Bull. Austral. Math. Soc. **26** (1982), 81-105, Retyped in 2011

The numerical solution of convective equations[1]

J. D. Fenton

Communicated by James M. Hill

Abstract

Finite-difference and spectral methods for the numerical solution of partial differential equations with convective terms are discussed. The finite accuracy and limited stability properties of such schemes are shown to follow from their non-recognition of the convective nature of the solutions which they seek, unlike schemes based on the use of characteristics. A numerical method for convective equations is proposed which incorporates the solution nature. The method is obvious and can be trivially derived, but seems not to have been exploited as it might have been. It uses interpolation only, rather than numerical differentiation, and for linear equations with constant coefficients it is exact and unconditionally stable. Although for more general equations the basic two-time level scheme is of relatively low accuracy, it can be simply used to generate a hierarchy of single-step multi-level methods of high accuracy.

1. Introduction

It is a widely-held view that numerical differentiation is an operation which should ideally be avoided, and yet the approximation of partial differential equations by finite-difference approximations to derivatives lies at the heart of computational fluid mechanics, particularly in the simulation of geophysical problems such as the motion of the sea and atmosphere. Such finite-difference expressions and methods are simply derived from the original equations, are computationally cheap, are capable of use in regions of arbitrary geometry, and they provide information in a convenient Eulerian sense at points fixed in space. While each finite difference approximation tries to mimic the differential equation, most make no attempt to incorporate the nature of the solution.

Most of the equations which are to be solved in fluid mechanics are of a rather similar convective nature, containing a linear time derivative term, plus convective terms in which velocities multiply spatial derivatives, then perhaps some terms due to pressure gradients, viscosity, rotational effects, and so on. The dominant feature of solutions to these equations is their essentially convective or travelling-wave nature.

A number of other methods, involving the use of characteristics, do attempt to build in the wave-like nature of the solutions. These, however, seem to find greater favour with theoreticians than with problem solvers. While characteristic-based methods have some very attractive features, such as usually-unrestricted stability, and the ability to describe the propagation of discontinuities, relatively little numerical analysis has been performed on them, often their accuracy is quite low and not made specific, and information may not be provided when and where it is wanted. Also, in many problems, characteristics do not exist.

This paper attempts to examine the relationships between the solution of a simple convective equation, and finite-difference and spectral approximations to that solution, as a model for rather more

[1] Received 28 January 1982. This paper is based on an invited lecture given at the Australian Mathematical Society Applied Mathematics Conference held in Bundanoon, February 7-11, 1982. Other papers delivered at this Conference appear in Volumes 25 and 26.

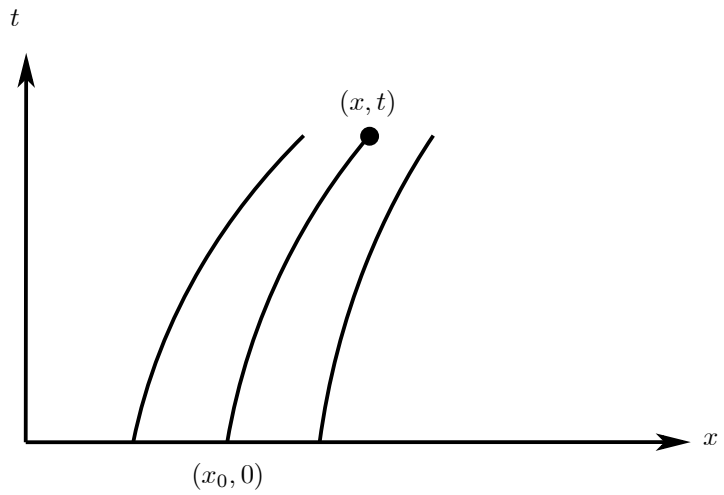

Figure 1. Three typical characteristics, including that passing through (x, t)

general systems whose solutions show the same behaviour. The equation is

$$\frac{\partial \theta}{\partial t} + u\frac{\partial \theta}{\partial x} = 0 \quad \text{for} \quad -\infty < x < \infty,\ t \geqslant 0 \tag{1}$$

which describes the variation of a scalar $\theta(x, t)$ in one space dimension x and time t, as it is carried by a velocity field $u(\theta, x, t)$, subject to the initial condition

$$\theta(x, 0) = f(x), \quad -\infty < x < \infty. \tag{2}$$

Subsequently a numerical method is developed which, although derived using characteristics, is given in a form so that θ may be solved at given fixed values of x and t , as with finite-difference methods. The method is unconditionally stable, and for certain problems is exact. Instead of numerical differentiation, the spatial operations are those of interpolation, which is less susceptible to error. It is suggested that approximation by cubic splines is the most robust, accurate and convenient means of interpolation. As far as representation in time is concerned, the method is nominally of first-order accuracy only. However, it is shown how solution methods of high order are easily generated.

2. An exact solution

The differential equation (1) shows that on a characteristic curve given by $dx/dt = u(\theta, x, t)$ the convective derivative of θ is zero, thus θ is a constant. To solve for $\theta(x, t)$ it is necessary only to find the value of x , x_0 say, through which the characteristic passes at $t = 0$. The situation is shown in Figure 1: for a given (x, t) one has to find which of the characteristics emanating from the x axis passes through (x, t) .

The solution is

$$\theta(x, t) = \theta(x_0, 0) = f(x_0), \tag{3}$$

from (2). The differential equation governing the characteristic is $dx/dt = u(\theta(x, t), x, t)$, in which θ is a constant, and the initial condition is $x(0) = x_0$, which is as yet unknown. If the solution is

