Chapter 8

The Von Neumann Method for Stability Analysis

Various methods have been developed for the analysis of stability, nearly all of them limited to linear problems. However, even within this restriction the complete investigation of stability for initial, boundary value problems can be extremely complicated, particularly in the presence of boundary conditions and their numerical representation.

The problem of stability for a linear problem with constant coefficients is now well understood when the influence of boundaries can be neglected or removed. This is the case either for an infinite domain or for periodic conditions on a finite domain. In the latter case we consider that the computational domain on the x-axis of length L is repeated periodically, and therefore all quantities, the solution, as well as the errors, can be developed in a finite Fourier series over the domain $2L$. This development in the frequency domain (in space) forms the basis of the *Von Neumann method* for stability analysis (Sections 8.1 and 8.2). This method was developed in Los Alamos during World War II by Von Neumann and was considered classified until its brief description in Cranck and Nicholson (1947) and in a publication in 1950 by Charney *et al.* (1950). At present this is the most widely applied technique for stability analysis, and furthermore allows an extensive investigation of the behaviour of the error as a function of the frequency content of the initial data and of the solution, as will be seen in Section 8.3. The generalization of the Von Neumann method to multidimensional problems is presented in Section 8.4.

If the problem of stability analysis can be treated generally for linear equations with constant coefficients and with periodic boundary conditions, as soon as we have to deal with non-constant coefficients and (or) non-linear terms in the basic equations the information on stability becomes very limited. Hence we have to resort to a local stability analysis, with frozen values of the non-linear and non-constant coefficients, to make the formulation linear. In any case, linear stability is a necessary condition for non-linear problems but it is certainly not sufficient. We will touch on this difficult problem in Section 8.5.

Finally, Section 8.6 presents certain general techniques in order to obtain the stability conditions from the Von Neumann analysis.

8.1 FOURIER DECOMPOSITION OF THE ERROR

If $\bar{u}_i^n$ is the exact solution of the difference equation and u_i^n the actual computed solution the difference might be due to round-off errors and to errors in the initial data. Hence,

$$u_i^n = \bar{u}_i^n + \varepsilon_i^n \tag{8.1.1}$$

where ε_i^n indicates the error at time level n in mesh point i. Clearly, any linear numerical scheme for u_i^n is satisfied exactly by $\bar{u}_i^n$, and therefore the errors ε_i^n are also solutions of the same discretized equation.

In order to present the essentials of the method we will first refer to the previous examples. Considering scheme (7.2.5) and inserting equation (8.1.1) leads to

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{\Delta t} + \frac{\varepsilon_i^{n+1} - \varepsilon_i^n}{\Delta t} = -\frac{a}{2\Delta x}(\bar{u}_{i+1}^n - \bar{u}_{i-1}^n) - \frac{a}{2\Delta x}(\varepsilon_{i+1}^n - \varepsilon_i^n) \tag{8.1.2}$$

Since $\bar{u}_i^n$ satisfies exactly equation (7.2.5) we obtain the equation for the errors ε_i^n:

$$\frac{\varepsilon_i^{n+1} - \varepsilon_i^n}{\Delta t} = -\frac{a}{2\Delta x}(\varepsilon_{i+1}^n - \varepsilon_{i-1}^n) \tag{8.1.3}$$

which is identical to the basic scheme. Hence the errors ε_i^n do evolve over time in the same way as the numerical solution u_i^n.

The general demonstration of this property is obvious when the operator form (equation (7.2.27)) is applied, considering the operator C to be linear. If e^n designates the column vector of the errors at time level n:

$$e^n = \begin{vmatrix} \cdot \\ \varepsilon_{i-1}^n \\ \varepsilon_i^n \\ \varepsilon_{i+1}^n \\ \cdot \end{vmatrix} \tag{8.1.4}$$

relation (8.1.1) can be written, with $\bar{U}^n$ indicating the exact solution,

$$U^n = \bar{U}^n + e^n \tag{8.1.5}$$

Inserting this equation into the basic scheme leads to

$$\bar{U}^{n+1} + e^{n+1} = C\bar{U}^n + Ce^n \tag{8.1.6}$$

or

$$e^{n+1} = Ce^n \tag{8.1.7}$$

by definition of $\bar{U}^n$ as a solution of

$$\bar{U}^{n+1} = C\bar{U}^n \tag{8.1.8}$$

Hence time evolution of the error is determined by the same operator C as the solution of the numerical problem.

If the boundary conditions are considered as periodic the error ε_i^n can be decomposed into a Fourier series in space at each time level n. Since the space domain is of a finite length we will have a discrete Fourier representation summed over a finite number of harmonics.

In a one-dimensional domain of length L the complex Fourier representation reflects the region $(0, L)$ onto the negative part $(-L, 0)$, and the fundamental frequency corresponds to the maximum wavelength of $\lambda_{max} = 2L$. The associated wavenumber $k = 2\pi/\lambda$ attains its minimum value $k_{min} = \pi/L$. On the other hand, the maximum value of the wavenumber k_{max} of the finite spectrum on the interval $(-L, L)$ is associated with the shortest resolvable wavelength on a mesh with spacing Δx. This shortest wavelength is clearly equal to $\lambda_{min} = 2\Delta x$ (see Figure 8.1.1), and consequently, $k_{max} = \pi/\Delta x$.

Therefore with the mesh index i, ranging from 0 to N, with $x_i = i \cdot \Delta x$ and

$$\Delta x = L/N \tag{8.1.9}$$

all the harmonics represented on a finite mesh are given by

$$k_j = jk_{min} = j\frac{\pi}{L} = j\frac{\pi}{N\,\Delta x} \qquad j = 0, 1, 2, \ldots, N \tag{8.1.10}$$

with the maximum value of j being associated with the maximum frequency. Hence with $k_{max} = \pi/\Delta x$ the highest value of j is equal to the number of mesh

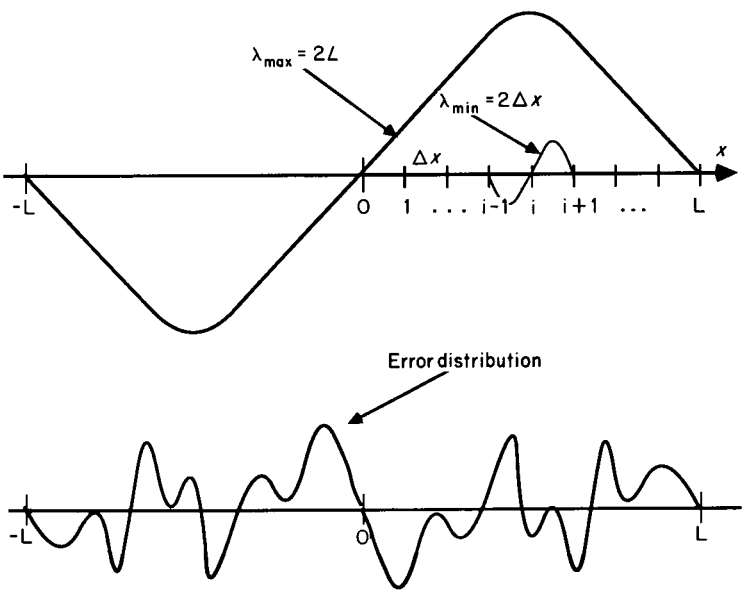

Figure 8.1.1 Fourier representation of the error on the interval $(-L, L)$

intervals N. Any finite mesh function, such as ε_i^n or the full solution u_i^n, will be decomposed into a Fourier series as

$$\varepsilon_i^n = \sum_{j=-N}^{N} E_j^n \, e^{Ik_j \cdot i\Delta x} = \sum_{j=-N}^{N} E_j^n \, e^{Iij\pi/N} \tag{8.1.11}$$

where $I = \sqrt{-1}$ and E_j^n is the amplitude of the jth harmonic.

The harmonic associated with $j = 0$ represents a constant function in space. The produce $k_j \Delta x$ is often represented as a phase angle:

$$\phi \equiv k_j \cdot \Delta x = \frac{j\pi}{N} \tag{8.1.12}$$

and covers the domain $(-\pi, \pi)$ in steps of π/N. The region around $\phi = 0$ corresponds to the low frequencies while the region close to $\phi = \pi$ is associated with the high-frequency range of the spectrum. In particular, the value $\phi = \pi$ corresponds to the highest frequency resolvable on the mesh, namely the frequency of the wavelength $2\Delta x$. Since we deal with linear schemes the discretized equation (8.1.7), which is satisfied by the error ε_i^n, must also be satisfied by each individual harmonic.

8.1.1 Amplification factor

Considering a single harmonic $E_j^n \, e^{Ii\phi}$, its time evolution is determined by the same numerical scheme as the full solution u_i^n. Hence inserting a representation of this form into equation (8.1.3) for the example considered we obtain, dropping the subscript j,

$$\frac{(E^{n+1} - E^n)}{\Delta t} e^{Ii\phi} + \frac{a}{2\Delta x} (E^n \, e^{I(i+1)\phi} - E^n \, e^{I(i-1)\phi})] = 0$$

or, dividing by $e^{Ii\phi}$,

$$(E^{n+1} - E^n) + \frac{\sigma}{2} E^n (e^{I\phi} - e^{-I\phi}) = 0 \tag{8.1.13}$$

where the parameter

$$\sigma = \frac{a\Delta t}{\Delta x} \tag{8.1.14}$$

has been introduced.

The stability condition (7.2.25) will be satisfied if the amplitude of any error harmonic E^n does not grow in time, that is, if the ratio

$$|G| \equiv \left| \frac{E^{n+1}}{E^n} \right| \leqslant 1 \qquad \text{for all } \phi \tag{8.1.15}$$

The quantity G, defined by,

$$G = \frac{E^{n+1}}{E^n} \tag{8.1.16}$$

is the *amplification factor*, and is a function of time step Δt, frequency and mesh size Δx. In the present case from equation (8.1.13) we have

$$G - 1 + \frac{\sigma}{2} \cdot 2I \sin \phi = 0$$

or

$$G = 1 - I\sigma \sin \phi \tag{8.1.17}$$

The stability condition (8.1.15) requires the modulus of G to be lower or equal to one. For the present example,

$$|G|^2 = 1 + \sigma^2 \sin^2 \phi \tag{8.1.18}$$

and is clearly never satisfied. Hence the centred scheme (7.2.5), for the convection equation with forward difference in time is *unconditionally unstable.*

Example of scheme (7.2.8): conditional stability

Inserting the single harmonic $E^n e^{Ii\phi}$ into scheme (7.2.8) written for the error we obtain

$$(E^{n+1} - E^n)e^{Ii\phi} + \sigma E^n(e^{Ii\phi} - e^{I(i-1)\phi}) = 0$$

or after division by $E^n e^{Ii\phi}$,

$$\begin{aligned} G &= 1 - \sigma + \sigma e^{-I\phi} \\ &= 1 - 2\sigma \sin^2 \phi/2 - I\sigma \sin \phi \end{aligned} \tag{8.1.19}$$

In order to analyse the stability of scheme (7.2.8), that is, the regions where the modulus of the amplification factor G is lower than one, a representation of G in the complex plane is a convenient approach. Writing ξ and η, respectively, for the real and imaginary parts of G we have

$$\begin{aligned} \xi &= 1 - 2\sigma \sin^2 \phi/2 = (1 - \sigma) + \sigma \cos \phi \\ \eta &= -\sigma \sin \phi \end{aligned} \tag{8.1.20}$$

which can be considered as parametric equations for G with ϕ as a parameter. We recognize the parametric equations of a circle centred on the real axis ξ at $(1 - \sigma)$ with radius σ.

In the complex plane of G the stability condition (8.1.15) states that the curve representing G for all values of $\phi = k\Delta x$ should remain within the unit circle (see Figure 8.1.2). It is clearly seen from Figure 8.1.2 that the scheme is stable for

$$0 < \sigma \leqslant 1 \tag{8.1.21}$$

Hence scheme (7.2.8) is conditionally stable and condition (8.1.21) is known as the *Courant–Friedrichs–Lewy* or *CFL* condition. The parameter σ is called

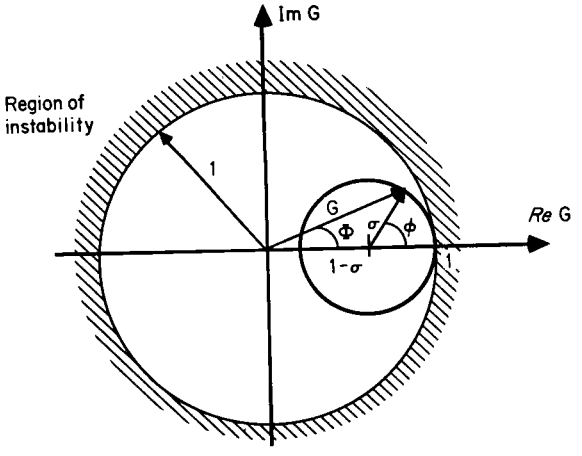

Figure 8.1.2 Complex G plane representation of upwind scheme (7.2.8), with unit circle defining the stability region

the *Courant number*. This condition for stability was introduced for the first time in 1928 in a paper by Courant *et al.* (1928), which can be considered as laying the foundations of the concepts of convergence and stability for finite difference schemes, although the authors were using finite difference concepts as a mathematical tool for proving existence theorems of continuous problems. Observe that the upwind scheme (7.2.8) is unstable for $a < 0$ (see also Problem 8.1).

