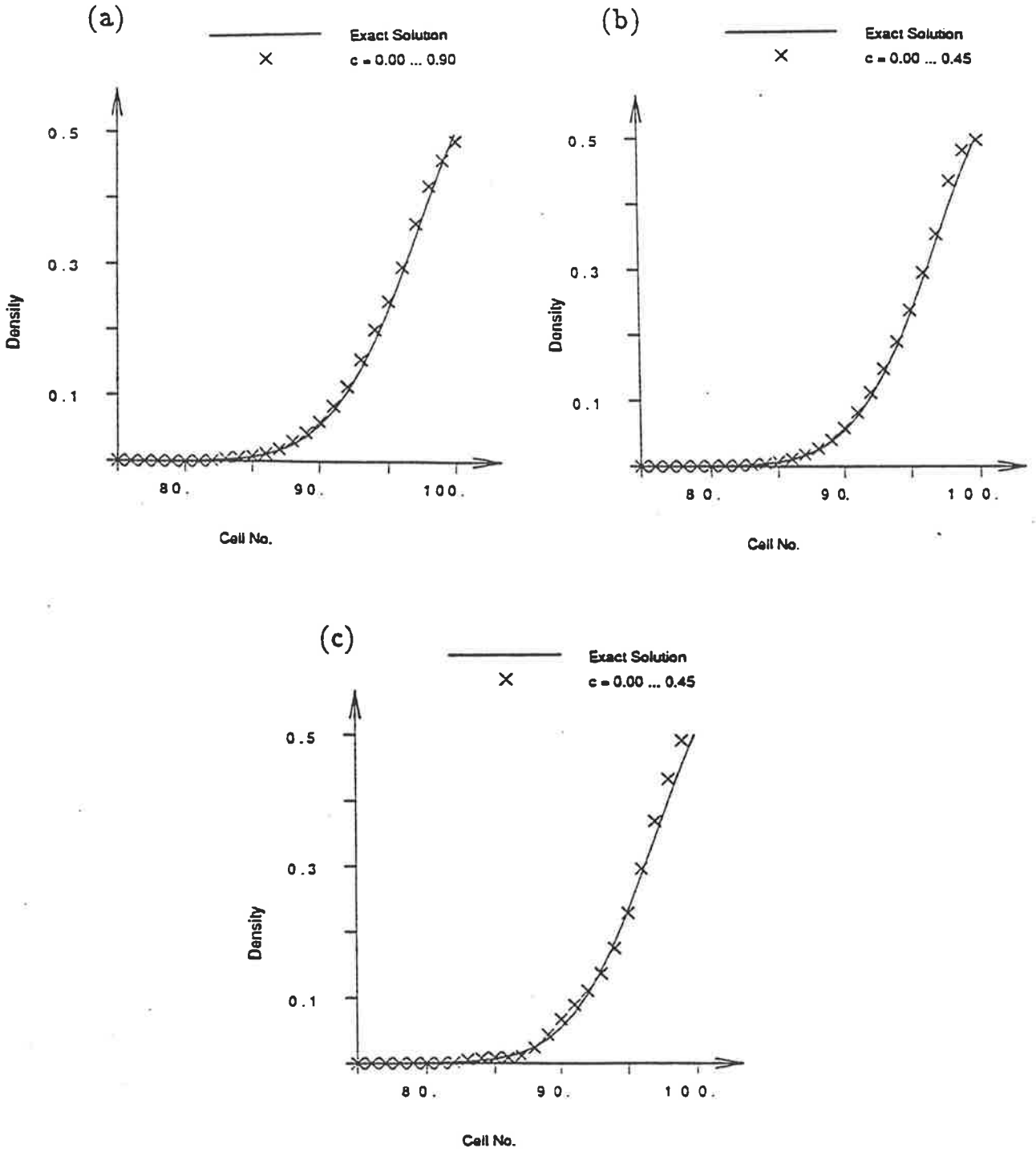

Figure 6.2: Illustration of the flow of material across an outflow boundary using (a) CN4FCT, (b) Implicit LPE SHASTA, and (c) Phoenical LPE SAHSTA.

6.0×10^{-3} for Phoenical LPE SHASTA. Comparisons between CN4FCT and either of the LPE SHASTA methods, using Fig. 6.2(a-c) and the RMS errors shows that the CN4FCT solution is both smooth and accurate.