$$x = x_0 + g(\theta(x, t), t), \tag{4}$$

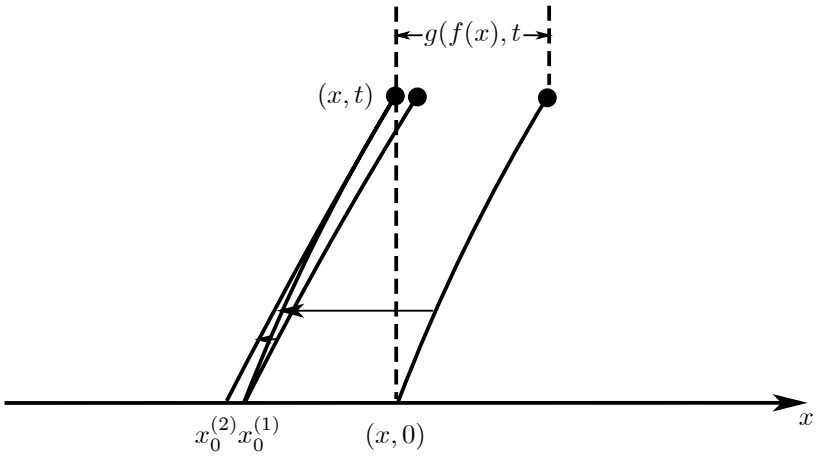

Figure 2. Iterative calculation of x_0

then eliminating x_0 between (3) and (4) gives the exact but implicit solution

$$\theta(x,t) = f(x - g(\theta(x,t),t)). \tag{5}$$

This is in a form suitable for fixed-point iteration, such that for given x and t , and some estimate θ_n for $\theta(x,t)$, another estimate is $\theta_{n+1} = f(x - g(\theta_n, t))$, the procedure being repeated successively. As shown in [2, §3.3], this will converge to a solution provided $|df/d\theta| < 1$. Perhaps the least arbitrary initial estimate θ_1 is $\theta(x,0)$, which is given by $f(x)$, in which case (5) can be written as the explicit iterated function

$$\theta(x,t) = \ldots f(x - (f(x - g(f(x),t)),t)) \ldots . \tag{6}$$

The physical significance of this is simple. From (4), $g(\theta(x,t),t) = x - x_0$ is the horizontal displacement of the characteristic as it travels from $(x_0, 0)$ to (x,t), so that $g(f(x),t)$ is the horizontal displacement of the characteristic through $(x,0)$, and $x_0^{(1)} = x - g(f(x),t)$ is the value of x at which that characteristic would intersect the t axis if it were displaced so as to pass through (x,t). The next iteration then uses this value of $x_0^{(1)}$ to calculate the next characteristic, subsequently displaced to give $x_0^{(2)}$ and so on. This process is shown geometrically in Figure 2.

As an example, consider the quasilinear problem

$$\frac{\partial\theta}{\partial t} + \theta\frac{\partial\theta}{\partial x} = 0,$$

with the initial condition $\theta(x,0) = f(x) = -\tanh x$. Solutions of such a problem show gradual steepening, as large values of θ travel with a corresponding large velocity θ. Using the method described above, it.is easily shown that $g(\theta,t) = t\theta$, so that the implicit solution is $\theta(x,t) = -\tanh(x - t\theta(x,t))$, and the explicit iterated solution is

$$\theta(x,t) = \ldots - \tanh\left(x - t\left(-\tanh\left(x - t\left(-\tanh x\right)\right)\right)\right) \ldots .$$

The condition that this iteration converges becomes $t\,\mathrm{sech}(x - t\theta) < 1$. Thus, the method will converge only for finite t. The significance of this can be shown by considering the origin $x = 0$,at which $\theta(0,t) = 0$, so that $t < 1$ for convergence. Locally, $\theta(x,0)$ is a straight line for x small, whose horizontal velocity increases linearly with θ, thus the profile of θ rotates as a straight line in the vicinity of the origin, until at $t = 1$ when the iteration method fails, the profile has a vertical tangent at the origin. It can be shown that this corresponds to two characteristics crossing at $(0,1)$.

3. Approximate solutions

For rather more general problems than the system (1) and (2), such as those involving more than one differential equation in more than one space variable, the luxury of an exact solution is not available. Here, two different types of approximations to (1) and (2) and to the solution (5) will be developed, which are applicable to more complicated problems.

A convective approximation

The most obvious approximation to the solution obtained through use of characteristics is to assume that all characteristics, locally at least, are straight and parallel so that $g \approx u(x,0)t$, and solution (5) becomes

$$\theta(x,t) \approx f(x - u(x,0)t), \tag{7}$$

If the velocity u is constant, all characteristics are parallel straight lines, and (7) is an exact solution. If the velocity is a function of x or of t , then it is an approximation. It does, however, incorporate the convective nature of the original differential equation and allows for travelling wavelike solutions. Henceforth in this paper it will be referred to as the convective approximation, on which convective methods are developed. It assumes that the value of θ at (x,t) is that which was upstream at time 0, at just the right distance to have been carried downstream at a mean velocity equal to that at $(x,0)$.

Taylor series approximations

Consider the exact infinite Taylor expansion for $\theta(x,t)$ in terms of the initial $\theta(x,0)$ and its derivatives:

$$\theta(x,t) = \theta(x,0) + t\,\theta_t(x,0) + \tfrac{1}{2}t^2\theta_{tt}(x,0) + \ldots.$$

If the differential equation (1) and the initial condition (2) are substituted into this, then

$$\theta(x,t) = f - tuf_x + \tfrac{1}{2}t^2\left(u^2 f_{xx} + u\,u_x\,f_x - u_t f_x\right) + O\left(t^3\right), \tag{8}$$

where $f = f(x)$ and $u = u(x,0)$. Finite difference and spectral methods use this expression or part of it, and approximate the spatial derivatives numerically.

Comparison of accuracy

The level of accuracy of the convective expression (7) can be found by writing it as the infinite Taylor series

$$\theta(x,t) = f - tuf_x + \tfrac{1}{2}t^2u^2 f_{xx} - \ldots. \tag{9}$$

Comparing (8) and (9) it is clear that (7) has the relatively low order error term $O(t^2)$; it makes no explicit allowance for a changing velocity, as shown in the second order terms in (8), which equation has made no attempt to include the convective nature of solutions of the differential equation. It will be shown that this latter omission has some important consequences.