8.1.2 Comment on the CFL condition

This fundamental stability condition of most explicit schemes for wave and convection equations expresses that the distance covered during the time

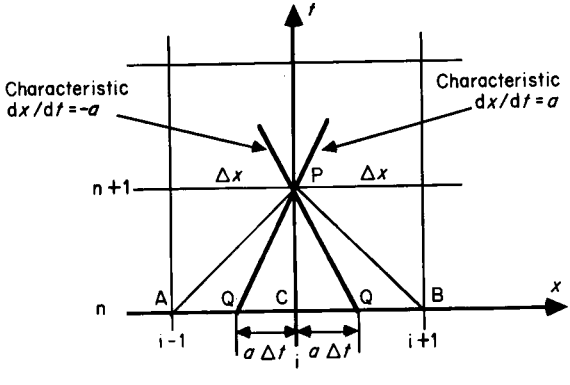

Figure 8.1.3 Geometrical interpretation of the CFL condition, $\sigma \leqslant 1$

interval Δt, by the disturbances propagating with speed a, should be lower than the minimum distance between two mesh points. Referring to Figure 8.1.3, the line PQ is the characteristic $dx/dt = a$, through P, and defines the domain of dependence of the differential equation in P. On the other hand, the difference equation defines a numerical domain of dependence of P which is the domain between PAC.

The CFL stability condition $\sigma \leqslant 1$ expresses that the mesh ratio $\Delta t/\Delta x$ has to be chosen in such a way that *the domain of dependence of the differential equation should be contained in the domain of dependence of the discretized equations.* In other words, the numerical scheme defining the approximation u_i^{n+1} in (mesh point i) must be able to include all the physical information which influences the behaviour of the system in this point.

Example of scheme (7.2.6): unconditional stability

The implicit, backward Euler scheme with central space differencing of the convection equation offers a third situation with respect to stability properties. Performing the same stability analysis with scheme (7.2.6), the error amplitude E^{n+1} becomes, after introduction of an harmonic of the form $E^n\, e^{I i\phi}$,

$$e^{I i\phi}(E^{n+1} - E^n) + \frac{\sigma}{2} E^{n+1}(e^{I\phi} - e^{-I\phi})e^{I i\phi} = 0$$

or

$$G - 1 + \frac{\sigma}{2} G(e^{I\phi} - e^{-I\phi}) = 0$$

leading to

$$G = \frac{1}{1 + I\sigma \sin \phi} \tag{8.1.22}$$

The modulus of G is always lower than one, for all values of σ, since

$$|G|^2 = G \cdot G^* = \frac{1}{1 + \sigma^2 \sin^2 \phi} \tag{8.1.23}$$

and therefore the implicit scheme (7.2.6) is unconditionally stable. Hence it is seen that schemes can have either conditional stability, unconditional stability or unconditional instability.

The Von Neumann method offers an easy and simple way of assessing the stability properties of linear schemes with constant coefficients when the boundary conditions are assumed periodic.

8.2 GENERAL FORMULATION OF VON NEUMANN'S METHOD: SYSTEM OF EQUATIONS

Referring to the second definition of stability (equation (7.2.35)), the Von Neumann method can be restated on the basis of the development of the

solution u_i^n in a Fourier series, that is, writing

$$u_i^n = \sum_{m=-N}^{N} v_m^n \, e^{Iik_m \Delta x} = \sum_{m=-N}^{N} v_m^n \, e^{Ii\phi} \tag{8.2.1}$$

where v_m^n is the amplitude of the mth harmonic of u_i^n. An arbitrary harmonic can be singled out and, when introduced into the scheme, stability requires that no harmonic should be allowed to increase in time without bound. Since u_i^n and the error ε_i^n satisfy the same numerical equation, the results obtained from equation (8.2.1) are identical to those obtained above. The amplification factor G is defined here as the ratio of the amplitudes v_m^n, that is, omitting the m subscript,

$$G = \frac{v^{n+1}}{v^n} = G(\phi, \Delta t, \Delta x) \tag{8.2.2}$$

and definition (7.2.35) leads to the stability condition (8.1.15).

In order to formulate the general Von Neumann stability condition it is necessary to write the discretized equations in operator and matrix forms.

8.2.1 Matrix and operator formulation

We consider that the numerical scheme is obtained in two steps: a space discretization, followed by a time integration.

(1) When a space discretization is applied (for instance, a finite difference method) the differential space operator is aproximated by a discretized space operator S, leading to the method of line formulation for the discrete values $u_i^n = u(\vec{x}_i, n\Delta t)$, where $\vec{x}_i$ is the co-ordinate of mesh point i:

$$\frac{du_i}{dt} = Su_i + q_i \tag{8.2.3}$$

The q_i term contains eventual sources and the contributions from boundary conditions. The matrix representation of the above system of ordinary differential equations in time is written with the vector U^n, defined by equation (7.2.26) as

$$\frac{dU}{dt} = SU + Q \tag{8.2.4}$$

where we use the same notation for the discretized space operator and its matrix representation.

(2) When a time-integration scheme is applied to the above space-discretized equations, corresponding to a two-level scheme connecting time levels $(n+1)$ and n, the numerical scheme associated with the differential problem generalizes equation (7.2.27):

$$u_i^{n+1} = C \cdot u_i^n + \bar{q}_i \tag{8.2.5}$$

or, in matrix form,

$$U^{n+1} = CU^n + \bar{Q} \tag{8.2.6}$$

where C can be considered as a discretization operator of the scheme.

For a two-level implicit scheme, of the form $B_1 U^{n+1} = B_0 U^n$ the difference operator C is defined by $C = B_1^{-1} B_0$. Note that for the Euler method we have $C = 1 + \Delta t S$. Some examples of the matrix representation of C have been given in Chapter 7 and we illustrate these various representations with a few additional examples.

The linear diffusion equation

$$\frac{\partial u}{\partial t} = a \frac{\partial^2 u}{\partial x^2} \tag{8.2.7}$$

The one-dimensional linearized shallow-water equations

These equations have been treated in Example 3.4.1 and we write here v for the x-component of the velocity, keeping the notation U for the column of the two dependent variables. The equations are linearized by setting v and h equal to v_0 and h_0 in the non-linear terms:

$$\left| \begin{array}{l} \dfrac{\partial h}{\partial t} + v_0 \dfrac{\partial h}{\partial x} + h_0 \dfrac{\partial v}{\partial x} = 0 \\ \dfrac{\partial v}{\partial t} + v_0 \dfrac{\partial v}{\partial x} + g \dfrac{\partial h}{\partial x} = 0 \end{array} \right. \tag{8.2.8}$$

Here the vector u is defined by

$$U = \left| \begin{array}{l} h \\ v \end{array} \right| \tag{8.2.9}$$

and the system is written as

$$\frac{\partial U}{\partial t} + A \frac{\partial U}{\partial x} = 0 \tag{8.2.10a}$$

where

$$A = \left| \begin{array}{ll} v_0 & h_0 \\ g & v_0 \end{array} \right| \tag{8.2.10b}$$

It is seen that, under this form, equation (8.2.10) generalizes the single convection equation.

Wave equation

$$\frac{\partial^2 w}{\partial t^2} - a^2 \frac{\partial^2 w}{\partial x^2} = 0 \tag{8.2.11a}$$

Table 8.1.

Differential equation	Space discretization operator S	Matrix representation of S (excluding boundary conditions)
Heat diffusion	**Second order central difference**	
$\dfrac{\partial u}{\partial t} = \alpha \dfrac{\partial^2 u}{\partial x^2}$	$\dfrac{du_i}{dt} = \dfrac{\alpha}{\Delta x^2}(u_{i+1} - 2u_i + u_{i-1})$	$S = \begin{vmatrix} +1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & \cdot & \end{vmatrix} \dfrac{\alpha}{\Delta x^2}$
$L = \alpha \dfrac{\partial^2}{\partial x^2}$	$= \dfrac{\alpha}{\Delta x^2}(E - 2 + E^{-1})u_i$	
	$S = \dfrac{\alpha}{\Delta x^2}(E - 2 + E^{-1})$	
Shallow water equation	**Central scheme**	
$\dfrac{\partial U}{\partial t} + A \dfrac{\partial U}{\partial x} = 0$	$\dfrac{du_i}{dt} = -A \dfrac{u_{i+1} - u_{i-1}}{2\,\Delta x}$	$U = \begin{vmatrix} \cdot \\ u_{i-1} \\ u_i \\ u_{i+1} \\ \cdot \end{vmatrix} = \begin{vmatrix} \cdot \\ \begin{vmatrix} v \\ h \end{vmatrix}_{i-1} \\ \begin{vmatrix} v \\ h \end{vmatrix}_{i} \\ \begin{vmatrix} v \\ h \end{vmatrix}_{i+1} \\ \cdot \end{vmatrix}$
$U = \begin{vmatrix} h \\ v \end{vmatrix}$	$S = \dfrac{-A}{2\,\Delta x}(E - E^{-1})$	$S = \begin{vmatrix} \cdot & & & \\ \dfrac{+A}{2\,\Delta x} & 0 & \dfrac{-A}{2\,\Delta x} & \\ & \dfrac{+A}{2\,\Delta x} & 0 & \dfrac{-A}{2\,\Delta x} \\ & & \cdot & \end{vmatrix}$
$A = \begin{vmatrix} v_0 & h_0 \\ g & v_0 \end{vmatrix}$		each element is a (2×2) matrix
$L = -A \dfrac{\partial}{\partial x}$		

Table 8.1. (*continued*)

Discretization operator C of time integrated scheme	Matrix representation of C	Amplification matrix
Euler method $\beta = \dfrac{\alpha\,\Delta t}{\Delta x^2}$ $u_i^{n+1} = u_i^n + \dfrac{\alpha\,\Delta t}{\Delta x^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n)$ $= [1 + \beta(E - 2 + E^{-1})]u_i^n$ $C \equiv 1 + \beta(E - 2 + E^{-1})$	$C = \begin{vmatrix} \cdot & \cdot & & & 0 \\ \beta & (1-2\beta) & \beta & & \\ & \beta & (1-2\beta) & \beta & \\ & & \beta & (1-2\beta) & \beta \\ 0 & & & \cdot & \cdot \end{vmatrix}$	$G = 1 + \beta(e^{I\phi} - 2 + e^{-I\phi})$ $G = 1 - 2\beta(1 - \cos\phi)$ $G = 1 - 4\beta \sin^2 \phi/2$
Trapezoidal (Cranck–Nicholson) method $u_i^{n+1} = u_i^n + \dfrac{\Delta t}{2} S u_i^n + \dfrac{\Delta t}{2} S u_i^{n+1}$ $\left(1 - \dfrac{\Delta t}{2} S\right) u_i^{n+1} = \left(1 + \dfrac{\Delta t}{2} S\right) u_i^{n+1}$ $B = 1 - \dfrac{\Delta t}{2} S,\ A = 1 + \dfrac{\Delta t}{2} S$ $C = B^{-1} \cdot A$	$A = \begin{vmatrix} \cdot & \cdot & \\ \beta/2 & (1-\beta) & \beta/2 \\ & \beta/2 & (1-\beta)\ \beta/2 \\ & & \cdot\ \cdot \end{vmatrix}$ $B = \begin{vmatrix} \cdot & \cdot & \\ -\beta/2 & (1+\beta) & -\beta/2 \\ & -\beta/2 & (1+\beta)\ -\beta/2 \\ & & \cdot\ \cdot \end{vmatrix}$ $C = B^{-1}A$	$G = \dfrac{1 + \dfrac{\beta}{2}(e^{I\phi} - 2 + e^{-I\phi})}{1 - \dfrac{\beta}{2}(e^{I\phi} - 2 + e^{-I\phi})}$
Euler method $u_i^{n+1} = u_i^n - \dfrac{A\,\Delta t}{2\,\Delta x}(u_{i+1}^n - u_{i-1}^n)$ $C = 1 - \dfrac{A\,\Delta t}{2\,\Delta x}(E - E^{-1})$ $C = \begin{vmatrix} 1 - \dfrac{v_0\,\Delta t}{2\,\Delta x}(E - E^{-1}) & -\dfrac{h_0\,\Delta t}{2\,\Delta x}(E - E^{-1}) \\ -\dfrac{g\,\Delta t}{\Delta x}\dfrac{1}{2}(E - E^{-1}) & 1 - \dfrac{v_0\,\Delta t}{2\,\Delta x}(E - E^{-1}) \end{vmatrix}$	$C = \begin{vmatrix} \cdot & \cdot & & \\ \dfrac{A\,\Delta t}{2\,\Delta x} & 1 & -\dfrac{A\,\Delta t}{2\,\Delta x} & \\ & \dfrac{A\,\Delta t}{2\,\Delta x} & 1 & -\dfrac{A\,\Delta t}{2\,\Delta x} \\ & & \dfrac{A\,\Delta t}{2\,\Delta x} & 1\ -\dfrac{A\,\Delta t}{2\,\Delta x} \\ & & & \cdot\ \cdot \end{vmatrix}$	$G = 1 - \dfrac{A\,\Delta t}{2\,\Delta x}(e^{I\phi} - e^{-I\phi})$ $G = \begin{vmatrix} 1 - I\dfrac{v_0\,\Delta t}{\Delta x}\sin\phi & -I\dfrac{h_0\,\Delta t}{\Delta x}\sin\phi \\ -I\dfrac{g\,\Delta t}{\Delta x}\sin\phi & 1 - I\dfrac{v_0\,\Delta t}{\Delta x}\sin\phi \end{vmatrix}$
Lax-Friedrichs scheme $u_i^{n+1} = (u_{i+1}^n + u_{i-1}^n)/2 - \dfrac{A\,\Delta t}{2\,\Delta x}(u_{i+1}^n - u_{i-1}^n)$ $C = \dfrac{1}{2}\left(1 - \dfrac{A\,\Delta t}{\Delta x}\right)E + \dfrac{1}{2}\left(1 + \dfrac{A\,\Delta t}{\Delta x}\right)E^{-1}$	$C = \begin{vmatrix} \cdot & \cdot & \cdot & \\ \dfrac{1}{2}\left(1 + \dfrac{A\,\Delta t}{\Delta x}\right) & 0 & \dfrac{1}{2}\left(1 - \dfrac{A\,\Delta t}{\Delta x}\right) & \\ & \dfrac{1}{2}\left(1 + \dfrac{A\,\Delta t}{\Delta x}\right) & 0\ \dfrac{1}{2}\left(1 - \dfrac{A\,\Delta t}{\Delta x}\right) & \\ & & \cdot\ \cdot\ \cdot & \end{vmatrix}$	$G = \dfrac{1}{2}\left(1 - \dfrac{A\,\Delta t}{\Delta x}\right)e^{I\phi} + \dfrac{1}{2}\left(1 + \dfrac{A\,\Delta t}{\Delta x}\right)e^{-I\phi}$ $G = \cos\phi - I\dfrac{\Delta t}{\Delta x} A \sin\phi$ $G = \begin{vmatrix} \cos\phi - \dfrac{v_0\,\Delta t}{\Delta x} I \sin\phi & -I\dfrac{\Delta t}{\Delta x} h_0 \sin\phi \\ -I\dfrac{\Delta t}{\Delta x} g \sin\phi & \cos\phi - v_0\dfrac{\Delta t}{\Delta x} I \sin\phi \end{vmatrix}$