This simple experiment demonstrates that the use of an implicit scheme need not have any effect on how smoothly an FCT algorithm advects material across an outflow boundary. The downstream boundary condition for the implicit fourth order scheme was treated very simply by using first order upwinding fluxes to represent the fluxes out of the last grid cell. The equation for ρ_J^{n+1} then involves values at the grid points x_{J-1} and x_J only, and the tri-diagonal system of equations will possess a unique solution. Any mismatch due to the difference between the fluxes of the implicit scheme and first order upwinding is resolved by the flux correction process. This approach is preferred to simply setting the value of ρ_J^{n+1} to be that of first order upwinding as the former approach causes the oscillations in the high order solution to reflect off the downstream boundary. In the latter case, where the value is specified, rather than the flux, the flux correction process will still eliminate oscillations. It was shown in Chapter 5, however, that the best results are obtained by minimizing the amount of work the flux limiter must do, i.e. by minimizing the use of the low order solution. The introduction of further oscillations into the high order solution can do nothing but degrade the final result.

6.4 Conclusion

In this chapter, the modelling of gas discharges was chosen as an example for a discussion on the development of a numerical method for modelling complex advective processes. The major problems as far as numerical modelling is concerned, namely the region of sudden speed change, the region of complex structure and the smooth transport of material both in and out of the solution domain occur not only in gas discharges, but in a wide variety of applications. The advantage of using gas discharges as an example is the sensitivity of this problem to errors in the modelling

of any of these features.

The implicit FCT algorithm CN4FCT, has been incorporated into a model of gas discharges and was found to provide a very fast and accurate solution. In a way, this provides the final proof of the validity of Implicit FCT. The success in accurately capturing the evolution of the electric field demonstrates not only that CN4FCT provides smooth results in idealized cases, but that this carries through to more complex examples. This also validates the systematic derivation of CN4FCT, from examination of the wave propagation parameters of high order implicit schemes, through to the derivation of conditions under which implicit finite difference schemes produce smooth results.

As far as modelling the development of the spatial distribution of electric charge, the three main components are the cathode sheath, the low field region and the absorption of ions. This is not to trivialize the calculation of the electric field from the distribution of charge or indeed any of the other calculations involved in gas discharge models, but the topic under discussion is the modelling of transport processes, and as such the first three components are the most relevant. The accuracy of CN4FCT in modelling each of these components was discussed in turn. That CN4FCT should be suitable for use in the low field region follows from the discussions in Chapter Five. Within this region it is important that the solution should accurately capture the sharp gradients, and at the same time contain no spurious oscillations. It was shown in Chapter Five that CN4FCT is capable of maintaining sharp gradients within a smooth solution for the case of constant velocity. In a full model of gas discharges, these sharp gradients imply that diffusion also becomes important. The presence of physical diffusion only enhances the accuracy of most finite difference schemes, CN4FCT included, and so it is only mentioned in passing. The ability of CN4FCT to adequately resolve the cathode sheath was demonstrated by considering the simplified case of linearly decreasing velocity. It was shown that CN4FCT accurately captures the growth of a peak, again while maintaining a

smooth numerical solution. Similar tests were used by Morrow (1981) to compare different schemes. The tests used here involved a more stringent initial condition, and the provision of error measures provided a quantitative measure of performance.

The final aspect, the flow of material across an outflow boundary, while not difficult, is still important. By use of appropriate fluxes at a downstream boundary, the solution of the implicit linear equations becomes unique and allows for the free movement of material out of the solution domain.

In problems such as this, where physical diffusion can become significant over part of the domain in which the fine structure is important, an algorithm such as CN4FCT becomes even more efficient relative to explicit schemes. This is because the overhead in calculating the high order, implicit advective fluxes becomes negligible, due to the need for implicit differencing of the electron diffusion. The diffusion could be differenced explicitly, but because of the fine mesh in the region where diffusion is significant, the time step for this part of the calculation becomes very small. This results in schemes based on implicit differencing of the electron diffusion requiring considerably less CPU time overall. By incorporating the calculation of the advective fluxes with that of the diffusive fluxes, the overhead of using implicit advective fluxes becomes insignificant relative to the calculation of explicit advective fluxes.