As the solutions (7) and (8) are valid only for small t, in any numerical solution a number of small time steps will have to be taken. The two schemes are here re-written so that if at a given time t, $\theta(x,t)$ and $u(x,t)$ are known, denoted by θ and u respectively, then $\theta(x,t+\Delta)$ after an increment of time Δ is given by

(i) the convective scheme as

$$\theta(x,t+\Delta) = \theta(x - u\Delta, t) + O\left(\Delta^2\right), \tag{10}$$

while (ii) the Taylor series becomes

$$\theta(x,t+\Delta) = \theta - u\Delta\theta_x + \tfrac{1}{2}\Delta^2\left(u^2\theta_{xx} + u\,u_x\,\theta_x - u_t\theta_x\right) + O\left(\Delta^3\right). \tag{11}$$

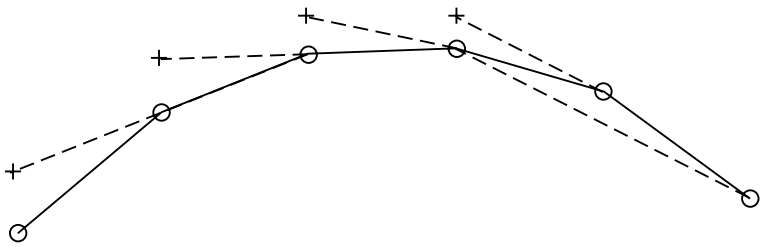

Figure 3. Instability of forward-time centred-space numerical scheme. Circles show initial profile. The crosses show the profile at the next time step, for $u\Delta = \delta$.

4. Finite difference methods

These are based on approximations to the Taylor series (11). Here, only explicit schemes will be described.

First-order schemes

Consider (11) to be truncated after the second term

$$\theta(x, t + \Delta) = \theta - u\Delta\theta_x + O\left(\Delta^2\right), \tag{12}$$

the error term is of the same order as in (10). If it is assumed that the value of $\theta_x(x, t)$ is given by the centred-difference expression $\theta_x(x, t) = (\theta(x+\delta, t)-\theta(x-\delta, t))/2\delta+O\left(\delta^2\right)$, in terms of the point values of θ at grid points $x \pm \delta$, then the scheme (12) can be represented as shown in Figure 3, for some points near a local extremum. According to (12) the value of $\theta(x, t+\Delta)$ is equal to that at $\theta(x, t)$ plus the change obtained by travelling along the tangent at x a horizontal displacement of $-\Delta u$. For a time step Δ such that $\Delta u = \delta$ the values thus predicted are shown by the crosses in Figure 3, the wave defined by the crosses being shifted to the right. It is obvious that the scheme exaggerates extrema and would be unstable. The failure of this simple and obvious scheme is well known and can be shown by the less-graphic but more rigorous von Neumann method of stability analysis to be unconditionally unstable (see Noye [4,§5.1]).

This instability has often been attributed to the use of downstream information through the use of $\theta(x + \delta, t)$,however the above geometric interpretation suggests that it is not the use of the downstream values *per se* that causes the instability. Rather, it is the poor attempt of the scheme (12) at extrapolating the upstream shape of the wave, to predict what it is at the next time step.

Now consider the scheme (12) but where θ_x is approximated by the backward-difference expression

$$\theta_x(x, t) = (\theta(x, t) - \theta(x - \delta, t))/\delta + O(\delta)$$

which actually is less accurate than the centred-difference expression. Predicted points now lie on the line joining $\theta(x, t)$ and $\theta(x - \delta, t)$. Results for $u\Delta = \frac{1}{2}\delta$ are shown in Figure 4(a). The method seems to be stable, however numerical diffusion or damping becomes apparent because of the relative poorness of straight line interpolation, and is emphasised by the results given for the succeeding time step. For a value of $u\Delta = \delta$ the scheme is exact, and corresponds the convective expression, because (12) now gives $\theta(x, t + \Delta) = \theta(x - \delta, t)$. However, in the general context of possibly quasi-linear equations this value of $u\Delta$ cannot be ensured, and the exactitude is not an important result. For $u\Delta = \frac{3}{2}\delta$ as shown in Figure 4(b) it is clear that the method is unstable, as it would be for any $u\Delta > \delta$, when the straight line interpolation becomes extrapolation. This stability criterion can be established by the von Neumann method [4, §5.17].

To compare various schemes a model problem was posed, that of a sharp-crested profile, defined by only five points, being convected by a constant velocity. Because of the gradient discontinuities this is a rather severe test of approximation methods, however it is a useful example as it shows in an

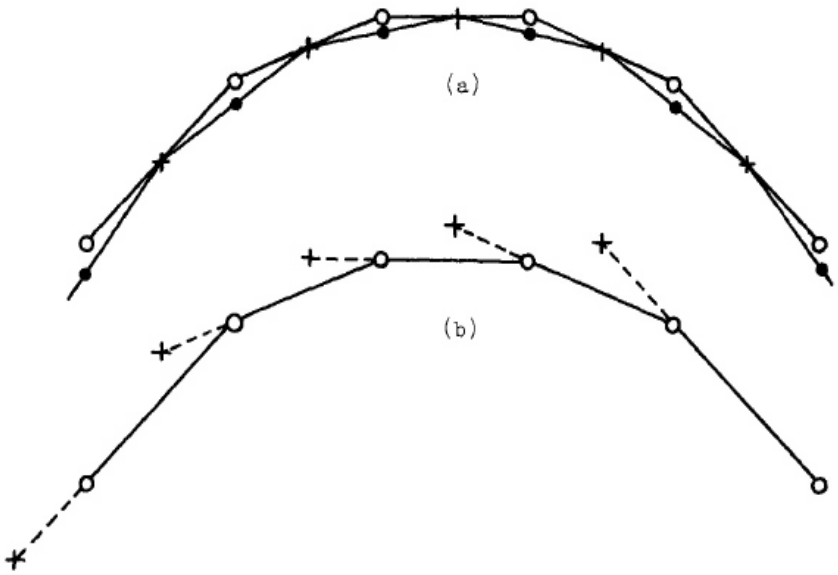

Figure 4. Conditional stability of upwind difference schemes. ○ Initial profile, + first time step, ● second time step; (a) is for $u\Delta = \frac{1}{2}\delta$, showing numerical damping of the solution; (b) is for $u\Delta = \frac{3}{2}\delta$, showing instability.

exaggerated the phenomena to be demonstrated. Results are shown in Figure 5, after a total time of $1.5\,\delta/u$ when the wave should have travelled a distance of 1.5 grid intervals. In Figure 5(a) results are shown for a time step such that $u\Delta = 0.5$, necessitating 3 steps to reach the stage required. The numerical damping is obvious. A value of $u\Delta = 1.5$ was used to obtain the results in Figure 5(b). After just one step, the dramatic instability can be seen.