(*continued*)

Table 8.1. (*continued*)

Differential equation	Space discretization operator S	Matrix representation of S (excluding boundary conditions)
Wave equation $\frac{\partial^2 w}{\partial t^2} - a^2 \frac{\partial^2 w}{\partial x^2} = 0$ or $\frac{\partial U}{\partial t} = A \frac{\partial U}{\partial x}$ $U = \begin{vmatrix} v \\ w \end{vmatrix}$ $A = \begin{vmatrix} 0 & a \\ a & 0 \end{vmatrix}$ $L = A \frac{\partial}{\partial x}$	**Central scheme** $\frac{du_i}{dt} = A \frac{u_{i+1} - u_{i-1}}{2\,\Delta x}$ $S = \frac{A}{2\,\Delta x}(E - E^{-1})$	$S = \begin{vmatrix} 0 & & \\ -\frac{A}{2\,\Delta x} & 0 & \frac{A}{2\,\Delta x} \\ & -\frac{A}{2\,\Delta x} 0 & \frac{A}{2\,\Delta x} \\ & 0 & \end{vmatrix}$
	Forward/backward scheme The two components v, w are discretized separately. $\frac{dv_i}{dt} = \frac{a}{\Delta x}(w_{i+1} - w_i)$ $\frac{dw_i}{dt} = \frac{a}{\Delta x}(v_i - v_{i-1})$ $S = \begin{vmatrix} 0 & a(E-1) \\ a(1-E^{-1}) & 0 \end{vmatrix}$ $S = E^{-1} \begin{vmatrix} 0 & 0 \\ -a & 0 \end{vmatrix} + \begin{vmatrix} 0 & -a \\ a & 0 \end{vmatrix} + \begin{vmatrix} 0 & a \\ 0 & 0 \end{vmatrix} E$ $= A_- E^{-1} + A_0 + A_+ E$	$S = \begin{vmatrix} A_- & A_0 & A_+ & \\ & A_- & A_0 & A_+ \\ & & A_- & A_0 \quad A_+ \end{vmatrix}$

with the initial boundary conditions, for $t = 0$,

$$w(x,0) = f(x)$$
$$\frac{\partial w}{\partial t}(x,0) = g(x) \tag{8.2.11b}$$

This wave equation is written as a system of first-order equations; for instance,

$$\frac{\partial v}{\partial t} = a \frac{\partial w}{\partial x}$$
$$\frac{\partial w}{\partial t} = a \frac{\partial v}{\partial x} \tag{8.2.12}$$

Defining

$$U = \begin{vmatrix} v \\ w \end{vmatrix} \tag{8.2.13}$$

we can write the system as

$$\frac{\partial U}{\partial t} = A \frac{\partial U}{\partial x} \tag{8.2.14a}$$

Table 8.1. (*continued*)

Discretization operator C of time integrated scheme	Matrix representation of C	Amplification matrix
Backward Euler		
$U_i^{n+1} = U_i^n + A\dfrac{U_{i+1}^{n+1} - U_{i-1}^{n+1}}{2\,\Delta x}$	$C^{-1} = \begin{vmatrix} \ddots & & & \\ \frac{A\,\Delta t}{2\,\Delta x} & 1 & \frac{-A\,\Delta t}{2\,\Delta x} & \\ & \frac{A\,\Delta t}{2\,\Delta x} & 1 & \frac{-A\,\Delta t}{2\,\Delta x} \\ & & \ddots & \end{vmatrix}$	$G^{-1} = 1 - \dfrac{A\,\Delta t}{2\,\Delta x}(e^{I\phi} - e^{-I\phi})$
$C^{-1} = 1 - \dfrac{A\,\Delta t}{2\,\Delta x}(E - E^{-1})$		$G^{-1} = 1 - I\dfrac{A\,\Delta t}{\Delta x}\sin\phi$
Forward Euler scheme—two step/ semi-implicit		
$v_i^{n+1} - v_i^n = \dfrac{a\,\Delta t}{\Delta x}(w_{i+1}^n - w_i^n)$	$C = \begin{vmatrix} \ddots & \ddots & & \\ C_- & C_0 & C_+ & \\ & C_- & C_0 & C_+ \\ & & \ddots & \ddots \end{vmatrix}$	$G = \begin{vmatrix} 1 & I\gamma e^{I\phi/2} \\ I\gamma e^{-I\phi/2} & 1-\gamma^2 \end{vmatrix}$
$w_i^{n+1} - w_i^n = \dfrac{a\,\Delta t}{\Delta x}(v_i^{n+1} - v_{i-1}^{n+1})$		$\gamma = 2\sigma\sin\phi/2$
or, with $\sigma = a\,\Delta t/\Delta x$	$C_0 = \begin{vmatrix} 1 & -\sigma \\ \sigma & 1-2\sigma^2 \end{vmatrix}$	
$C = \begin{vmatrix} 1 & \sigma(E-1) \\ (1-E^{-1})\sigma & 1+\sigma^2(1-E^{-1})(E-1) \end{vmatrix}$	$C_- = \begin{vmatrix} 0 & 0 \\ -\sigma & \sigma^2 \end{vmatrix}$	
$C = \begin{vmatrix} 1 & -\sigma \\ \sigma & 1-2\sigma^2 \end{vmatrix} + \begin{vmatrix} 0 & 0 \\ -\sigma & \sigma^2 \end{vmatrix} E^{-1}$	$C_+ = \begin{vmatrix} 0 & \sigma \\ 0 & \sigma^2 \end{vmatrix}$	
$+ \begin{vmatrix} 0 & \sigma \\ 0 & \sigma^2 \end{vmatrix} E$		

with

$$A = \begin{vmatrix} 0 & a \\ a & 0 \end{vmatrix} \tag{8.2.14b}$$

These operators are summarized in Table 8.1 for some representative schemes and the operators S and C are expressed as a function of the shift operator E defined in Chapter 4. Note that the matrix representation of the operators S and C of Table 8.1 do not contain the boundary points. This will be dealt with in Chapter 10.

8.2.2 The general Von Neumann stability condition

When a single harmonic is applied to scheme (8.2.5) the operator C will act on the space index i, since C can be considered as a polynomial in the displacement operator E, as can be seen from Table 8.1. Hence we obtain, inserting

$$u_i^n = v^n\, e^{Ii\phi} \tag{8.2.15}$$

into the homogeneous part of scheme (8.2.5),

$$e^{Ii\phi}\cdot v^{n+1} = C(E)\, e^{Ii\phi}\cdot v^n \equiv G(\phi)\cdot v^n\cdot e^{Ii\phi}$$

and after division by $e^{Ii\phi}$,

$$v^{n+1} = G(\phi) \cdot v^n = [G(\phi)]^n v^1 \tag{8.2.16}$$

with

$$G(\phi) = C(e^{I\phi}) \tag{8.2.17}$$

The matrix $G(\phi)$ is called the *amplification* matrix, and reduces to the previously defined amplification factor when there is only one equation to be discretized. Observe that $G(\phi)$ or $G(k)$ can be considered as the discrete Fourier symbol of the discretization operator C, and is obtained from C by replacing E^j by $e^{Ij\phi}$ (see Table 8.1 for several examples).

The stability condition (7.2.35) requires that the matrix $[G(\phi)]^n$ remains uniformly bounded for all values of ϕ. The bound of a matrix G is defined by the maximum value of the ratio of the two vector magnitudes

$$\| G \| = \underset{u \neq 0}{\text{Max}} \frac{| G \cdot u |}{| u |} \tag{8.2.18}$$

where $| u |$ is any vector norm. For instance, the L_2 norm is defined by the square root of the sum of the components squared $| u |_{L_2} = (| u_1 |^2 + \cdots + | u_p |^2)^{1/2}$ if u is a vector with p components.

Since G is a $(p \times p)$ matrix with p eigenvalues $\lambda_1, \ldots, \lambda_j, \ldots, \lambda_p$ obtained as solutions of the polynomial

$$\det | G - \Lambda I | = 0 \tag{8.2.19}$$

its spectral radius is defined by the modulus of the largest eigenvalue:

$$\rho(G) = \underset{j=1,p}{\text{Max}} | \lambda_j | \tag{8.2.20}$$

We have the following properties (see, for instance, Varga, 1962):

$$\| G \| \geqslant \underset{j}{\text{Max}} \frac{| Gg_j |}{| g_j |} = \underset{j}{\text{Max}} | \lambda_j | = \rho(G) \tag{8.2.21}$$

where g_i are the eigenvectors of G, and

$$\| G \|^n \geqslant \| G^n \| \geqslant \rho^n(G) \tag{8.2.22}$$

The Von Neumann necessary condition for stability can be stated as the condition that the spectral radius of the amplification matrix satisfies (Richtmyer and Morton, 1967)

$$\rho(G) \leqslant 1 + 0(\Delta t) \tag{8.2.23}$$

for finite Δt and for *all* values of ϕ, in the range $(-\pi, \pi)$. This condition is less severe than the previous one (equation (8.1.15)), which corresponds to a

condition

$$\rho(G) \leqslant 1 \tag{8.2.24}$$

The possibility for the spectral radius to be slightly higher than one for stability allows the treatment of problems where the exact solution grows exponentially (for instance, equation (7.1.5), with a source term q proportional to the temperature, $q = bT$, $b > 0$). However, in other cases condition (8.2.23) allows numerical modes to grow exponentially in time for finite values of Δt. Therefore the practical, or strict, stability condition (8.2.24) is recommended in order to prevent numerical modes growing faster than physical modes solution of the differential equation. (We will return to this important aspect in Chapter 10.) In this connection, when some eigenvalues are equal to one they would generate a growth of the form $\Delta t^{(m-1)}$, where m is the multiplicity. Hence eigenvalues $\lambda_j = 1$ should be simple.

Conditions (8.2.23) or (8.2.24) are also sufficient for stability if G is a normal matrix, that is, if G commutes with its Hermitian conjugate. In this case, equation (8.2.22) is valid with an equality sign in the L_2-norm, that is, $\| G \|_{L_2} = \rho(G)$ and $\| G^2 \|_{L_2} = \rho^2(G)$. In particular, for a single equation this is satisfied, and therefore condition (8.2.24) is sufficient and necessary for the stability of two-level schemes of linear equations with constant coefficients. Other cases for which the above condition is also sufficient for stability can be found in Richtmyer and Morton (1967).

Properties

(1) If G can be expressed as a polynomial of a matrix A, $G = P(A)$, then the spectral mapping theorem (Varga, 1962) states that

$$\lambda(G) = P(\lambda(A)) \tag{8.2.25}$$

where $\lambda(A)$ are the eigenvalues of A. For example, if G is of the form

$$G = 1 - I\alpha A + \beta A^2$$

then

$$\lambda(G) = 1 - I\alpha\lambda(A) + \beta\lambda^2(A)$$

(2) If G can be expressed as a function of several *commuting* matrices the above property remains valid. That is, if

$$G = P(A, B) \qquad \text{with } AB = BA \tag{8.2.26}$$

the two matrices have the same set of eigenvectors, and

$$\lambda(G) = P(\lambda(A), \lambda(B)) \tag{8.2.27}$$

This property ceases to be valid when the matrices do not commute. Unfortunately this is the case for the system of flow equations in two and three dimensions. Therefore additional conjectures have to be introduced in order to

derive stability conditions for schemes applied to the linearized flow equations in multi-dimensions. More details will be found in Volume 2 when dealing with the discretization of Euler equations.

Note that this condition of strict stability is called *zero* stability by Lambert (1973) when applied to the discretization of initial value problems in systems of ordinary differential equations (see also Chapter 11).

Example 8.2.1 Shallow-water equations

Referring to Table 8.1 we deduce readily the amplification matrix for the two schemes considered. The steps can easily be followed and we leave it to the reader to reproduce this table as an exercise.