The second method of increasing the efficiency of CN4FCT is by using the fact that the fourth order CTCS scheme for modelling the full transport equation becomes positive definite when

$$\left|\frac{1}{2}c_{j+\frac{1}{2}}\right| \leq \left|s_{j+\frac{1}{2}}\right| \leq 1 \quad . \tag{6.10}$$

This was shown in Section 5.3. In such regions, where this is true for two or more adjacent grid cells, the flux correction process may be switched off. The exact gains produced by this reduction in the number of flux correction calculations depends very strongly on the nature of the problem, but they can be significant. This extension must be used with some care as it is possible in some very non-linear problems for

the small oscillatory modes discussed in Section 5.4 to produce some measurable oscillations.

Such behaviour also demonstrates the advantage of performing the advection and diffusion calculations together, rather than as two separate steps as would be the case in a split operator scheme. As was seen in Section 4.1 the addition of a diffusive step to an advective calculation is equivalent to applying a linear filter to the advected solution, and that this does not guarantee non-oscillatory results. Thus, at some stage of the calculations in a split operator method some form of smoothing must still be performed at every time-step and at every grid point. The statement holds, irrespective of the order of the advection and diffusion calculations within each time step since all linear high order advective schemes are capable of producing numerical oscillations as was shown in Chapters Two and Three.

Apart from reiterating that CN4FCT has been very successfully included in a full non-linear model of gas dynamics by Morrow, and pointing out that this success is as expected from its behaviour in idealized systems and the systematic approach used to derive the algorithm, no more of the results will be discussed. Suffice to say that the speed and accuracy of the CN4FCT method confirms its suitability for modelling transport processes in complicated flow fields and should be applicable to a wide variety of such problems.

Chapter 7

Conclusion

This thesis has demonstrated that high order implicit finite difference schemes can be successfully incorporated into a model involving complicated advective processes. These schemes are frequently overlooked in discussions of such models due to a perceived inability to produce smooth results. It has been shown that, on the contrary, such schemes can be used in such a way as to provide positive definite algorithms which are both accurate and efficient in comparison with similar explicit algorithms.

The accuracy and efficiency of implicit schemes as applied to simple, linear problems have been discussed at length within the literature. The greater task has been to adapt these schemes for use in non-linear problems, where it is often crucial to maintain positivity. In order to do this, it was necessary to provide a detailed analysis of the differences between numerical solutions obtained from implicit schemes, as opposed to those obtained from explicit schemes. This involved analyzing the behaviour of different schemes as applied to the constant velocity case. The comparisons between the different high order schemes were also used in deciding which schemes could be used most profitably in an FCT algorithm to provide a robust and accurate model for non-linear advection.

It has been a widely held view that the numerical oscillations produced by implicit schemes are inherently different from those of explicit schemes and that they are therefore unsuitable for use within local adjustment algorithms, such as FCT. The

basic difference is, that after one time step explicit schemes will produce only local oscillations, but implicit schemes will produce global oscillations. It was noted that, in this respect, implicit schemes could be considered to be similar to explicit schemes with very wide stencils, and thus, techniques that only consider local adjustments may still be applicable.

By examining the sources of the oscillations produced by implicit difference schemes, it was shown that some of the problems encountered by other workers within the field could be attributed to errors in the calculation of the low order fluxes. By applying suitable conditions to the calculation of these fluxes, it was then demonstrated that implicit schemes could be incorporated successfully into positive definite algorithms that only use local adjustments to produce the final, smooth solution.

Chapter Two discussed approaches for comparing the performance of a variety of schemes, using the wave propagation parameters, the modified equivalent equation and some numerical tests. The equivalence between the order of the modified equivalent equation and the accuracy of the wave propagation parameters was demonstrated in theory and practice. The theoretical results established the equivalence in the limit as the grid-spacing approaches zero ($N_\lambda \rightarrow \infty$). The numerical experiments demonstrated that this equivalence also holds for finite N_λ. These methods of comparing schemes were shown to provide an appropriate measure of the overall performance of a selection of explicit schemes.

In Chapter Three it was shown how the modified equivalent equation can also be used in the development of schemes, not just as a diagnostic tool. Due to the correspondence between the modified equivalent equation and overall accuracy, a scheme that is constructed so as to maximize the order of the modified equivalent equation will also be very accurate. Implicit differencing was chosen as it has two advantages. Firstly, very high order schemes can be obtained without recourse to exceptionally wide stencils and secondly it is possible to guarantee unitary amplitude

reponse. While it is acknowledged that this second feature is a drawback when implicit schemes are used by themselves to advect sharp profiles, it was shown in Chapter Five that when used in conjunction with some smoothing techniques, such an amplitude response was an advantage. It was also demonstrated that schemes with no damping are particularly well suited for the advection of smooth profiles, such as a Gaussian pulse.