Second-order (Lax-Wendroff) schemes

The defects of the first-order schemes have been seen to flow from the inadequacy of their spatial approximation, thus the use of higher-order schemes is suggested. If all the quadratic terms in (11) are retained the methods are generally known as Lax-Wendroff, whereas if the terms containing variation in u are not included this is known as Leith's method. Both are described in Roache [15, pp. 75-83, pp. 244-250]. Typically these methods interpolate over two grid intervals by a parabola, defined by the three points $\theta(x-\delta,t)$, $\theta(x,t)$, and $\theta(x+\delta,t)$ to give values of $\theta_x(x,t)$ and $\theta_{xx}(x,t)$.

The Leith/ Lax-Wendroff method for $u\Delta/\delta = 1/2$ was used on the test problem described above, with results shown in Figure 5(c). The approximation is better, the numerical damping is less, but the well-known phase error of the method is apparent, as the numerical solution lags behind the exact solution. The rather irregular curves joining the computational points are the interpolating parabolae, drawn upstream over the interval in which they are required to interpolate. The computational points at the next time step would be those at the mid-point of each curve for $u\Delta/\delta = 1/2$. It can be how the use of piecewise- quadratic approximation can lead to irregularities in the profile. The situation is worse in Figure 5(d), for $u\Delta/\delta = 3/2$, where the plotted points arose from a single time step and the initial parabolae used to extrapolate upstream, outside the interval containing the three defining points. The results are even more wildly divergent than those shown in Figure 5(b) for linear extrapolation, not so surprising when one considers the behaviour of parabolae compared with that of straight lines. The method seems to be unstable for $u\Delta > \delta$, for the same geometric reasons as for linear extrapolation, and this stability criterion is the case [5, p. 78].

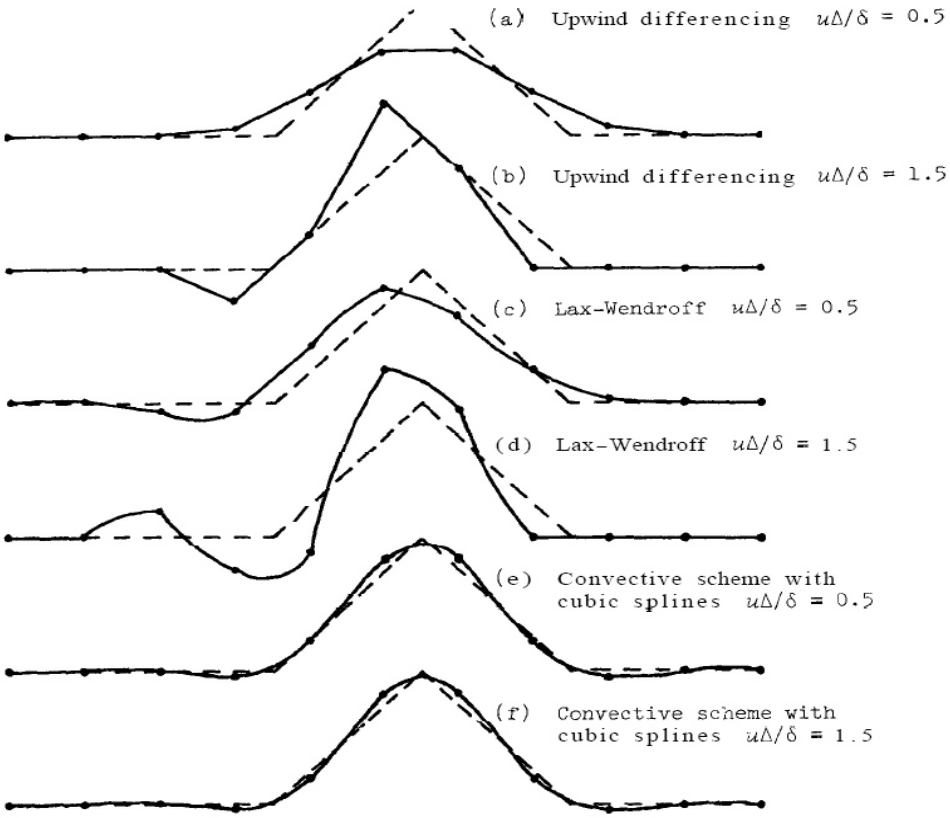

Figure 5. Comparison of different schemes for the problem of a sharp-crested profile being convected by a constant velocity, after a time such that the profile has moved 1.5 grid intervals. The dashed line shows the exact solution., the small closed circles show the numerical results at the grid points; they are connected by the interpolating function used in each case.

Higher order schemes

Kreiss and Oliger [3] studied several higher order schemes and concluded that higher-order schemes are indeed more accurate. They considered a sharp profile as in Figure 5, but they took grid intervals 1/10 of those in Figure 5, and time steps such that $u\Delta/\delta = 0.1$; that is, they interpolated upstream only 1/10 of a grid interval, in which case high-order interpolation should be safe and accurate. For demanding cases, results from the present work suggest that the use of higher-order methods with their attendant higher-degree polynomial interpolation may be hardly worthwhile. The above geometric arguments show that for any local polynomial approximation, the stability limit of $u\Delta/\delta = 1$ remains the same, and there seems to be the danger that higher order schemes are less robust.

Other timestepping schemes

There are many other schemes in addition to the two time level (that is, t, $t + \Delta$) one step explicit schemes discussed above. For example, by writing the Taylor expansion (11) for $t - \Delta$ and subtracting that from (11) the "leap-frog" expression is obtained, in which the error terms are third order:

$$\theta(x, t + \Delta) = \theta(x, t - \Delta) - 2\Delta u\, \theta_x(x, t) + 0(\Delta^3). \tag{13}$$

If the $\theta_x(x, t)$ is approximated by a centred-difference expression, (13) is stable provided $u\Delta/\delta \leqslant 1$. Geometrical demonstration of stability for such a scheme is rather more complicated than for the two time level schemes, and will not be presented here. schemes, which involve the solution of a number of equations at each step, are generally stable. However stable they are, they still make use of lower order spatial approximation and numerical differentiation, and have much in common with the methods already described.