Euler method: For the Euler method in time the amplification factor is

$$G = \begin{vmatrix} 1 - I\dfrac{v_0 \Delta t}{\Delta x} \sin \phi & -h_0 \dfrac{\Delta t}{\Delta x} I \sin \phi \\ -g \dfrac{\Delta t}{\Delta x} I \sin \phi & 1 - I\dfrac{v_0 \Delta t}{\Delta x} \sin \phi \end{vmatrix} \tag{E8.2.1}$$

The stability condition (8.2.24) requires a knowledge of the eigenvalues of G, and these are obtained from

$$[\lambda - (1 - I\sigma_0 \sin \phi)]^2 + \sigma^2 \sin^2 \phi = 0 \tag{E8.2.2}$$

where

$$\sigma_0 = \frac{v_0 \Delta t}{\Delta x} \tag{E8.2.3}$$

$$\sigma = (gh_0)^{1/2} \frac{\Delta t}{\Delta x} \tag{E8.2.4}$$

Hence the two eigenvalues are

$$\lambda_{\pm} = 1 - I(\sigma_0 \pm \sigma) \sin \phi \tag{E8.2.5}$$

and the spectral radius is given by

$$\rho(G) = |\lambda_+| = 1 + \left(\frac{\Delta t}{\Delta x}\right)^2 (v_0 + \sqrt{(gh_0)})^2 \sin^2 \phi \geqslant 1 \tag{E8.2.6}$$

The scheme is therefore *unstable*, as might be expected from the previous analysis of the central, Euler scheme for the convection equation.

Lax–Friedrichs scheme: This scheme was introduced by Lax (1954) as a way of stabilizing the unstable, forward in time, central scheme of the previous example. It consists of replacing u_l^n in the right-hand side by the average value $(u_{l+1}^n + u_{l-1}^n)/2$, maintaining the scheme as first order in time and space. It is

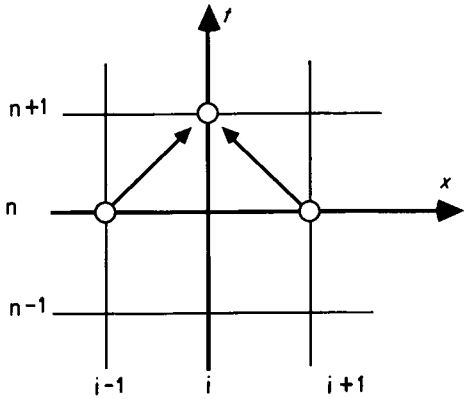

Figure 8.2.1 Lax–Friedrichs scheme for convection equations

schematically represented in Figure 8.2.1:

$$U_i^{n+1} = \frac{1}{2}(U_{i+1}^n + U_{i-1}^n) - \frac{\Delta t}{2\Delta x} A(U_{i+1}^n - U_{i-1}^n) \qquad \text{(E8.2.7)}$$

The reader can deduce the amplification matrix following the steps of Table 8.1, obtaining

$$G = \begin{vmatrix} \cos\phi - \sigma_0 I \sin\phi & -I\dfrac{\Delta t}{\Delta x} h_0 \sin\phi \\ -I\dfrac{\Delta t}{\Delta x} g \sin\phi & \cos\phi - \sigma_0 I \sin\phi \end{vmatrix} \qquad \text{(E8.2.8)}$$

The eigenvalues λ of G are given by

$$(\lambda - \cos\phi + \sigma_0 I \sin\phi)^2 + \sigma^2 \sin^2\phi = 0$$

or

$$\lambda_\pm = \cos\phi - I(\sigma_0 \pm \sigma)\sin\phi \qquad \text{(E8.2.9)}$$

The spectral radius is given by

$$\rho(G) = |\lambda_+| = [\cos^2\phi + (\sigma_0 + \sigma)^2 \sin^2\phi]^{1/2} \qquad \text{(E8.2.10)}$$

The stability condition $\rho(G) \leqslant 1$ will be satisfied if (for $v_0 > 0$)

$$(\sigma_0 + \sigma) \leqslant 1$$

or

$$(v_0 + \sqrt{(gh_0)})\frac{\Delta t}{\Delta x} \leqslant 1 \qquad \text{(E8.2.11)}$$

This is the CFL condition for the wave speed $(v_0 + \sqrt{(gh_0)})$, which is the largest eigenvalue of A.

Example 8.2.2 Second-order wave equation $(\partial^2 w/\partial t^2) - a^2(\partial^2 w/\partial x^2) = 0$

The forward–backward scheme, with semi-implicit time integration, of Table 8.1:

$$\begin{aligned} v_i^{n+1} - v_i^n &= \frac{a\Delta t}{\Delta x}(w_{i+1}^n - w_i^n) \\ w_i^{n+1} - w_i^n &= \frac{a\Delta t}{\Delta x}(v_i^{n+1} - v_{i-1}^{n+1}) \end{aligned} \tag{E8.2.12}$$

is equivalent to the three-level, centred scheme for the second-order wave equation, that is, to the scheme

$$w_i^{n+1} - 2w_i^n + w_i^{n-1} = \sigma^2(w_{i+1}^n - 2w_i^n + w_{i-1}^n) \tag{E8.2.13}$$

where $\sigma = a\Delta t/\Delta x$ (see also Problems 8.3 and 8.4). The amplification matrix is obtained from Table 8.1 as

$$G = \begin{vmatrix} 1 & I\gamma\, e^{I\phi/2} \\ I\gamma\, e^{-I\phi/2} & 1-\gamma^2 \end{vmatrix} \tag{E8.2.14}$$

where

$$\gamma = 2\sigma \sin \phi/2 \tag{E8.2.15}$$

The eigenvalues of G are obtained from

$$(1-\lambda)(1-\gamma^2-\lambda) + \gamma^2 = 0$$

leading to the two solutions

$$\lambda_\pm = \tfrac{1}{2}[(2-\gamma^2) \pm I\gamma\sqrt{(4-\gamma^2)}] \tag{E8.2.16}$$

For $\gamma^2 > 4$, that is, for $|\sigma \sin \phi/2| > 1$ or $|\sigma| > 1$, the spectral radius

$$\rho(G) = |\lambda_+| > 1$$

and the scheme is unstable. On the other hand, when $\gamma^2 \leqslant 4$, that is, for

$$\begin{aligned} |\sigma| &\leqslant 1 \\ \rho(G) &= |\lambda_+| = 1 \end{aligned} \tag{E8.2.17}$$

the scheme is stable, although only marginally, since the norm of G is equal to one.

For negative values of a^2, that is, for negative values of σ^2, the wave equation becomes elliptic:

$$\frac{\partial^2 w}{\partial t^2} + |a^2| \frac{\partial^2 w}{\partial x^2} = 0 \tag{E8.2.18}$$

and the scheme

$$(w_{i+1}^{n+1} - 2w_i^n + w_{i-1}^{n-1}) + |\sigma^2|(w_{i+1}^n - 2w_i^n + w_{i+1}^n) = 0 \tag{E8.2.19}$$

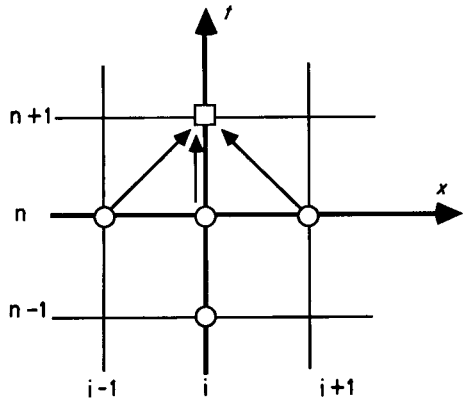

Figure 8.2.2 Unstable resolution scheme for Laplace equation

is *unstable*. Indeed, when a^2 is negative, the positive eigenvalue λ_+ becomes

$$\lambda_+ = 1 + 2|\sigma|^2\sin^2\phi/2 + 2|\sigma|\sin\phi/2\sqrt{(1+|\sigma|^2\sin^2\phi/2)} \geqslant 1 \quad \text{(E8.2.20)}$$

This shows that an elliptic problem cannot be treated numerically as an initial value problem. This is not surprising, since it is known that the Cauchy or initial value problem is not well posed for an elliptic equation (see, for instance, Courant and Hilbert, 1962, Volume II).

Observe that the above scheme, with $|a^2| = 1$, is the five-point difference operator for the Laplace equation, in the space (x, t). This scheme, as it stands, can be solved in a stable way for the associated boundary value problem, say on a rectangle $0 \leqslant x \leqslant L, 0 \leqslant t \leqslant T$, with any of the methods to be described in Chapter 12.

What the above results show is that the numerical solution of the elliptic problem cannot be obtained by a propagation from the points indicated by a circle in Figure 8.2.2 towards the point $(i, n+1)$. Such an algorithm is basically unstable. A resolution method for elliptic equations based on this marching scheme has nevertheless been developed by Roache (1971) and is called the error vector propagation method (EVP). This is based on a computation of the error generated in the marching procedure from $t = 0$ to $t = T$ and a comparison with the imposed boundary condition on $t = T$. However, this method cannot be stabilized when the number of grid points in the marching direction increases (McAveney and Leslie, 1972). The reader will find a recent account of this approach in the monograph edited by Book (1981), chapter 7 by Madela and McDonald.

8.3 THE SPECTRAL ANALYSIS OF NUMERICAL ERRORS

The amplification matrix G allows, next to an assessment of stability, an evaluation of the frequency distribution of the discretization errors generated by the numerical scheme. Definition (8.2.16) of the amplification matrix

defines the numerical representation of the time evolution of the solution, and the amplitude v^n of the harmonic corresponding to the wavenumber k can be written as

$$v^n = \hat{v}\, e^{-I\omega t_n} = \hat{v}\, e^{-I\omega \cdot n\Delta t} \tag{8.3.1}$$

where $\omega = \omega(k)$ is a complex function of the real number k, representing the numerical dispersion relation. The function $\hat{v}(k)$ is obtained from the Fourier decomposition of the initial solution, since for $u(x, 0) = f(x)$ at $t = 0$ we have, assuming that the initial solution is represented exactly in the numerical scheme, with the exception of round-off errors:

$$\hat{v}(k) = \frac{1}{2L} \int_{-L}^{L} f(x)\, e^{-Ikx}\, dx \tag{8.3.2}$$

Actually, this defines the harmonic k of the solution u_i^n following equation (8.2.15) as

$$(u_i^n)_k = \hat{v}(k)\, e^{-I\omega(n\Delta t)}\, e^{Ik(i\Delta x)} \tag{8.3.3}$$

and is a discrete formulation of the single-wave representation applied in equation (3.4.13). In this latter form the exact solution is represented as

$$\tilde{u}_i^n = \hat{v}\, e^{-I\tilde{\omega}\Delta t\, n}\, e^{Ik(i\Delta x)} \tag{8.3.4}$$

As seen in Chapter 3, the exact dispersion relation $\tilde{\omega} = \tilde{\omega}(k)$ can be obtained from the differential system as a solution of the eigenvalue equation (3.4.20), while the approximate relation between ω and k, obtained from the amplification matrix G, is the numerical dispersion relation of the scheme.

From equation (8.2.16) we have

$$v^n = G^n \cdot v^0 = G^n \cdot \hat{v} = e^{-I\omega n\Delta t} \cdot \hat{v} \tag{8.3.5}$$

and G can be written as

$$G = e^{-I\omega\Delta t} \tag{8.3.6}$$

A comparison with the exact amplification function

$$\tilde{G} = e^{-I\tilde{\omega}\Delta t} \tag{8.3.7}$$

will allow us to investigate the nature and frequency spectrum of the numerical errors. Since ω is a complex function the amplification matrix can be separated into an amplitude $|G|$ and a phase Φ. With

$$\omega = \xi + I\eta \tag{8.3.8}$$

we have

$$\begin{aligned} G &= e^{+\eta\Delta t} \cdot e^{-I\xi\Delta t} \\ &= |G|\, e^{-I\Phi} \end{aligned} \tag{8.3.9a}$$

where

$$\begin{aligned} |G| &= e^{\eta\Delta t} \\ \Phi &= \xi\Delta t \end{aligned} \tag{8.3.9b}$$

A similar decomposition, performed for the exact solution

$$\tilde{\omega} = \tilde{\xi} + I\tilde{\eta} \tag{8.3.10}$$

following equation (3.4.24), leads to

$$|\tilde{G}| = e^{\tilde{\eta}\Delta t} \qquad \text{and} \qquad \tilde{\Phi} = \tilde{\xi}\Delta t \tag{8.3.11}$$

The error in amplitude, called the *diffusion* or *dissipation* error, is defined by the ratio of the computed amplitude to the exact amplitude:

$$\varepsilon_D = \frac{|G|}{e^{\tilde{\eta}\Delta t}} \tag{8.3.12}$$

The error on the phase of the solution, the *dispersion* error, can be defined as the difference

$$\varepsilon_\phi = \Phi - \tilde{\Phi} \tag{8.3.13}$$

suitable for pure parabolic problems, where $\tilde{\Phi} = 0$, in the absence of convective terms. For convection-dominated problems the definition

$$\varepsilon_\phi = \Phi/\tilde{\Phi} \tag{8.3.14}$$

is better adapted. In particular, for hyperbolic problems such as the scalar convection equation (7.2.1) the exact solution is a single wave propagating with the velocity a. Hence

$$\tilde{\phi} = ka\Delta t \tag{8.3.15}$$

8.3.1 Error analysis for parabolic problems

Let us consider as an example the error analysis for the explicit central discretization of the heat diffusion equation (8.2.7). Consider the explicit scheme, with space-centred differences

$$u_i^{n+1} = u_i^n + \frac{\alpha\Delta t}{\Delta x^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n) \tag{8.3.16}$$

the amplification factor is obtained from Table 8.1 as

$$G = 1 - 4\beta \sin^2\phi/2 \tag{8.3.17}$$

with

$$\beta = \frac{\alpha\Delta t}{\Delta x^2} \tag{8.3.18}$$

The stability condition is

$$|1 - 4\beta \sin^2\phi/2| \leqslant 1$$

which is satisfied for

$$1 \geqslant (1 - 4\beta \sin^2\phi/2) \geqslant -1$$

that is,

$$0 \leqslant \beta \leqslant 1/2 \tag{8.3.19}$$

Hence the above scheme is stable for

$$\alpha \geqslant 0 \qquad \text{and} \qquad \beta = \frac{\alpha \Delta t}{\Delta x^2} \leqslant \frac{1}{2} \tag{8.3.20}$$

The first condition expresses the stability of the physical problem, since for $\alpha < 0$ the analytical solution is exponentially increasing with time.