It was also shown in Chapters Two and Three that high order schemes produce results of a specified accuracy faster than low order schemes and for this reason they can be considered to be more efficient. One of the major difficulties in quantifying efficiency is how to match gains in accuracy and increased CPU time. This was overcome by comparing the CPU time needed by different schemes to obtain similar levels of accuracy which automatically provides the correct weighting of CPU time relative to accuracy. Being able to attain a certain level of accuracy with lower resolution is also a significant advantage when dealing with multi-dimensional problems, for then it is often the capacity of the machine to store large arrays that is the major constraint on the problem, rather than the amount of CPU time required. In such cases high order schemes are clearly advantageous.

The comparisons between the high order schemes obtained via the modified equivalent equation approach and those based on Padé approximants demonstrated that high levels of accuracy are attained only when the modified equivalent equation is of high order in Δx and Δt. Khaliq and Twizell's schemes are of high order in Δt but only second order in Δx. The error measures associated with these methods were much closer to those of other second order schemes than the higher order schemes NS4 and NS5. This is due to the errors associated with the spatial discretization dominating the solution and persist despite the use of either higher order Padé approximants or the suggested extrapolations. Using the modified equivalent equation directly in the development of the difference scheme ensures that there is no such inconsistency between the spatial and temporal discretizations.

Implicit schemes also possess the additional advantage of unconditional stability in the von Neumann sense. A comparison between the CPU times of high order implicit and explicit schemes shows that the timestep need only be increased by a factor of two or so before the implicit schemes become as fast or faster than the explicit schemes. Since for some schemes, the timestep can be almost trebled before the solution begins to significantly deteriorate, the implicit schemes can be considered to be highly efficient.

The unconditional von Neumann stability of implicit schemes does not however guarantee that the algorithm will be stable, due to the presence of other instabilities. These instabilities are often ignored in discussions of implicit schemes but are still important because an instability of any form will eventually cause a numerical model to give physically unreasonable results. An example of ignoring other instabilites was associated with Khaliq and Twizell's schemes, where the individual steps were stable, but the extrapolation to provide high order time discretization was not necessarily stable. This was not discussed in their paper, but the instability can be seen to appear reasonably quickly in the case of the (2,0) scheme. Other instabilities are associated with matrix inversion by schemes based on the Thomas algorithm or with the marching of solutions across the domain. It was shown that the fourth order CTCS scheme can be used in such a way as to guarantee stable results for any size timestep. The unconditional stability and overall accuracy of the fourth order CTCS scheme makes for a very efficient method of modelling advection (although the accuracy of the solution deteriorates as the timestep becomes large).

A further complication with using high order schemes arises when boundary conditions are considered. The brief discussion in Chapter Three demonstrated that interpolation was inappropriate and that using the fluxes from first order upwinding was as good as any of the alternatives. Using these fluxes has two further advantages over other schemes namely simplicity and the instantaneous response to sudden changes in boundary values. Other schemes, which use values at interior gridpoints

necessarily cause a lag in the propagation of information from the boundary to the interior of the solution domain.

Another problem with high order implicit finite difference schemes is the production of high frequency noise by sharp gradients in the advected profile. This was highlighted by experiments involving a semi-elliptical waveform. It was noted however, that the overall accuracy of these schemes was still quite high despite the oscillations. The sensitivity of the high order implicit schemes to sharp gradients resulted in a reduced improvement in accuracy over explicit schemes of similar order. The high order implicit schemes were now comparable to third or fifth order upwinding. The point was made that since all high order schemes (explicit or implicit) produce oscillations and as such require the use of some form of smoothing technique in order to produce a positive definite algorithm, the high frequency oscillations seen in implicit solutions are not necessarily a major problem.

A variety of smoothing techniques were discussed in Chapter Four where it was shown that linear filtering is inappropriate, since diffusion is continuously being added to the numerical solution, and nothing is done about minimizing the residual diffusion. Various other methods were considered, some being more successful than others. The most successful were those that used a high order solution over most of the solution domain, but reverted to a low order solution in regions of high gradients. These included schemes such as van Leer's extension of Fromm's scheme, Harten's Self Adjusting Hybrid Scheme and the two FCT algorithms, Phoenical LPE SHASTA and Implicit LPE FCT. The remaining non-linear techniques were not considered as appropriate for combination with the high order implicit schemes since Smolarkiewicz's scheme admitted oscillatory overshoots and the TVD schemes of Yee were seen to be not only very CPU intensive but also rather inaccurate. Of the more successful schemes, FCT is the most general in that it makes the least assumptions about the nature of the high and low order solutions.