5. Spectral methods

All finite difference methods are based on local approximation. In this section the application of global approximation by spectral methods will be discussed, Fourier series the simplest example. The methods described are pseudospectral, in that most operations are performed in physical rather than spectral space. will be shown that conventional spectral methods have few advantages over finite difference methods.

Consider a finite computational region $-L/2 \leqslant x \leqslant L/2$ divided into N equal intervals, and the solution $\theta(x,t)$ represented by the point values $\theta_m(t) = \theta(mL/N, t)$, for $m = -N/2, ..., N/2$. Because of the implied periodicity in the use of Fourier series, $\theta_{-N/2} = \theta_{N/2}$. The N values can be transformed by the (inverse) discrete Fourier transform

$$\Theta_j(t) = \mathcal{D}^{-1}(\theta_m(t); j) = \frac{1}{N} \sum_{m=-N/2}^{N/2}{}'' \; \theta_m(t) \; \exp(-\mathrm{i}2\pi mj/N) \tag{14}$$

for $j = -N/2, \ldots, +N/2$. The summation over m has a factor of $1/2$ multiplying the contribution at $\pm N/2$. This 'trapezoidal' summation is denoted by Σ''. The inverse transformation can be obtained in relatively few operations using standard fast transform techniques and exploiting the fact that the θ_m are real only, so that the number of operations is of order $N/2 \log_2 (N/2)$ if N is equal to some power of two.

The coefficients $\Theta_j(t)$can be transformed using the discrete Fourier transform to recover the point values (see [1, §6.37], which has notational differences):

$$\theta_m(t) = \mathcal{D}(\Theta_j(t); m) = \sum_{j=-N/2}^{N/2}{}'' \; \Theta_j(t) \; \exp(+\mathrm{i}2\pi mj/N) \tag{15}$$

for $m = -N/2$ to $+N/2$, the summation in j also being over these values. The interpolating Fourier series which takes the values $\theta_m(t)$ at the points $x = mL/N$ is simply

$$\theta(x,t) = \sum_{j=-N/2}^{N/2}{}'' \; \Theta_j(t) \; \exp(+\mathrm{i}jkx), \tag{16}$$

where $k = 2\pi/L$.

Simple considerations based on the result that coefficients Θ_j of infinite Fourier series vary like $j^{-(n+1)}$,where the function $\theta(x,t)$ has a discontinuity in the nth derivative, suggest that the truncated Fourier series (16) has errors of the order of $((N/2)+1)^{-(n+1)}$ or roughly δ^n, where $\delta = L/N$. For sufficiently-continuous functions the approximation is very accurate. However, for functions which are discontinuous themselves or in their lower-order derivatives, the spatial approximation accuracy is comparable with that of polynomial approximation only.

Now consider the first order solution scheme (12):

$$\theta(x, t+\Delta) = \theta - u\Delta\theta_x + O\left(\Delta^2\right).$$

If $\theta(x,t)$ is approximated by the Fourier series, then substituting (16) gives

$$\theta(x, t+\Delta) = \sum_{j}{}'' \; \Theta_j(t) \; (1 - \mathrm{i}jk\Delta u\,(x,t)) \exp(+\mathrm{i}jkx),$$

and at the node point $x_m = mL/N$, with $u_m = u(x_m, t)$, the value of θ_m at the next time step is

predicted to be

$$\theta_m(t+\Delta) = \sum_j{}'' \Theta_j(t)\ (1 - \mathrm{i}jk\Delta u_m) \exp(+\mathrm{i}2\pi mj/N) + O\left(\Delta^2\right). \qquad (17)$$

The results of applying this scheme are shown for the sharp-crested wave in Figure 6. The scheme (17) predicts that the solution value after a time step Δ will be that obtained by extrapolating back along the tangent at the grid point a horizontal displacement $-\Delta u$. However accurate the value of θ_x obtained from the interpolating function, the results are catastrophic, for the method is unstable in precisely the same way as finite difference approximations to the derivative.

That the method is unstable is easily shown by the von Neumann method using the present spectral approach. If u_m in (17) i s a constant, u, then the scheme can be written in spectral form $\Theta_j(t+\Delta) = \Theta_j(t)(1-\mathrm{i}jk\Delta u)$ for each j, and it is obvious that the magnitude of the factor on the right is greater than unity: the Fourier coefficients grow exponentially with time and the method is unstable. All the extra trouble of developing a global method has gained nothing.

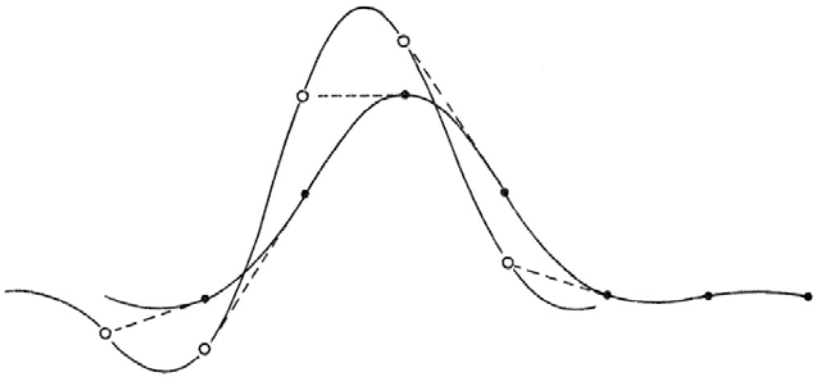

Figure 6. Instability of first-order scheme with spectral method. Despite the initial accurate interpolation by the Fourier series through the solid circles, the use of a first-order scheme extrapolates backwards along the local tangents shown dashed, giving the solution shown by the open circles, which will subsequently diverge even more wildly.

If the second-order scheme (11) is used, then for constant u the predicted value of $\theta(x, t+\Delta)$ is simply that obtained by backwards interpolation on a parabola which has the required first and second derivatives at $\theta(x,t)$, given by the Fourier approximation. For the case of the two points at the centre of the sides in Figure 6, the local. curvature of the approximation is zero, and the predicted values of$\theta(x, t+\Delta)$ would be precisely those as shown in the figure, suggesting that the method is unstable. This instability is easily shown, by considering the Fourier coefficients of the scheme (11), which gives

$$\frac{\Theta_j(t+\Delta)}{\Theta_j(t)} = 1 - \mathrm{i}jk\Delta u + \tfrac{1}{2}\left(\mathrm{i}jk\Delta u\right)^2,$$

the right side has a magnitude greater than unity, and the method is unconditionally unstable, unlike the Lax-Wendroff method of finite-difference approximation to (11).