The exact solution corresponding to a wavenumber k is obtained by searching a solution of the type

$$\tilde{u} = \hat{v}\, e^{-I\tilde{\omega}t}\, e^{Ikx} \tag{8.3.21}$$

Inserting into equation (8.2.7) we have

$$\tilde{\omega}(k) = -I\alpha k^2 = -I\beta \cdot \phi^2/\Delta t \tag{8.3.22}$$

The exact solution of this parabolic problem is associated with a purely imaginary eigenvalue $\tilde{\omega}$, that is, with an exponential decay in time of the initial amplitude if $\alpha > 0$:

$$\tilde{u} = \hat{v}\, e^{Ikx}\, e^{-\alpha k^2 t} \tag{8.3.23}$$

Hence the error in the amplitude is measured by the ratio

$$\varepsilon_D = \frac{1 - 4\beta \sin^2 \phi/2}{e^{-\beta \cdot \phi^2/\Delta t}} \tag{8.3.24}$$

Expanding in powers of ϕ we obtain

$$\varepsilon_D = \frac{1 - \beta\phi^2 + \beta\phi^4/12 + \cdots}{1 - \beta\phi^2 + (\beta^2\phi^4/2) + \cdots} \approx 1 - \frac{\beta^2\phi^4}{2} + \frac{\beta\phi^4}{12} + \cdots$$

$$\approx 1 - \frac{\alpha^2 k^4 \Delta t^2}{2} + \frac{\alpha k^4}{12}\, \Delta t \Delta x^2 \tag{8.3.25}$$

For the low frequencies ($\phi \approx 0$) the error in amplitude remains small; while at high frequencies ($\phi \approx \pi$) the error could become unacceptably high, particularly for the larger values of $\beta \leqslant 1/2$. However, for $\beta = 1/6$ the two first terms of the expansion cancel, and the error is minimized, becoming of higher order, namely of the order $0(\Delta t^2, \Delta x^4)$ for constant values of $\beta = \alpha \Delta t/\Delta x^2$ and proportional to k^6.

Since G is real there is no error in phase, that is, there is no dispersive error for this scheme. It is seen that the error is proportional to the fourth and sixth power of the wavenumber, indicating that the high frequencies are computed with large errors. However, the amplitudes of these high frequencies are strongly damped since they are equal to $e^{-\alpha k^2 t}$. Therefore this will generally not greatly affect the overall accuracy, with the exception of situations where the initial solution $u(x, 0)$ contains a large number of high-frequency components (see also Problem 8.5).

8.3.2 Error analysis for hyperbolic problems

A hyperbolic problem such as the convection equation $u_t + au_x = 0$ represents a wave travelling at constant speed without damping, that is, with constant amplitude. The exact solution for a wave of the form $u = e^{-I\tilde{\omega}t} e^{Ikx}$ is given by

$$\tilde{u} = e^{-Ikat} e^{Ikx} \tag{8.3.26}$$

Hence the exact amplification function is defined by the real value of $\tilde{\omega}$:

$$\begin{aligned} \tilde{\omega} &= ka = \tilde{\xi} \\ \tilde{\eta} &= 0 \end{aligned} \tag{8.3.27}$$

The error in amplitude will be given by the modulus of the amplification factor

$$\varepsilon_D = |G| \tag{8.3.28}$$

and the error in phase (the dispersive error) is defined by

$$\varepsilon_\phi = \frac{\Phi}{ka\,\Delta t} = \frac{\Phi}{\sigma\phi} \tag{8.3.29}$$

An initial sinusoidal wave will be damped in the numerical simulation by a factor $|G|$ per time step and its propagation speed will be modified by the dispertion error ε_ϕ. When this ratio is larger than one ($\varepsilon_\phi > 1$) the phase error is a *leading* error and the numerical computed wave speed, $\bar{a}$, is larger than the exact speed, since

$$\bar{a} = \Phi/(k\Delta t) = a\Phi/(\sigma\phi) \tag{8.3.30}$$

and

$$\varepsilon_\phi = \bar{a}/a \tag{8.3.31}$$

This means that the computed waves appear to travel faster than the physical waves. On the other hand, when $\varepsilon_\phi < 1$ the phase error is said to be a *lagging* error, and the computed waves travel at a lower velocity than the physical ones.

Example 8.3.1 Lax–Friedrichs scheme for the convection equation

Applying the Lax–Friedrichs scheme to the single convection equation (see Table 8.1) leads to

$$u_i^{n+1} = \frac{1}{2}(u_{i+1}^n + u_{i-1}^n) - \frac{\sigma}{2}(u_{i+1}^n - u_{i-1}^n) \tag{E8.3.1}$$

The amplification factor is obtained by inserting a single harmonic $v^n e^{Iki\Delta x}$:

$$G = \cos\phi - I\sigma \sin\phi \tag{E8.3.2}$$

leading to the CFL stability condition $|\sigma| \leqslant 1$.

The accuracy of the scheme is obtained from the modulus and phase of the

amplification factor:

$$|G| = |\cos^2\phi + \sigma^2\sin^2\phi|^{1/2}$$
$$\Phi = \tan^{-1}(\sigma \tan \phi) \tag{E8.3.3}$$

This defines the dissipation error

$$\varepsilon_D = |G| = |\cos^2\phi + \sigma^2\sin^2\phi|^{1/2} \tag{E8.3.4}$$

and the dispersion error

$$\varepsilon_\phi = \frac{\Phi}{\sigma\phi} = \frac{\tan^{-1}(\sigma \tan \phi)}{\sigma\phi} \tag{E8.3.5}$$

As can be seen, the choice $\sigma = 1$ gives the exact solution, but lower values of σ will generate amplitude and phase errors.

Two equivalent graphical representations for the amplification factor are applied in practice. Cartesian representation of $|G|$ and ε_ϕ as a function of the parameter $\phi = k\Delta x$, ranging from 0 to π or a polar representation for $|G|$ and

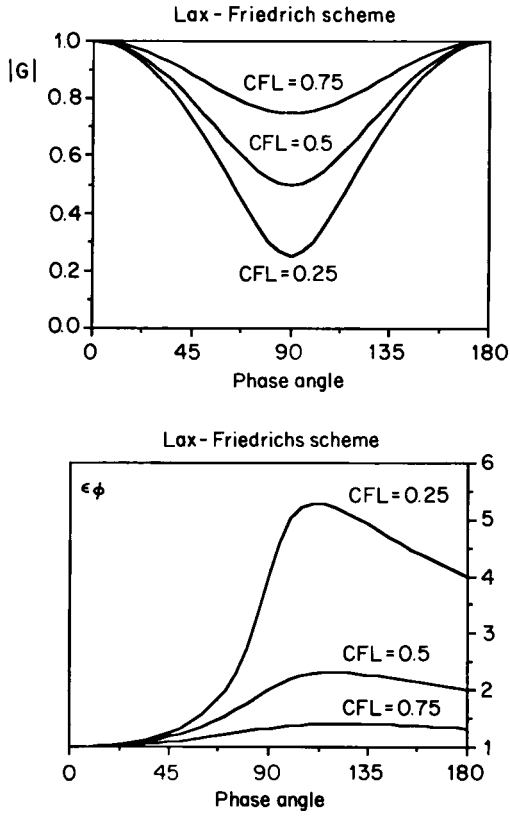

Figure 8.3.1 Amplitude and phase errors for Lax–Friedrichs scheme applied to the convection equation

ε_ϕ, where ϕ is represented as the polar angle. Figure 8.3.1 shows the Cartesian representation of $|G|$ and ε_ϕ for the Lax–Friedrichs scheme. For small values of σ the waves are strongly damped, indicating that this scheme is generating a strong numerical dissipation. The phase error is everywhere larger or equal to one, showing a leading phase error, particularly for $\phi = \pi$, $\varepsilon_\phi = 1/\sigma$ (see also Problem 8.6).

Example 8.3.2 Explicit upwind scheme (7.2.8)

The amplification factor for this scheme is defined by equation (8.1.19). Its modulus is given by

$$|G| = [(1 - \sigma + \sigma \cos \phi)^2 + \sigma^2 \sin^2 \phi]^{1/2} = [1 - 4\sigma(1 - \sigma)\sin^2 \phi/2]^{1/2} \tag{E8.3.6}$$

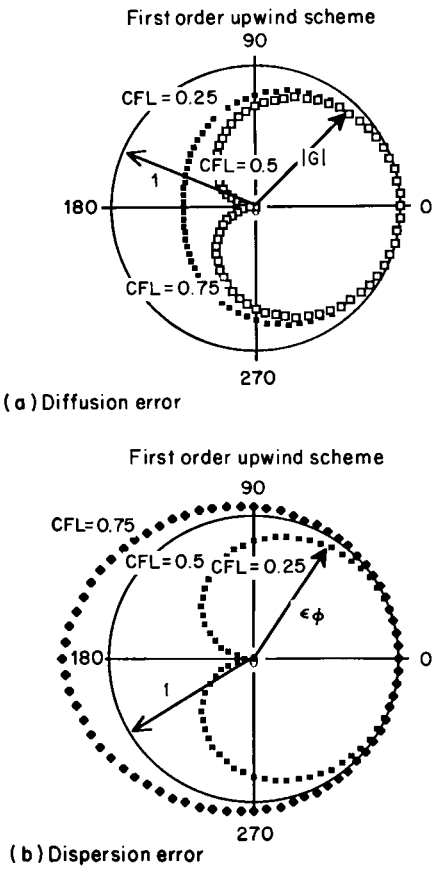

Figure 8.3.2 Polar representation of amplitude and phase errors for the upwind scheme applied to the convection equation

and the phase error is

$$\varepsilon_\phi = \frac{\tan^{-1}[(\sigma \sin \phi)/(1 - \sigma + \sigma \cos \phi)]}{\sigma\phi} \tag{E8.3.7}$$

A polar representation is shown in Figure 8.3.2.

For $\sigma = 0.5$ the phase error $\varepsilon_\phi = 1$, but for $\sigma < 0.5, \varepsilon_\phi < 1$, indicating a lagging error, while the numerical speed of propagation becomes larger than the physical speed, $\varepsilon_\phi > 1$ for Courant numbers $\sigma > 0.5$ (see also Problem 8.7).

Example 8.3.3 The Lax–Wendroff scheme for the convection equation

The schemes of the two previous examples are of first-order accuracy, which is generally insufficient for practical purposes. The first second-order scheme for the convection equation with two time levels is due to Lax and Wendroff (1960). The original derivation of Lax and Wendroff was based on a Taylor expansion in time up to the third order such to achieve second-order accuracy. In the development

$$u_i^{n+1} = u_i^n + \Delta t(u_t)_i + \frac{\Delta t^2}{2}(u_{tt})_i + 0(\Delta t^3) \tag{E8.3.8}$$

the second derivative is replaced by

$$u_{tt} = a^2 u_{xx} \tag{E8.3.9}$$

leading to

$$u_i^{n+1} = u_i^n - a\,\Delta t(u_x)_i + \frac{a^2\,\Delta t^2}{2}(u_{xx})_i + 0(\Delta t^3) \tag{E8.3.10}$$

When this is discretized centrally in mesh point i we obtain

$$u_i^{n+1} = u_i^n - \frac{\sigma}{2}(u_{i+1}^n - u_{i-1}^n) + \frac{\sigma^2}{2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n) \tag{E.8.3.11}$$

As can be seen, the third term, which stabilizes the instability generated by the first two terms, is the discretization of an additional dissipative term of the form $(a^2\,\Delta t/2)u_{xx}$.

The amplification matrix from the Von Neumann method is

$$G = 1 - I\,\sigma \sin \phi - \sigma^2(1 - \cos \phi) \tag{E.8.3.12}$$

In the complex G-plane this represents an ellipse centred on the real axis at the abscissa $(1 - \sigma^2)$ and having a semi-axis length of σ^2 along the real axis and σ along the vertical axis. Hence this ellipse will always be contained in the unit circle if the CFL condition is satisfied (Figure 8.3.3). For $\sigma = 1$ the ellipse becomes identical to the unit circle. The stability condition is therefore

$$|\sigma| \leqslant 1 \tag{E8.3.13}$$

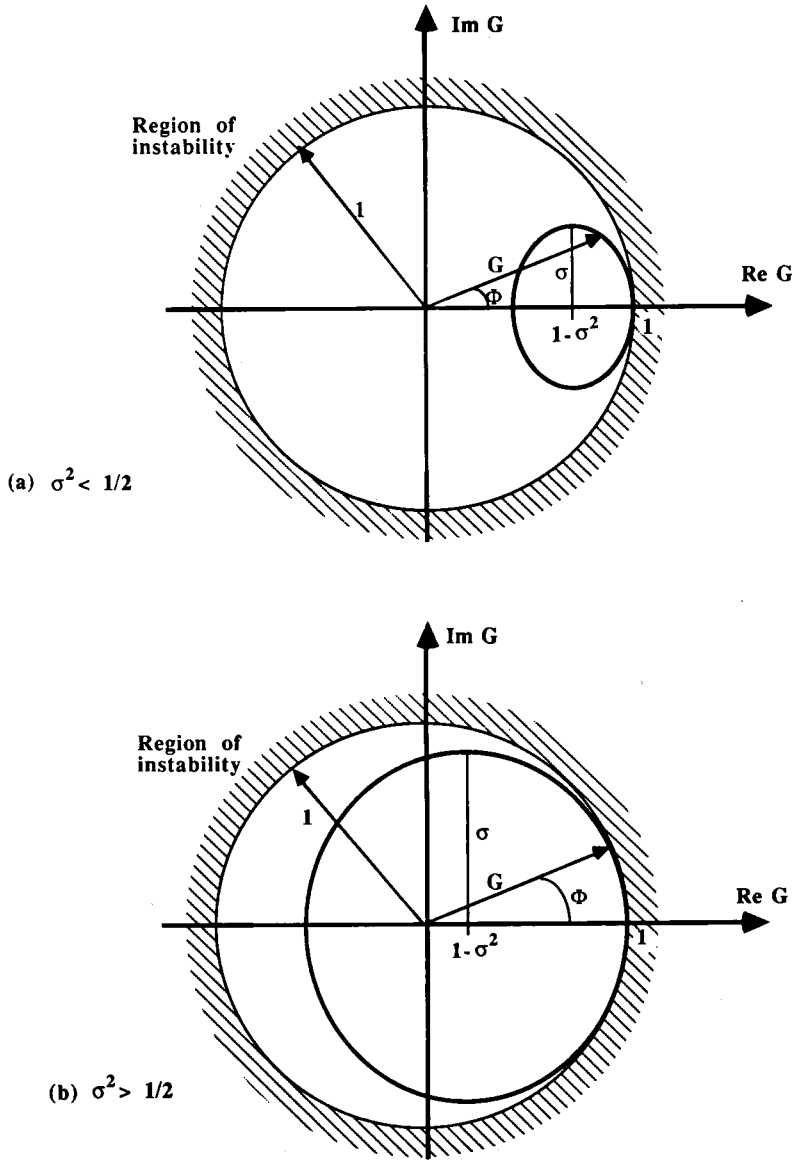

Figure 8.3.3 Polar representation of the amplification factor for Lax–Wendroff scheme. (a) $\sigma < 1$ and $\sigma^2 < 1/2$; (b) $\sigma < 1$ and $\sigma^2 > 1/2$