In the discussions on the behaviour of these methods two important points

emerged. Firstly, that non-linear smoothing techniques, such as FCT, are very effective at removing oscillations with wavelength of about three grid points. Secondly, sharp profiles rapidly become smooth and so the the problem of advecting a sharp profile degenerates to one of advecting a smooth approximation to the original profile. In the light of these observations it is apparent that high order implicit schemes should be well suited to use within a general smoothing technique.

If implicit schemes were to be used within an FCT algorithm an answer had to be provided to Boris and Book's statements about the incompatibility of using a local flux limiter to handle the global oscillations present in implicit high order solutions. A reanalysis of REVFCT demonstrated that this algorithm is less than ideal for passive advection, for even in the case of constant velocity, there are still oscillations in the results. Although these oscillations are small, they are a genuine product of the numerical algorithm and not due to round-off error. This difficulty was originally attributed to numerical precursors, but it was shown here that the low order solution did not always produce positive definite results which violates the basic assumption of FCT. For the case of constant velocity advection, the numerical oscillations in the low order solution will also appear in the corrected solution until removed by residual diffusion. In more complex applications, however, it is quite feasible that these oscillations could produce the problems encountered by Boris and Book.

To demonstrate that it was only the low order solution producing these oscillations it was necessary to alter the calculation of the low order solution of REVFCT, and retain the remainder of the algorithm. Since the corrected form of REVFCT was positive definite, then it seems that it was the low order solution producing the oscillations rather than any problems associated with using a local flux limiter to control the global oscillations produced by implicit schemes. The correction to REVFCT required the derivation of a set of conditions under which implicit finite difference schemes were positive definite. These constraints not only meant that a

different low order scheme had to be used, but also the method of calculation had to be changed. The original method used to calculate a low order solution from a high order implicit solution is only an approximation to calculating the low order scheme directly. This approximation need not preserve the positive definite nature of the low order scheme. With these corrections to both the low order scheme and the method of calculation of the low order solution, the oscillations present in the REVFCT corrected solution disappeared. This led to two algorithms that may be considered as corrections to REVFCT, namely CN2POS and REVPOS. The calculations involved in both of these schemes closely followed those of REVFCT apart from those associated with the low order solution. The results from these corrected forms of REVFCT were not particularly encouraging, as far as obtaining an accurate, positive definite finite difference scheme was concerned. In the context of using implicit high order solutions within an FCT algorithm, however, they were very encouraging as they demonstrated that it was possible to use FCT and high order implicit schemes to produce a positive definite result, provided that sufficient care was taken with the calculation of the anti-diffusive fluxes.

The main reason for the significant loss of accuracy in going from REVFCT to either REVPOS or CN2POS was the excessive diffusion present in the low order schemes, as was evident from the corresponding modified equivalent equations. To overcome this problem, Zalesak's formulation of FCT was used which allows complete freedom in the choice of the low order scheme.

The remaining problem was to find an appropriate low order solution. Godunov (1959) showed that first order upwinding is the highest order, positive definite explicit finite difference scheme. By expanding a general difference scheme on a (3,3) stencil and applying the constraints of Section 5.2, it was found that first order upwinding was also the least diffusive scheme on this stencil. Hence, this scheme was used as the low order solution, giving the algorithm CN4FCT. This was shown to be positive definite and of comparable accuracy to REVFCT. The point was also made

that this scheme was superior in efficiency to the other positive definite algorithms due to its overall accuracy and unconditional stability.

In a similar fashion a family of implicit FCT schemes were obtained and these were compared to ascertain the most accurate and efficient algorithm. There was a substantial gain in accuracy when fourth (or higher) order schemes were used to calculate the high order solution, as opposed to using second order schemes, such as the Lax Wendroff and Crank Nicolson schemes. The gain in accuracy from using fourth order CTCS rather than a lower order explicit scheme more than compensates for the computational expense of using an implicit scheme Comparisons were also made using fifth order upwinding, NS4 (sixth order) and NS5 (eighth order). Substituting fifth order upwinding as the high order solution did little if anything to enhance the accuracy of CN4FCT. This can be explained by the damping of the explicit scheme reinforcing the diffusion of the flux limiting process, which will not occur when a fully centred finite difference scheme is used. The higher order schemes, NS4 and NS5 are more efficient for a given Courant number, but the limited timestep (due to the stability of the matrix inversion) means these schemes are in general, no more efficient than CN4FCT.