If, instead of the two time level approach, a leapfrog method is used, then the scheme (13) becomes, in spectral space:

$$\Theta_j(t+\Delta) = \Theta_j(t-\Delta) - \mathrm{i}2jk\Delta u\Theta_j(t) + O\left(\Delta^3\right).$$

By supposing that the ratio of Θ_j between any two successive time levels is a constant, r, the quadratic equation is obtained:

$$r^2 + 2r(\mathrm{i}jku\Delta) - 1 = 0,$$

with solutions

$$r = -\mathrm{i}jku\Delta \pm \left(1 - (,jku\Delta)^2\right)^{1/2}.$$

Provided $|jku\Delta| \leqslant 1$, $|r| = 1$ and the scheme is stable. As the maximum value of j is $N/2$, and $k = 2\pi/L$, this criterion becomes $u\Delta/\delta \leqslant 1/\pi$ for stability, which is more demanding than that found for finite difference leapfrog methods, however the scheme is at least *conditionally* stable. In view of the better stability properties of the leapfrog scheme (13) for both finite-difference and spectral methods, it does seem that it is a rather more natural way of dealing with hyperbolic equations and is clearly much to be preferred over two time level (Euler) schemes such as (11).

Finally a comment can be made on the computational cost of using Fourier methods. While fast Fourier transforms can be used the coefficients Θ_j at each step, if the velocity u is a function of x, the u_m are not constant, and the N-term series must be evaluated directly at each of the N points, giving a effort of $O\left(N^2\right)$ compared with the $O\left(N\right)$ of finite difference methods.

6. Convective methods

Consider the scheme (10):

$$\theta(x, t+\Delta) = \theta(x - u\left(x,t\right)\Delta, t) + O\left(\Delta^2\right).$$

This is exact if u is constant, but otherwise has the low-order error term shown. However, it involves no numerical differentiation; rather, it is only necessary to *interpolate* to evaluate the right side. That this is consistent with the differential equation (1) in the limit $\Delta \to 0$ can easily be shown by writing each side as a Taylor expansion about (x,t) and then taking the limit. Stability of the scheme can be studied for the case of constant u, and examining one component of (10) in spectral space. The Fourier coefficient at the next step $\Theta_j(t+\Delta)$ is given by $\Theta_j(t+\Delta) = \Theta_j(t)\exp\left(-\mathrm{i}jk\Delta u\right)$ for any j, This further demonstrates the nature of the scheme – the coefficients are unchanged in magnitude but are changed in phase by an amount $jk\Delta u$, precisely the amount by which the component $\exp(-\mathrm{i}jkx)$ should change in time Δ: the scheme is unconditionally stable. It is interesting that the first and second order spectral schemes in §5 are simply low order approximations to this:

$$\exp(-ijkx) = 1 - \mathrm{i}jk\Delta u + \tfrac{1}{2}\left(\mathrm{i}jk\Delta u\right)^2 - \ldots,$$

but whose magnitudes at each level of truncation are greater than unity, unlike the left side, and the methods are unstable.

As the convective scheme is *stable*, and it is *consistent* with the differential equation, then *convergence* of numerical solutions to the exact solution in the limit as $\Delta \to 0$ is indicated [4, §3.4]. Instead of having to consider a number of different ways of approximating derivatives, attention can now be fixed on means of interpolating values of θ, given values at a finite number of grid points.

Piecewise-polynomial approximation

In §4 it was shown that approximation of the Taylor expansion (11) by low-order polynomials had disadvantages. If the polynomials are used purely for interpolation, some of these problems do not occur. For instance, instead of extrapolating back a distance of Δ along a local tangent, it is now a matter of locating the interval in which $x - u(x,t)\Delta$ falls, between x_m and x_{m+1} say, and then obtaining the local approximation. Linear interpolation would give, for example,

$$\theta\left(x - u\Delta, t\right) \approx \theta\left(x_m, t\right) + \frac{\left(x - u\Delta - x_m\right)}{x_{m+1} - x_m}\left(\theta\left(x_{m+1}, t\right) - \theta\left(x_m, t\right)\right)$$

It is not necessary to have a constant grid spacing. While the method is unconditionally stable for constant u , the linear approximation may not be very accurate. For the problem of a sharp-crested

wave shown in Figure 5, application of this method for a time step of $\Delta = (n + 1/2)\delta/u$, for any integer n, would give the results in Figure 5(a) after three such steps - the method has the same amount of diffusion per time step as the forwards-time backwards-space finite difference method.

A more accurate method of piecewise polynomial approximation is by Cubic Splines, where third degree polynomials are used which have continuous derivatives at node points. A complete set of FORTRAN subprograms is given in [2, §6.7] in which some of the theory of cubic spline interpolation given. It can be shown that the error of a cubic spline interpolant over an interval $[a, b]$ is bounded by $5\left|\theta^{(4)}(\xi)\right|\delta^4/384$, where ξ is in $[a, b]$ and δ is the maximum step length. Importantly, the computational effort is proportional to N, the number of computational points, even if u is a function of x. This compares favourably with spectral methods, $O(N^2)$ for a non-constant convective velocity.

The model problem of the sharp-crested wave was solved using the convective scheme (10) with cubic splines as interpolating functions. Results are shown in Figure 5(e) for $u\Delta/\delta = 1/2$, the case in Figures 5(a) and (c) for finite difference methods. Some numerical diffusion has occurred, because of the finite accuracy of the spline interpolation, however, there is no phase error, and the solution is much more accurate than the other methods. In Figure 5(f) the results are shown for $u\Delta/\delta = 1.5$, when the finite difference methods were unstable. The results are even more accurate than in Figure 5(e) because fewer time steps have been taken, with smaller total diffusion. This problem, however, is a very easy one for the convective method to solve, as the scheme is exact, the only approximation being in the spatial representation. A more demanding quasilinear problem is that solved analytically in §2:

$$\frac{\partial\theta}{\partial t}e + \theta\frac{\partial\theta}{\partial x} = 0, \qquad \theta(x, 0) = -\tanh x,$$

in which the profile of θ becomes vertical and then multi-valued after $t = 1$.