The dissipation error is given by

$$|G|^2 = 1 - 4\sigma^2(1-\sigma^2)\sin^4\phi/2 \qquad \text{(E8.3.14)}$$

and the phase error by

$$\varepsilon_\phi = \frac{\tan^{-1}[(\sigma \sin\phi)/(1-2\sigma^2\sin^2\phi/2)]}{\sigma\phi} \qquad \text{(E8.3.15)}$$

To the lowest order we have

$$\varepsilon_\phi \approx 1 - \tfrac{1}{6}(1-\sigma^2)\phi^2 + 0(\phi^4) \qquad \text{(E8.3.16)}$$

This relative phase error is mostly lower than one, indicating a dominating lagging phase error. On the high-frequency end the phase angle Φ goes to zero if $\sigma^2 < 1/2$ and tends to π if $\sigma^2 > 1/2$. These diffusion and dispersion errors are represented in Figure 8.3.4.

The phase error is the largest at the high frequencies, hence this will tend to accumulate high-frequency errors (for instance, those generated at a moving

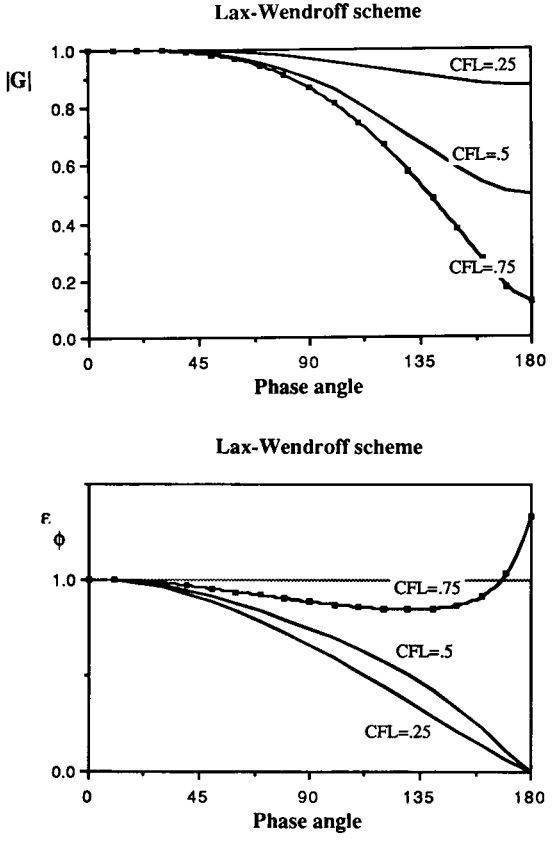

Figure 8.3.4 Dispersion and diffusion errors for Lax–Wendroff scheme

discontinuity). When the linear equation $u_t + au_x = 0$ is solved for a propagating discontinuity, oscillations will appear being the shock as can be seen from Figure 8.3.6 (to be discussed in Section 8.3.4, which compares the results computed with four different schemes).

8.3.3 Extension to three-level schemes

The properties of the amplification factor in the previous sections were based on two-level schemes, allowing a straightforward definition of G. However, many schemes can be defined which involve more than two time levels, particularly when the time derivatives are discretized with central difference formulas. A general form, generalizing equations (8.2.6), would be

$$U^{n+1} + b_0 U^n + b_1 U^{n-1} = CU^n + \tilde{Q} \tag{8.3.32}$$

For instance, for the convection equation $u_t + au_x = 0$ and a central difference in space we can define a scheme

$$\frac{u_i^{n+1} - u_i^{n-1}}{2\Delta t} = -\frac{a}{2\Delta x}(u_{i+1}^n - u_{i-1}^n) \tag{8.3.33}$$

which is second-order accurate in space and time. This scheme is known as the *leapfrog* scheme, because of the particular structure of its computational molecule (Figure 8.3.5) where the nodal value u_i^n does *not* contribute to the computation of u_i^{n+1}.

This scheme treats three levels simultaneously and, in order to start the calculation, two time levels $n = 0$ and $n = 1$ have to be known. In practical computations this can be obtained by applying another, two-level, scheme for the first time step. The method applied for the determination of the amplification matrix, consists of replacing the multi-level scheme by a two-step

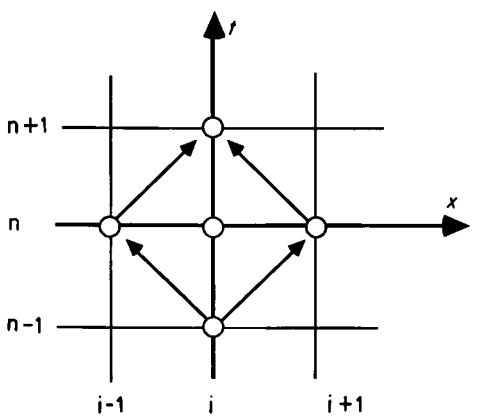

Figure 8.3.5 Computational molecule for the leapfrog scheme

system through the introduction of a new variable Z:

$$Z^n = U^{n-1} \tag{8.3.34}$$

Equation (7.3.32) then becomes

$$\begin{aligned} U^{n+1} &= -b_1 Z^n + (C - b_0)U^n + \tilde{Q} \\ Z^{n+1} &= U^n \end{aligned} \tag{8.3.35}$$

and by defining a new variable

$$W = \begin{vmatrix} U \\ Z \end{vmatrix} \tag{8.3.36}$$

the system is rewritten as

$$W^{n+1} = \bar{C}W^n + \bar{Q} \tag{8.3.37}$$

and analysed as in the previous cases.

Alternatively, the method of introducing an additional variable is fully equivalent to a more direct approach, whereby we write for the amplitudes v^n of a single harmonic

$$v^{n-1} = G^{-1} \cdot v^n \tag{8.3.38}$$

and

$$v^{n+1} = G \cdot v^n \tag{8.3.39}$$

When this is introduced into the three-level scheme a quadratic equation for G is obtained.

Example 8.3.4 The leapfrog scheme for the convection equation

Scheme (8.3.33) will be written with the new variable Z as follows:

$$\begin{vmatrix} u_i^{n+1} = z_i^n - \sigma(u_{i+1}^n - u_{i-1}^n) \\ z_i^{n+1} = u_i^n \end{vmatrix} \tag{E8.3.17}$$

and as a function of the vector W we obtain the system

$$w_i^{n+1} = \bar{C}w_i^n \tag{E8.3.18}$$

With the introduction of the shift operator E the operator $\bar{C}$ becomes

$$\bar{C} = \begin{vmatrix} -\sigma(E - E^{-1}) & 1 \\ 1 & 0 \end{vmatrix} \tag{E8.3.19}$$

The amplification matrix becomes

$$G = \begin{vmatrix} -\sigma(e^{I\phi} - e^{-I\phi}) & 1 \\ 1 & 0 \end{vmatrix} \tag{E8.3.20}$$

The eigenvalues of G are readily obtained as

$$\lambda_{\pm} = -I\sigma \sin \phi \pm \sqrt{(1 - \sigma^2 \sin^2 \phi)} \tag{E8.3.21}$$

and are to be considered as the *amplification factors of the three-level scheme*. Indeed, applying the second approach (8.3.38) and (8.3.39) to equation (8.3.33), for a harmonic $\phi = k\,\Delta x$, leads to

$$(G - 1/G) = -\sigma(e^{I\phi} - e^{-I\phi}) \tag{E8.3.22}$$

with the two solutions

$$G = -I\sigma \sin\phi \pm \sqrt{(1 - \sigma^2 \sin^2\phi)} \tag{E8.3.23}$$

If $\sigma > 1$ the scheme is unstable, since the term under the square root can become negative, and for these values G is purely imaginary and in magnitude larger than one. This is best seen for the particular value $\phi = \pi/2$.

For $|\sigma| \leqslant 1$ the scheme is *neutrally stable*, since

$$|G| = 1 \qquad \text{for} \qquad |\sigma| \leqslant 1 \tag{E8.3.24}$$

The phase error is given by

$$\varepsilon_\phi = \frac{\pm \tan^{-1}[\sigma \sin\phi / \sqrt{(1 - \sigma^2 \sin^2\phi)}]}{\sigma\phi} = \pm \frac{\sin^{-1}(\sigma \sin\phi)}{\sigma\phi} \tag{E8.3.25}$$

Hence the leapfrog scheme should give accurate results when the function u has a smooth variation, since the amplitudes are correctly modelled, so much that for low frequencies the phase error is close to one since $\varepsilon_\phi = \pm 1$ for $\phi \to 0$. However, high-frequency errors tend to remain stationary since $\varepsilon_\phi \to 0$ for $\phi \to \pi$ and, since they are undamped, they can accumulate and destroy the accuracy of the numerical solution. This is clearly seen in Figure 8.3.6.

Example 8.3.5 Du Fort and Frankel scheme for the heat-conduction equation

The scheme of Du Fort and Frankel (1953) is obtained from the unstable 'leapfrog' explicit scheme applied to the diffusion equation (8.2.7) (see Problem 8.9):

$$u_i^{n+1} - u_i^{n-1} = 2\left(\frac{\alpha\,\Delta t}{\Delta x^2}\right)(u_{i+1}^n - 2u_i^n + u_{i-1}^n) \tag{E8.3.26}$$

by averaging out the term u_i^n in time as $(u_i^{n+1} + u_i^{n-1})/2$. This leads to the scheme $\beta = \alpha\,\Delta t/\Delta x^2$:

$$u_i^{n+1} - u_i^{n-1} = 2\,\frac{\alpha\,\Delta t}{\Delta x^2}\,(u_{i+1}^n - u_i^{n+1} - u_i^{n-1} + u_{i-1}^n) \tag{E8.3.27a}$$

or

$$u_i^{n+1}(1 + 2\beta) = u_i^{n-1}(1 - 2\beta) + 2\beta(u_{i+1}^n + u_{i-1}^n) \tag{E8.3.27b}$$

The amplification matrix is obtained from the system, with $Z^n = U^{n-1}$,

$$w_i^{n+1} = \bar{C} \cdot w_i^n \tag{E8.3.28}$$

where

$$\bar{C} = \begin{vmatrix} \dfrac{2\beta(E+E^{-1})}{1+2\beta} & \dfrac{1-2\beta}{1+2\beta} \\ 1 & 0 \end{vmatrix} \tag{E8.3.29}$$

Hence

$$G = \begin{vmatrix} \dfrac{4\beta\cos\phi}{1+2\beta} & \dfrac{1-2\beta}{1+2\beta} \\ 1 & 0 \end{vmatrix} \tag{E8.3.30}$$

and the two eigenvalues, representing the amplification factors of the scheme, are given by

$$\lambda_{\pm} = \frac{2\beta\cos\phi \pm \sqrt{(1-4\beta^2\sin^2\phi)}}{1+2\beta} \tag{E8.3.31}$$

A plot of the eigenvalues $\lambda_{\pm}$ for different values of β as a function of ϕ, or a direct calculation of the condition $|\lambda_{\pm}| < 1$, shows that the scheme of Du Fort and Frankel is unconditionally stable for $\beta > 0$. This is very unusual for an explicit scheme. However, as will be seen in Chapter 10, this scheme is not always consistent.

Note that, for three-level schemes, there are two amplification factors, although the exact solution has a single value of the amplification. For the leapfrog scheme applied to the wave equation it can be observed that one of the two solutions has a negative phase error, that is, it propagates in the wrong direction. Hence the solution with the + sign corresponds to the physical solution, while the other is a *spurious* solution generated by the scheme. More insight into this aspect will appear from the stability analysis of Chapter 10 dealing with the matrix method.

8.3.4 A comparison of different schemes for the linear convection equation

It is instructive to compare the results obtained with the four schemes described in Examples 8.3.1–8.3.4 when applied to the linear convection equation. The effects of the diffusion and dispersion errors can be demonstrated, as a function of frequency, with the following two test cases, a propagating discontinuity and a sinusoidal wave packet.