One of the aims of this thesis was to develop a positive definite advective algorithm for use in a gas discharge model. It was because of this that the discussions have concentrated on the abilities of different schemes to model the advection of pulses. The comparisons suggested that the most suitable algorithm was CN4FCT, which was successfully included in the model described in Chapter Six. While the results of the CN4FCT version of this model appeared reasonable, when compared with other numerical and idealized models it was not straightforward to determine whether or not CN4FCT was performing as expected. To this end the three regions that could cause most of the numerical difficulties were examined more closely.

The modelling difficulties associated with the low field region were one of the major reasons for developing implicit FCT algorithms. For it is here that spurious

oscillations cause most difficulties and it also the region where implicit differencing of the diffusion term is required due to the fine grid spacing necessary to resolve the sharp gradients which can form. Since the diffusive fluxes were to be calculated implicitly it was decided also to try to calculate the advective fluxes implicitly. The reason was that if the two steps are performed separately, then the physical diffusion is added to any numerical diffusion. Whereas, if the two steps are performed simultaneously, the high order advected solution is necessarily smoother, helping to reduce the residual diffusion in the system. The results presented in Chapter Six, demonstrate that in the low field region, the CN4FCT algorithm performs acceptably well.

Other areas of concern were the regions where the electrons rapidly slow down, and where particles leave the domain in an increasing velocity field. The performance of CN4FCT in these regions was quantified by using an idealized situation where the velocity fields varied linearly in space. Comparisons were made between CN4FCT and other schemes, and it was seen that the comparisons made in the constant velocity case were still valid when the velocity was allowed to vary.

The success of implicit FCT algorithms (and CN4FCT in particular) in providing robust and accurate schemes for modelling advective processes opens up many avenues for further research. More work needs to be done on the discretization of non-linear terms, within the implicit calculations. In all of the experiments discussed here, the velocity has been determined before the advective calculations were performed, although there was still a non-linear interaction between the advected fields and the velocity fields in the gas discharge model. Also, other adjustment strategies could be examined, as the major source of errors in the linear test-problems was the inability of the flux-limiter to adequately preserve the peak. It has been shown here that local adjustment strategies can be adapted for use with implicit equations, and so work done on explicit algorithms should carry over to implicit algorithms. It would also be interesting to incorporate the conditions for implicit schemes to

be positive definite into Grandjouan's work on modified equations and FCT. This may go some way to reducing the errors near the peak, although these errors are dominated by the inability of the flux limiter to preserve peaks. Nevertheless, the algorithms developed here have been shown to be a significant improvement over many existing schemes.

Bibliography

Book, D.L., Boris, J.P. and Hain, K. (1975), "Flux Corrected Transport II. Generalizations of the Method", *Journal of Computational Physics*, Vol. 18, pp. 248-283.

Boris, J.P. and Book, D.L. (1973), "Flux Corrected Transport I. SHASTA, A Fluid Transport Algorithm That Works.", *Journal of Computational Physics*, Vol. 11, pp. 38-60.

Boris, J.P. and Book, D.L. (1976a), "Solution of Continuity Equations by the Method of Flux Corrected Transport", *Methods in Computational Physics*, Vol. 16, pp. 85-129.

Boris, J.P. and Book, D.L. (1976b), "Flux Corrected Transport III. Minimal-Error FCT Algorithms", *Journal of Computational Physics*, Vol. 20, pp. 397-431.

Braglia, G.L. and Lowke, J.J (1979), "Comparisons of Monte Carlo and Boltzmann Calculations of Electron Diffusion to Absorbing Electrodes", *Journal of Physics D.: Applied Physics*, Vol. 12, p. 1831

Cathers, B. and O'Connor, B.A. (1985), "The Group Velocity of Some Numerical Schemes", *International Journal for Numerical Methods in Fluids*, Vol. 5, pp. 201-224.

Cunge, J.A., Holly, F.M., Jr., and Verwey, A. (1980), Practical Aspects of Compu-

tational River Hydraulics, Pitman.