A computational region $(-4, 4)$ was divided into 20 equal intervals, and various methods implemented. After 10 steps of $\Delta = 0.1$, by which time the analytical solution develops a vertical tangent at $x = 0$, the results are as shown in Figures 7 and 8.

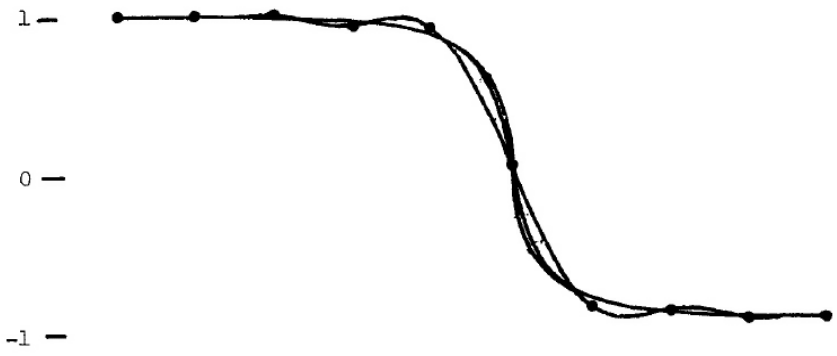

Figure 7. Solution to quasilinear equation at the instant the profile develops a vertical tangent. The steepest curve is the exact solution, that close to it is the solution from the convective method with cubic splines and clustered grid points, while the line passing through the solid circles was obtained from equispaced points shown.

Figure 7 shows that the convective method (10) with cubic splines developed oscillations reminiscent of Gibbs' phenomenon in Fourier approximation, because the gradient, as shown by the analytical solution, became very steep, and the point spacing was too coarse to describe the region of high curvature. The numerical solution is, however, oscillating about the analytical solution as if it were attempting to describe it in a minimum least-squares error sense. With a trivial modification to the computer program, a variable grid spacing was used, points being distributed according to

a cubic power law. Results are very close to the analytical solution on Figure 7. It seems that the freedom to use variable mesh spacing might be a useful feature of this method.

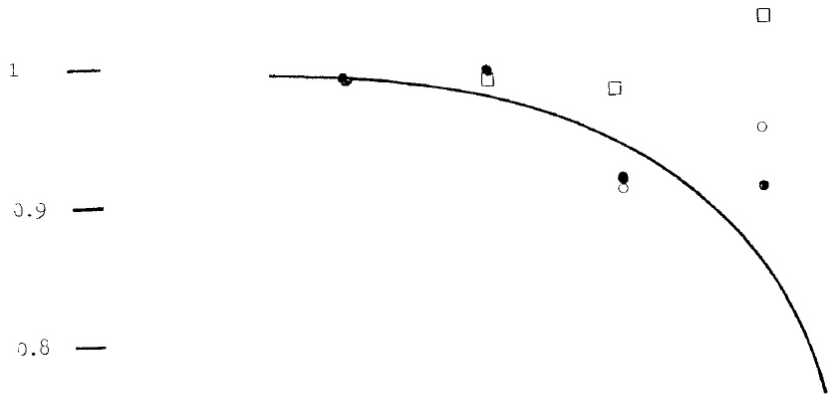

Figure 8. Comparison of different methods after 10 steps of interval 0.1 for equispaced grid points: Line – exact solution; □ – conventional Lax-Wendroff; ○ – Lax-Wendroff with cubic splines; ● – convective method with cubic splines.

In Figure 8 the results of three schemes for equispaced points are compared over part of the solution region. It can be seen that the convective method is the most accurate, even though it has lower order error terms! It seems that the absence of numerical differentiation contributes more to accuracy than does the inclusion of higher-order terms, where derivatives become large. In view of the success of the cubic spline/convection method, the author could not resist the temptation to use the Taylor expansion correct to second order, but where the derivatives were obtained from a cubic spline fit to the grid points. This was still not as accurate as the pure convection method however, providing further support for the view that numerical differentiation should be avoided. The least accurate scheme of the three tested was the nominally more-accurate conventional Lax-Wendroff method, however, it did perform satisfactorily until the solution became very steep.

Higher-order convective schemes

Consider the two time level convective scheme with the error terms shown as an infinite power series, and where the coefficients a_n are unknown:

$$\theta(x, t+\Delta) = \theta(x - u\,(x,t)\,\Delta, t) + \sum_{n=2}^{\infty} a_n \Delta^n. \tag{18}$$

A hierarchy of higher-order schemes can be generated, theoretically without limit, by writing (18) for different integer multiples of Δ and eliminating the a_n to a certain order. Thus, for example, the leapfrog convective scheme is obtained:

$$\theta(x, t+\Delta) = \theta(x, t-\Delta) + \theta(x - u\Delta, t) - \theta(x + u\Delta, t) + O\left(\Delta^3\right), \tag{19}$$

where $u = u(x,t)$. The scheme is accurate to second order. If the terminology

$$\theta_n\,(j) = \theta\,(x + j\Delta u\,(x,t)\,, t + n\Delta)$$

is adopted, the scheme (10) becomes

$$\theta_1(0) = \theta_0(-1) + O\left(\Delta^2\right)$$

and (19) becomes

$$\theta_1(0) = \theta_{-1}(0) + \theta_0(-1) - \theta_0(1) + O\left(\Delta^3\right)$$

It is easily shown that time level expression, with errors of fifth order is

$$\theta_2(0) = \theta_{-2}(0) + \theta_0(-2) - \theta_0(2) + 8\left[\theta_1(0) - \theta_{-1}(0) + \theta_0(1) - \theta_0(-1)\right] + O(\Delta^5). \quad (20)$$

Such schemes cost little more in computing resources, as the interpolation need only be done at one time level (t) per step in each method. Most effort is incurred in setting up the coefficients for the spline approximation, which must be done whatever the order of the method The number of subsequent interpolations (that is, polynomial evaluations) for each grid point in each scheme is proportional to the order of accuracy. For example, the fourth-order scheme involves four evaluations at each point, $\theta_0(-2),\ 0_0(-1),\ 0_0(1)$ and $0_0(2)$. At other time levels, only the data at x is required, the number of time levels to be stored being equal to the order of accuracy of the scheme. For example, the scheme (20) involves the four node values $\theta_{-2}(0)$, $\theta_{-1}(0)$, $\theta_1(0)$, and although not shown explicitly $\theta_0(0)$ must be stored to enable the splines to be fitted.