The former is typical of a signal with a high-frequency content, since the Fourier decomposition of a discontinuity contains essentially high-order harmonics. On the other hand, the sinusoidal wave packet can be chosen to correspond to a selected value of the wavenumber and hence to a fixed value of the phase angle ϕ for a given mesh size Δx.

Figure 8.3.6 compares the computed results for the propogating discontinuity at a Courant number of 0.8 after 50 time steps on a mesh size

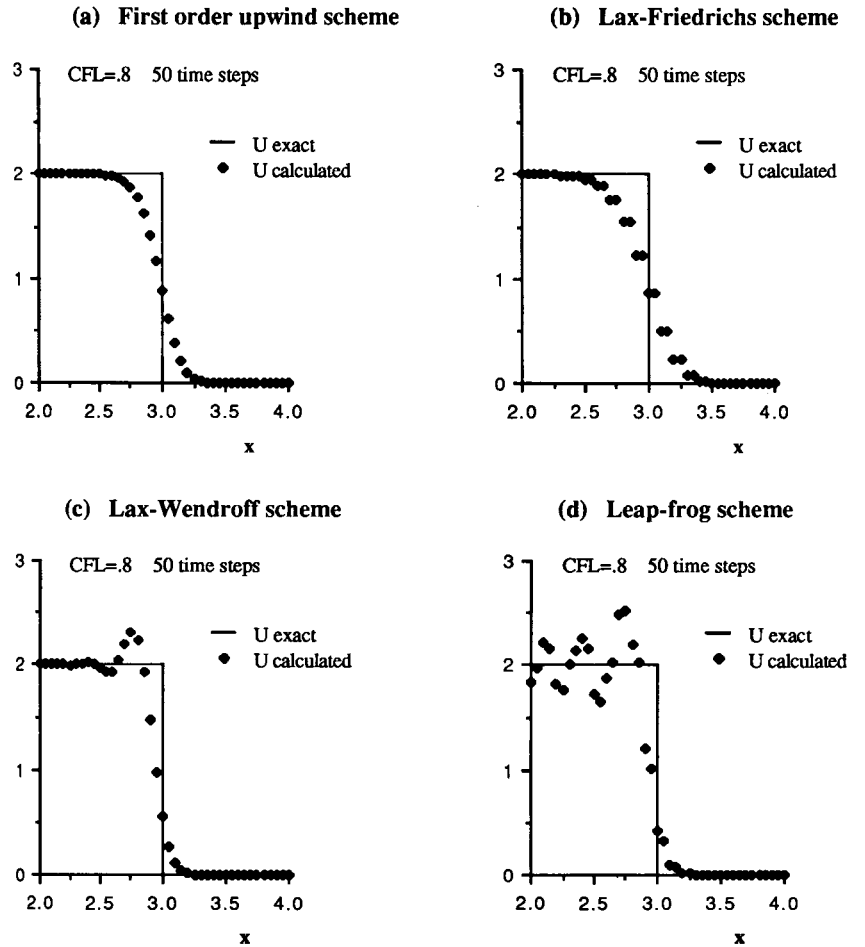

Figure 8.3.6 Comparison of four schemes on the linear convection equation for a propagating discontinuity

$\Delta x = 0.05$. The strong dissipation of the first-order upwind and Lax–Friedrichs schemes is clearly seen from the way the discontinuity is smoothed out. Observe also the 'double' solution obtained with the Lax–Friedrichs scheme, illustrating the odd–even decoupling discussed in Section 4.4 (Figure 4.4.4). Looking at Figure 8.2.1 it can be seen that u_i^{n+1} does not depend on u_i^n but on the neighbouring points u_{i-1}^n and u_{i+1}^n. These points also influence the solutions u_{i+2}^{n+1}, u_{i+4}^{n+1}, ..., while u_i^n will influence independently the points $u_{i+1}^{n+1}, u_{i+3}^{n+1}, \ldots$ The solutions obtained at the even- and odd-numbered points can therefore differ by a small constant without preventing convergence and such a difference appears on the solution shown in Figure 8.3.6(b).

The second-order Lax–Wendroff and leapfrog schemes generate oscillations due to the dominating high-frequency dispersion errors, which are mostly

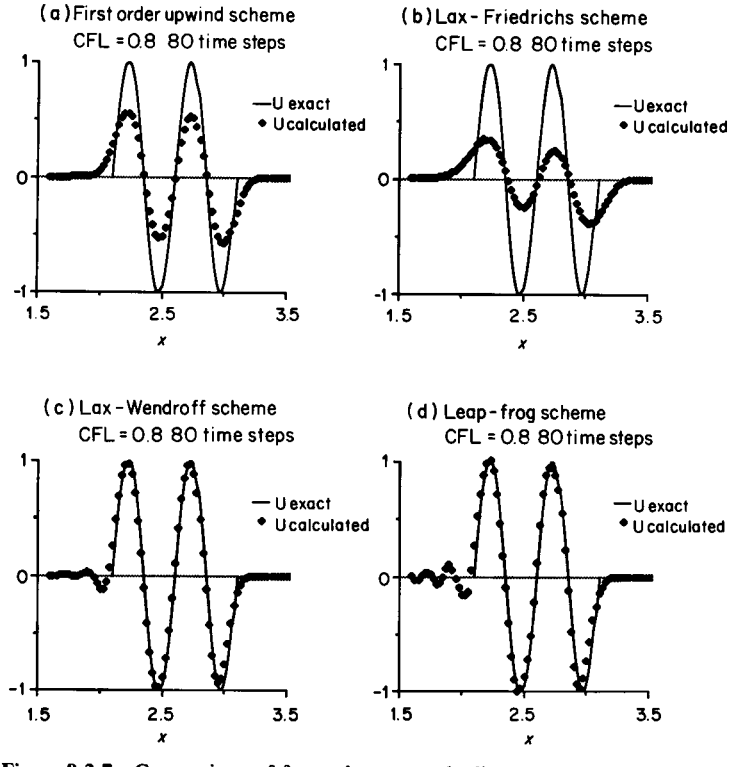

Figure 8.3.7 Comparison of four schemes on the linear convection equation for a propagating wave packet for $\phi = \pi/10$

lagging. The leapfrog scheme, which has no damping, generates stronger high-frequency oscillations compared with the Lax–Wendroff scheme, whose amplification factor is lower than one at the phase angle $\phi = \pi$, where $G(\pi) = 1 - 2\sigma^2$.

The test cases of the moving wave packet allow us to experiment with the frequency dependence of the schemes at the low end of the spectrum. Figure 8.3.7 compares the four schemes for a phase angle ϕ equal to $\pi/10$ at a Courant number of 0.8 after 80 time steps on a mesh $\Delta x = 0.025$. The strong diffusion error of the first-order schemes is clearly seen, showing that they are useless for time-dependent propagation problems of this kind. The second-order schemes give accurate results at these low frequencies, the oscillations at the beginning of the wave packet being created by the high-frequency errors generated by the slope discontinuity of the solution at this point. Hence a behaviour similar to the propagating discontinuity of the previous figure appears.

The same computations performed at a higher frequency corresponding to a phase angle of $\phi = \pi/5$, are shown in Figure 8.3.8. The first-order schemes are

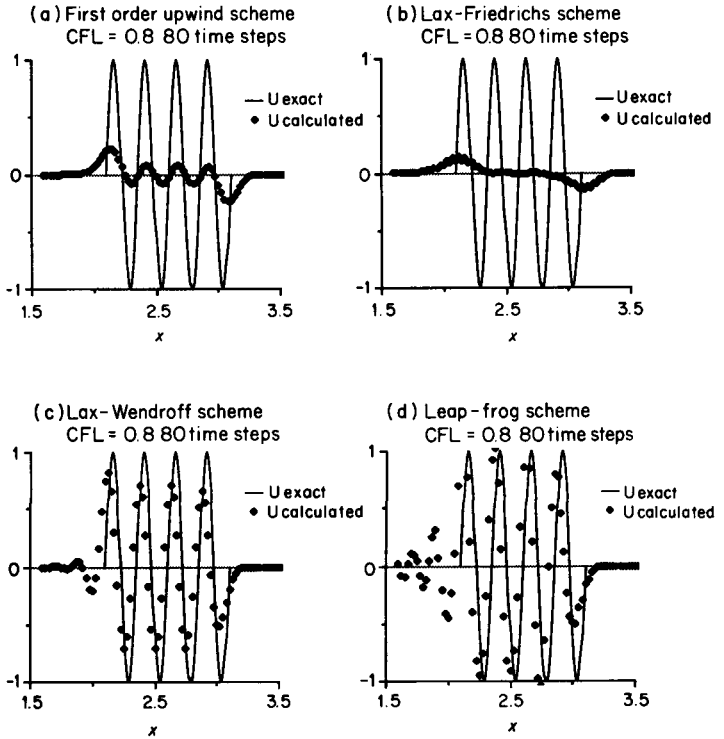

Figure 8.3.8 Comparison of four schemes on the linear convection equation for a propagating wave packet for $\phi = \pi/5$

more severely damped while the increasing, lagging dispersion errors of the two- second-order schemes can be seen by the phase shift of the computed solutions. The Lax–Wendroff scheme has a diffusion error which increases with frequency, as can be seen in Figure 8.3.4, and an amplitude error develops. The leapfrog scheme has a better behaviour with regard to the amplitude of the wave, as can be seen from the amplitudes of the second and third periods, although the first period of the wave is spoiled by the high-frequency oscillations generated at the initial slope discontinuity.

8.3.5 The numerical group velocity

The group velocity of a wave packet, containing more than one frequency, has been defined in Chapter 3 (equation (3.4.35)) and is also the velocity at which the energy of the wave is travelling. For a one-dimensional wave we have

$$\tilde{v}_G(k) = \frac{d\tilde{\omega}}{dk} \tag{8.3.40}$$

defining the group velocity as the derivative of the time frequency with respect to the wavenumber k. For a linear wave it is seen from equation (8.3.27) that the group velocity is equal to the phase speed a.

By writing the amplification factor as equation (8.3.9) the numerical dispersion relation $\omega = \omega(k) = \xi + I\eta$ can be defined, and the numerical group velocity

$$v_G(k) = \frac{d\xi}{dk} = \mathrm{Re}\left(\frac{d\omega}{dk}\right) \tag{8.3.41}$$

will represent the travelling speed of wave packets centred around the wavenumber k. Since the errors generated by a numerical scheme generally contain a variety of frequencies it is more likely that they will travel at the numerical group velocity instead of the numerical phase speed $\bar{a}$, defined by equation (8.3.30).

For the leapfrog scheme (equation (8.3.33)) the introduction of equation (8.3.6) into equation (E8.3.23) leads to the numerical dispersion relation:

$$\sin \omega\Delta t = \sigma \sin \phi \tag{8.3.42}$$

from which we derive

$$v_G = a \frac{\cos \phi}{\cos \omega\Delta t} = a \frac{\cos \phi}{(1 - \sigma^2 \sin^2 \phi)^{1/2}} \tag{8.3.43}$$

For low frequencies the group velocity is close to the phase speed a, but for the high frequencies ($\phi \approx \pi$) the group velocity is close to $-a$, indicating that the high wavenumber packets will travel in the opposite direction to the wave phase speed a. This can be observed in Figure 8.5.2, where it is seen that the high-frequency errors, generated upstream of the stationary shock, travel in the upstream direction.

An instructive example is provided by the exponential wave packet

$$u(x, t = 0) = \exp(-\alpha x^2) \sin 2\pi k_w x \tag{8.3.44}$$

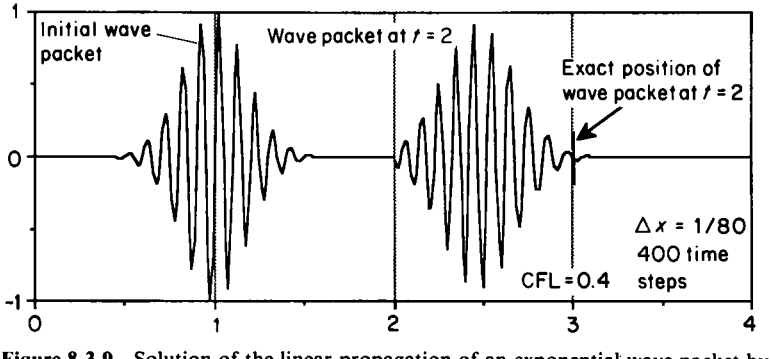

Figure 8.3.9 Solution of the linear propagation of an exponential wave packet by the leapfrog scheme, after 400 time steps, for $\phi = \pi/4$

shown in Figure 8.3.9 for a phase angle $\phi = k\Delta x = \pi/4$, corresponding to a wavelength of $\lambda = 8\Delta x$. The solution of the linear wave equation $u_t + au_x = 0$ with the leapfrog scheme is shown in the same figure after 400 time steps for a Courant number of 0.4 and $\Delta x = 1/80$ for $a = 1$. If the initial solution is centred at $x = 1$ the exact solution should be located around $x = 3$ at time $t = 400\Delta t = 2$. However, the numerical solution is seen to have travelled only to the point $x \approx 2.475$, which indicates a propagating speed of 0.7375 instead of the phase speed $a = 1$. This corresponds exactly to the computed group velocity from equation (8.3.43), which gives a value of $v_G = 0.7372$ at $\phi = \pi/4$.

These properties of the group velocity should be kept in mind when analysing numerical data associated with high-frequency solutions. More details on the applications of the concept of group velocity to the analysis of numerical schemes can be found in Vichnevetsky and Bowles (1982), Trefethen (1982) and Cathers and O'Connor (1985). The last reference presents a detailed, comparative analysis of the group velocity properties of various finite element and finite difference schemes applied to the one-dimensional, linear convection equation. Trefethen (1983, 1984) has derived some important relations between group velocity and the stability of numerical boundary conditions of hyperbolic problems. His results can be expressed by the condition that the numerical boundary treatment should not allow group velocities at these boundaries to transport energy into the computational domain. We refer the reader to the original references for more details and derivations.