Godunov, S.K. (1959), "Finite Difference Method for Numerical Computation of Discontinuous Solutions of the equations of Fluid Dynamics", *Mathematicheskii Sbornik*, Vol. 47, pp. 271-306.

Grandjouan, N. (1990), "The Modified Equation Approach to Flux-Corrected Transport", *Journal of Computational Physics*, Vol. 91, pp. 424-440.

Harten, A. (1978), "The Artificial Compression Method for Computation of Shocks and Contact Discontinuities III. Self Adjusting Hybrid Schemes", *Mathematics of Computation* , Vol. 32, pp. 363-389.

Harten, A. (1983), "A High Resolution Scheme for the Computation of Weak Solutions of Hyperbolic Conservation Laws", *Journal of Computational Physics*, Vol. 49, pp. 357-393.

Harten, A. (1984), "On a Class of High Resolution Total-Variation-Stable Finite-Difference Schemes", *SIAM Journal of Numerical Analysis*, Vol. 21, No. 1, pp. 1-23.

Harten, A. and Osher, S. (1987)' "Uniformly High-Order Non-oscillatory Schemes I", *SIAM Journal of Numerical Analysis*, Vol. 24, pp. 279-309.

Holly, F.M., Jr. and Preissmann, A. (1977), "Accurate calculation of transport in two dimensions", *Journal of the Hydraulics Division, A.S.C.E.*, Vol. 103, pp. 1259-1277.

Khaliq, A.Q.M. and Twizell, E.H. (1982), "The extrapolation of stable finite difference schemes for first order hyperbolic equation", *International Journal of Computer Mathematics*, Vol. 11, pp. 155-167.

Kreyszig, E. (1983) Advanced Engineering Methods, J. Wiley and Sons.

Leonard, B.P. (1984), "Third order upwinding as a Rational Basis for Computational Fluid Dynamics", *Computational Techniques and Applications: CTAC-83*, editors B.J. Noye and C.A.J.Fletcher, North-Holland Publishing Co., pp. 106-120.

Leonard, B.P. and Niknafs, H.S. (1990), "The Ultimate CFD Scheme with Adaptive Discriminator for High Resolution of Narrow Extrema", *Computational Techniques and Applications: CTAC-90*, editors W.L. Hogarth and B.J. Noye, Hemisphere Publishing Corp., pp. 303-310.

Leslie, L.M., Mills, G.A., Logan, L.W., Gauntlett, D.J., Kelly, G.A., Manton, M.J., McGregor, J.L. and Sardie, J.M. (1985), " A High Resolution Primitive Equation NWP Mode for Operations and Research", *Australian Meteorological Magazine*, Vol. 33, pp. 11-35.

Martin, B. (1975), "Numerical Representations which Model Properties of the Solution to the Diffusion Equation", *Journal of Computational Physics*, Vol. 17, pp. 358-383.

McGregor, J.L. and Leslie, L.M. (1977), "On the Selection of Grids for Semi-Implicit Schemes", *Monthly Weather Review*, Vol. 105, pp. 236-238.

Morrow, R. (1981), "Numerical Solution of Hyperbolic Equations for Electron Drift in Strongly Non-Uniform Electric Fields", *Journal of Computational Physics*, Vol. 43, pp. 1-15.

Morrow, R. (1982), "Space-Charge Effects in High-Density Plasmas", *Journal of Computational Physics*, Vol. 46, pp. 454-461.

Morrow, R. (1985), " Theory of negative corona in oxygen", *Physical Review, Series A*, Vol. 32, pp. 1799-1809.

Morrow, R. (1991), "Theory of Positive Corona in SF_6 due to a Voltage Impulse",

IEEE Transactions on Plasma Science, Vol. 19, No. 2, pp. 86-94.

Morton, K.W. and Sweby, P.K. (1985), "A Comparison of Flux-Limited Difference Methods and Characteristic Galerkin Methods", Oxford University Computing Laboratory Report, No. 84/1.

Noye, B.J. (1984), "Analysis of Explicit Finite Difference Methods used in Computational Fluid Dynamics", *Contributions of Mathematical Analysis to the Numerical Solution of Partial Differential Equations*, edited by A.Miller, Proceedings of the Centre for Mathematical Analysis, Australian National University, pp. 106-118.

Noye, B.J. (1987), Lecture Series at University of Malaya (unpublished).