Multi-time level schemes do require some starting, however, the five time level scheme (20) requiring initial values at four time levels. This can conveniently be done by using lower order schemes with smaller time steps.

Fourier approximation

If $\theta(x,t)$ is represented by a Fourier series (16), where the coefficients $\Theta_j(t)$ are obtained from the point values $\theta_m(t)$ by (14), then the convective scheme gives, for the point values $\theta_m(t+\Delta)$:

$$\theta_m(t+\Delta) = \sum_j{}'' \Theta_j(t)\ \exp\left(-\mathrm{i}jk\Delta u_m\right) \exp(\mathrm{i}2\pi mj/N) + O\left(\Delta^2\right). \quad (21)$$

It is clear that the effect of the convective velocity is simply to change the phase of the Fourier coefficients.

It was shown in §5 that for functions which are discontinuous or which have discontinuous low-order

Figure 9. Interpolation near a discontinuity by: - - - finite Fourier series, – cubic splines.

derivatives, that Fourier approximation may be little better than low-order polynomial methods. To examine this for the case of a discontinuous function, a simple step discontinuity was approximated by a 20-term Fourier series and cubic splines. The results are shown in Figure 9. At the jump, the two methods agree closely, however, it is clear that the oscillations of the Gibbs phenomenon in the Fourier series are larger than in the spline approximation and they persist for much further away from the discontinuity. While such discontinuities may not exist in the interior, the possibly-artificial periodicity imposed by the Fourier approximation may impose discontinuities the ends, such as would be the case in the example shown in Figure 7, where $\theta(-4^+,t) = 1$ and $\theta(4^-,t) = -1$, however, a Fourier scheme of period 8 would introduce $\theta(-4^-,t) = -1$ and $\theta(4^+,t) = 1$, giving jumps at each end. To eliminate such discontinuities it would be necessary to use some form of polynomial subtraction or artificial extension of the computational interval and the possible matching of a mirror image of $\theta(x,t)$. This would not be necessary if, for example, a Chebyshev spectral scheme were used.

As mentioned in §5, for problems of non-constant a another serious disadvantage exists, that the

series (21) have to be evaluated with a computational effort proportional to N^2 , rather than the N of piecewise polynomials. To conclude, it does seem that the Fourier approximation has little to recommend it for general quasilinear hyperbolic problems.

7. Some other applications of convective schemes

In this section some applications to rather more general problems are briefly discussed.

Three space dimensions

Consider the quasilinear convective equation in three dimensions:

$$\frac{\partial \theta}{\partial t} + \mathbf{u}\cdot\nabla\theta = 0,$$

where the velocity field $\mathbf{u}$ is a vector of position and time. The following scheme is immediately obvious, following from (10):

$$\theta(\mathbf{r}, t+\Delta) = \theta(\mathbf{r} - \mathbf{u}\Delta, t) + O\left(\Delta^2\right),$$

where $\mathbf{r}$ is the position vector. All the higher-order schemes for the timestepping are immediately applicable. By expressing θ as a triple Fourier series in x, y and z and taking the case $\mathbf{u}$ = constant it is trivially shown that the scheme is stable, and by writing a Taylor series expansion about $(\mathbf{r}, t)$ for the left and the right sides it can be shown that the scheme converges to the differential equation in the limit $\Delta \to 0$. The scheme is independent of the co-ordinates used, and in its vector form could be written in terms of any orthogonal co-ordinate system.

In three dimensions the problem of interpolation becomes considerably more complex. Greater reliance may have to be placed on piecewise polynomial methods. Such methods may be quite problem-specific.

Long waves in canals

It can be shown that the equations governing the motion of long waves in rectangular canals are capable of being expressed in characteristic form:

$$\begin{aligned}\left(\frac{\partial}{\partial t} + (u+c)\frac{\partial}{\partial x}\right)(u+2c) &= 0, \quad \text{and} \\ \left(\frac{\partial}{\partial t} + (u-c)\frac{\partial}{\partial x}\right)(u-2c) &= 0,\end{aligned}$$

where $u(x,t)$ is the horizontal fluid velocity, and

$$c(x,t) = (\text{gravitational acceleration} \times \text{local depth})^{1/2}$$

is a measure of the local depth. The convective scheme follows immediately:

$$\begin{aligned}u(x,t+\Delta) + 2c(x,t+\Delta) &= u(x-(u+c)\Delta, t) + 2c(x-(u+c)\Delta, t) + O\left(\Delta^2\right), \quad \text{and} \\ u(x,t+\Delta) - 2c(x,t+\Delta) &= u(x-(u-c)\Delta, t) - 2c(x-(u-c)\Delta, t) + O\left(\Delta^2\right).\end{aligned}$$

The right sides of these equations can be evaluated by some interpolation method, and the pair of equations solved to give explicit expressions for $u(x,t+\Delta)$ and $c(x,t,+\Delta)$. All the high-order timestepping schemes may be implemented from these expressions.

References

1. E. Oran Brigham, The fast Fourier transform (Prentice-Hall, Englewood Cliffs, New Jersey, 1974).
2. S.D. Conte and C. de Boor, Elementary numerical analysis, 3rd edition (McGraw-Hill Kogakusha, Tokyo, 1980).
3. Heinz-Otto Kreiss and Joseph O1iger, "Comparison of accurate methods for the integration of hyperbolic equations", Tellus 24 (1972), 199-215.
4. B.J. Noye,"An introduction to finite difference techniques", Numerical simulation of fluid motion, 1-112 (Proc. Internat. Conf. Numerical Simulation of Fluid Dynamic Systems, Monash University, Clayton, Victoria, 1976. North-Holland, Amsterdam, New York, 1978).
5. Patrick J. Roache, Computational fluid dynamics (Hermosa, Albuquerque, New Mexico, 1972).

Department of Theoretical and Applied Mechanics,
School of Mathematics, University of New South Wales,
PO Box 1,
Kensington,
New South Wales 2033, Australia.