Noye, B.J. and Hayman, K.H. (1986), "Accurate Finite Difference Methods for Solving the Advection-Diffusion Equation", *Computational Techniques and Applications: CTAC-85*, editors B.J. Noye and R.L.May, North-Holland Publishing Co., pp. 137-158.

Noye, B.J. and Steinle, P.J. (1986), "Improved 5-point implicit methods for solving the 1-dimensional advection equation", *Computational Techniques and Applications : CTAC-85*, editors B.J. Noye and R.L. May, North-Holland Publishing Co., pp. 193-204

O'Brien, G.G., Hyman, M.A. and Kaplan, S. (1950), " A Study of the Numerical Solution of Partial Differential Equations", *Journal of Mathematics and Physics*, Vol. 29, pp. 223-251.

Owen, A. (1984), "Artificial Diffusion in the Numerical Modelling of the Advective Transport of Salinity", *Applied Mathematical Modelling*, Vol. 8, pp. 116-120.

Patnaik, G., Guirguis, R.H., Boris, J.P. and Oran, E.S. (1987), "A Barely Implicit

Correction for Flux-Corrected Transport", *Journal of Computational Physics*, Vol. 71, pp. 1-20.

Richardson, C.F. and Gaunt, J.A. (1927), "The Deferred Approach to the Limit", *Philosophical Transactions of the Royal Society*, London, Series A, Vol. 226, pp. 299-361.

Rusanov, V.V. (1970), "On Difference Schemes of Third Order Accuracy for Nonlinear Hyperbolic Systems", *Journal of Computational Physics*, Vol. 5, pp. 507-516.

Shapiro, R. (1970), "Smoothing, Filtering and Boundary Effects", *Reviews of Geophysics and Space Physics*, Vol. 8, No. 2, pp. 359-387.

Smolarkiewicz, P.K. (1983), "A Simple Positive Definite Advection Scheme with Small Implicit Diffusion", *Monthly Weather Review*, Vol. 111, pp. 479-486.

Sweby, P.K. (1984), "High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws", *SIAM Journal of Numerical Analysis*, Vol. 21, No. 5, pp. 995-1011.

van Leer, B. (1973), "Towards the Ultimate Conservative Difference Scheme : I. The Quest for Monotonicity", Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics, *Lecture Notes in Physics*, Vol. 18, pp. 163-168.

van Leer, B. (1974), "Towards the Ultimate Conservative Difference Scheme : II. Monotonicity Combined in a Second order Scheme", *Journal of Computational Physics* , Vol. 14, pp. 361-370.

van Leer, B. (1977), "Towards the Ultimate Conservative Difference Scheme :IV. A New Approach to Numerical Convection", *Journal of Computational Physics* , Vol. 23, pp. 276-299.

Steinle, P.J. and Morrow, R. (1989), "An Implicit Flux Corrected Transport Algorithm", *Journal of Computational Physics*, Vol. 80, pp. 61-71.

Takewaki, H. and Yabe, T. (1987), "The Cubic-Interpolated Pseudo-Particle (CIP) Method: Application to Nonlinear and Multi-Dimensional Hyperbolic Equations", *Journal of Computational Physics*, Vol. 70, pp. 355-372.

Thomas, L.H. (1949), "Elliptic Problems in Linear Difference Equations over a Network", Watson Scientific Computing Laboratory, Columbia University, New York.

Yee, H.C., Warming, Y.F. and Harten, A. (1985), "Implicit Total Variation Diminishing (TVD) Schemes for Steady-State Calculations", *Journal of Computational Physics*, Vol. 57, pp. 327-360.

Yee, H.C. (1986), "Linearized Form of Implicit TVD Schemes for the Multi-dimensional Euler and Navier-Stokes Equations", *An International Journal of Computers and Mathematics with Applications*, Vol. 12, pp. 413-432.

Yee, H.C. (1987), "Construction of Explicit and Implicit Symmetric TVD Schemes and Their Applications", *Journal of Computational Physics*, Vol. 68, pp. 151-179.

Warming, R.F. and Hyett, B.J. (1974), "The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods", *Journal of Computational Physics*, Vol. 14, pp. 159-179.

Zalesak, S.T. (1979), "Fully Multidimensional Flux-Corrected Transport Algorithms for Fluids", *Journal of Computational Physics*, Vol. 31, pp. 335-362.