8.4 MULTI-DIMENSIONAL VON NEUMANN ANALYSIS

For problems in more than one space dimension the Fourier decomposition at the basis of the Von Neumann stability analysis can be performed separately in each space direction through the introduction of a wavenumber vector $\vec{\varkappa}$. For instance, the solution $u(\vec{x}, t)$ will be represented as a superposition of harmonics of the form

$$u(\vec{x}, t) \sim \hat{v}\, e^{-I\omega t}\, e^{I\vec{\varkappa}\cdot\vec{x}} \tag{8.4.1}$$

where the scalar product $\vec{\varkappa} \cdot \vec{x}$ is defined as

$$\vec{\varkappa} \cdot \vec{x} = \varkappa_x x + \varkappa_y y + \varkappa_z z \tag{8.4.2}$$

In discretized form, with mesh point indexes i, j, k, we have

$$(\vec{\varkappa} \cdot \vec{x})_{i,j,k} = i(\varkappa_x \Delta x) + j(\varkappa_y \Delta y) + k(\varkappa_z \Delta z) \equiv i \cdot \phi_x + j \cdot \phi_y + k \cdot \phi_z \tag{8.4.3}$$

and the three parameters ϕ_x, ϕ_y, ϕ_z range from $-\pi$ to π is each of the three space directions. The further determination of the amplification matrix remains unchanged from the one-dimensional case.

8.4.1 Parabolic equations

Let us consider as an example the two-dimensional heat diffusion equation (7.1.15). The obvious generalization of the one-dimensional explicit central scheme (8.3.16) for the parabolic equation, written as

$$\frac{\partial u}{\partial t} = \alpha\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{8.4.4}$$

is

$$u_{ij}^{n+1} - u_{ij}^{n} = \alpha\,\Delta t\left[\frac{u_{i+1,j}^{n} - 2u_{ij}^{n} + u_{i-1,j}^{n}}{\Delta x^2} + \frac{u_{i,j+1}^{n} - 2u_{ij}^{n} + u_{i,j-1}^{n}}{\Delta y^2}\right] \tag{8.4.5}$$

A discrete Fourier decomposition is defined by

$$u_{ij}^{n} = \sum_{\varkappa_x, \varkappa_y} v^n\, \mathrm{e}^{I\varkappa_x i\,\Delta x}\, \mathrm{e}^{I\varkappa_y j\,\Delta y} \tag{8.4.6}$$

where the range of $\varkappa_x$ and $\varkappa_y$ is defined separately for each direction, as in the one-dimensional case. Inserting a single component into the discretized scheme, the amplification matrix G is still defined, as in the one-dimensional case, as

$$v^{n+1} = Gv^n \tag{8.4.7}$$

We obtain, from equation (8.4.5), after division by $v^n\, \mathrm{e}^{iI\phi_x}\, \mathrm{e}^{jI\phi_y}$,

$$\begin{aligned} G - 1 &= \beta[(\mathrm{e}^{I\phi_x} + \mathrm{e}^{-I\phi_x} - 2) + \left(\frac{\Delta x}{\Delta y}\right)^2 (\mathrm{e}^{I\phi_y} + \mathrm{e}^{-I\phi_y} - 2)] \\ G - 1 &= -4\beta\left(\sin^2\phi_x/2 + \left(\frac{\Delta x}{\Delta y}\right)^2 \sin^2\phi_y/2\right) \end{aligned} \tag{8.4.8}$$

The strict stability condition becomes

$$\left| 1 - 4\beta\left(\sin^2\phi_x/2 + \left(\frac{\Delta x}{\Delta y}\right)^2 \sin^2\phi_y/2\right) \right| \leqslant 1 \tag{8.4.9}$$

which leads to

$$\alpha > 0 \tag{8.4.10}$$

and

$$\beta\left(1 + \left(\frac{\Delta x}{\Delta y}\right)^2\right) \leqslant \frac{1}{2}$$

or

$$\alpha\left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\right) \Delta t \leqslant \frac{1}{2} \tag{8.4.11}$$

This stability condition is necessary and sufficient and is analogous to condition (8.3.20) but puts a more severe requirement on the time step. For

instance, if $\Delta x = \Delta y$ the time step is reduced by a factor of two, compared with the one-dimensional case:

$$\Delta t \leqslant \frac{\Delta x^2}{4\alpha} \tag{8.4.12}$$

8.4.2 The two-dimensional convection equation

Consider the system of p equations

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} + B\frac{\partial U}{\partial y} = 0 \tag{8.4.13}$$

where A and B are $(p \times p)$ constant matrices, with the property $AB = BA$. Applying a Lax–Friedrichs scheme to this system leads to

$$U_{ij}^{n+1} = \frac{1}{4}(U_{i,j+1}^n + U_{i+1,j}^n + U_{i-1,j}^n + U_{i,j-1}^n) - \frac{\Delta t}{2\Delta x} A(U_{i+1,j}^n - U_{i-1,j}^n)$$

$$-\frac{\Delta t}{2\Delta y} B(U_{i,j+1}^n - U_{i,j-1}^n) \tag{8.4.14}$$

With the decomposition (8.4.6) for a single harmonic the amplification matrix becomes

$$G = \frac{1}{2}(\cos\phi_x + \cos\phi_y) - \frac{\Delta t}{\Delta x} A\, I \sin\phi_x - \frac{\Delta t}{\Delta y} B\, I \sin\phi_y \tag{8.4.15}$$

The spectral radius ρ can be obtained from equation (8.2.22) and the fact that G is a normal matrix. Hence with

$$\| G \|_{L_2}^2 = \rho(G^*G) = \rho^2(G)$$

$$\rho(G^*G) = \tfrac{1}{4}(\cos\phi_x + \cos\phi_y)^2 + (\sigma_x \sin\phi_x + \sigma_y \sin\phi_y)^2 \tag{8.4.16}$$

where

$$\sigma_x = \frac{\Delta t}{\Delta x}\rho(A) \qquad \sigma_y = \frac{\Delta t}{\Delta y}\rho(B) \tag{8.4.17}$$

A necessary condition is obtained by looking at the most unfavourable situation, namely ϕ_x and ϕ_y independent but small. Expanding the sine and cosine functions up to higher order.

$$\| G \|_{L_2}^2 = 1 - [(\tfrac{1}{2} - \sigma_x^2)\phi_x^2 + (\tfrac{1}{2} - \sigma_y^2)\phi_y^2 - 2\sigma_x\sigma_y\phi_x\phi_y] + 0(\phi_x^4, \phi_y^4) \tag{8.4.18}$$

The quadratic form in ϕ_x, ϕ_y between parentheses has to be positive for stability. Thus if the discriminant is negative the quadratic form never goes through zero avoiding a change of sign. This will occur if

$$(\sigma_x^2 + \sigma_y^2) \leqslant \tfrac{1}{2} \tag{8.4.19}$$

representing the interior of a circle in the (σ_x, σ_y) plane of radius $\sqrt{2}/2$, centred at the origin. This condition is also shown to be sufficient in Section 8.6. Here again, this condition is far more severe than the corresponding one-dimensional case.

As can be seen from these examples, it is much more difficult to obtain the stability limits for multi-dimensional problems, even for linear equations, and several non-sufficient stability conditions can be found in the literature. Actually, even for one-dimensional problems, controversial results from Von Neumann analysis have appeared in the literature (see Chapter 10 for a discussion of a famous example concerning the convection–diffusion equation).

8.5 STABILITY CONDITIONS FOR NON-LINEAR PROBLEMS

Most of the mathematical models describing the various approximations to a flow problem contain non-linear terms, or eventually non-constant coefficients. In these cases the Von Neumann method for stability analysis based on the Fourier expansion cannot strictly be applied since we can no longer isolate single harmonics. Nevertheless, if we introduced a complete Fourier series into the discretized scheme with non-constant coefficients the amplification matrix would become a function of all wavenumbers, instead of a linear superposition of amplification matrices for single harmonics. In addition, for non-linear problems the amplification matrix would also become a function of the amplitude of the solutions and not only of their frequency as in the constant-coefficient, linear case. Hence these contributions could generate instabilities, even with schemes which are basically linearly stable.

8.5.1 Non-constant coefficients

Consider a linear problem with non-constant coefficients, for instance, the one-dimensional, parabolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right) \tag{8.5.1}$$

or the hyperbolic problem

$$\frac{\partial u}{\partial t} + a(x)\frac{\partial u}{\partial x} = 0 \tag{8.5.2}$$

A two-step numerical scheme applied to these equations will be written as

$$u_i^{n+1} = C(x, E)u_i^n \tag{8.5.3}$$

For instance, for an explicit, central scheme the parabolic equation (8.5.1) becomes

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x^2}\left[\alpha_{i+1/2}(u_{i+1}^n - u_i^n) - \alpha_{i-1/2}(u_i^n - u_{i-1}^n)\right] \tag{8.5.4}$$

where

$$\alpha_{i+1/2} = \alpha(x_{i+1/2}) \tag{8.5.5}$$

Hence

$$C(x, E) = 1 + \frac{\Delta t}{\Delta x^2} [\alpha(x_{i+1/2})(E - 1) - \alpha(x_{i-1/2})(1 - E^{-1})] \tag{8.5.6}$$

The hyperbolic equation (8.5.2) with an explicit, upwind scheme for $a > 0$ will be written as

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} a(x_{i-1/2})(u_i^n - u_{i-1}^n) \tag{8.5.7}$$

or

$$C(x, E) = 1 - \frac{\Delta t}{\Delta x} a(x_{i+1/2})(1 - E^{-1}) \tag{8.5.8}$$

The amplification matrix is now a function of x and not only of the wavenumber k. Indeed, introducing a single harmonic $(u_i^n)_k = v^n \, e^{Iik\,\Delta x}$, a *local* amplification matrix can be defined by

$$G(x, k) = C(x, e^{I\phi}) \tag{8.5.9}$$

where the variable coefficients are formally retained as functions of x.

In the two examples above we have

$$G(x, \phi) = 1 + \frac{\Delta t}{\Delta x^2} [\alpha(x)(e^{I\phi} - 1) - \alpha(x)(1 - e^{-I\phi})] \tag{8.5.10}$$

and for the hyperbolic example

$$G(x, \phi) = 1 - \frac{\Delta t}{\Delta x} a(x)(1 - e^{-I\phi}) \tag{8.5.11}$$

Under general conditions (see Richtmyer and Morton, 1967) it can be proved that for linear, non-constant coefficient problems a *local* Von Neumann analysis will provide a necessary condition for stability. That is, freezing the coefficients at their value at a certain point and applying the Von Neumann method provides a local stability condition. However, in order to obtain also sufficient conditions for stability, additional restrictions on the amplification matrix G have to be introduced. These conditions are connected with the generation of high-frequency harmonics due to the non-linear behaviour and to the necessity of damping these frequencies in order to maintain stability. This is particularly urgent for non-linear hyperbolic problems, since they describe essentially propagating waves without physical damping. Even with parabolic problems, where such a physical damping exists an additional condition on the amplification matrix is required. This is provided by the concept of *dissipative schemes*.

8.5.2 Dissipative schemes (Kreiss, 1964)

A scheme is called dissipative (in the sense of Kreiss) of order $2r$, where r is a positive integer, if there exists a constant $\delta > 0$ such that for wavenumbers $\vec{\kappa}$ with $\phi_j = (\kappa_j \Delta x_j) \leqslant \pi$ for each space component j ($j = 1, 2, 3$ in a three-dimensional space) the eigenvalues λ of the amplification matrix satisfy the condition

$$| \lambda(\vec{\kappa}, \Delta t, \vec{x}) | \leqslant 1 - \delta | \vec{\kappa} \cdot \Delta \vec{x} |^{2r} \tag{8.5.12}$$

for all $\vec{\kappa}$ and for $0 < \Delta t < \tau$. This condition ensures that for $\phi = \pi$, that is, for the high frequencies associated with the $(2\Delta x_j)$ waves (the shortest waves to be resolved on the mesh), enough dissipation is provided by the discretization to avoid their negative impact on the stability.

For parabolic problems we can show, under fairly general conditions (Richtmyer and Morton, 1967), that if a $\delta > 0$ exists such that

$$| G(x, \phi) | \leqslant 1 - \delta\phi^2 \qquad \text{for } -\pi \leqslant \phi \leqslant \pi \tag{8.5.13}$$

the corresponding schemes are stable. In particular, a scheme with an amplification matrix, such that the spectral radius $\rho(G) = 1$ for $\phi = \pi$, is not dissipative in the sense of Kreiss.

For hyperbolic problems we have the following theorem of Kreiss (1964): If the matrix A is Hermitian, uniformly bounded and Lipshitz continuous in x, then if the scheme is dissipative of order $2r$ and accurate of order $(2r - 1)$, it is stable.

This is a sufficient condition for stability, but many schemes applied in practice do not satisfy this condition.

Lax–Friedrichs scheme

From the amplification factor of the Lax scheme (equation (E8.3.2))

$$G(\phi) = \cos \phi - I\sigma \sin \phi$$

we deduce that $G(\pi) = 1$ for all σ. Hence the Lax scheme is not dissipative in Kreiss's sense. However, since this scheme damps strongly all frequencies, as seen earlier, it remains generally stable even for non-linear problems such as the Euler equations (Di Perna, 1983).

Upwind scheme

According to equation (E8.3.6) the modulus of the amplification factor becomes, for small values of ϕ,

$$| G | \approx 1 - \sigma(1 - \sigma)\phi^2 + \cdots \tag{8.5.14}$$

and is dissipative of order 2 for $0 < \sigma < 1$. Since

$$| G(\pi) | = | 1 - 2\sigma | \tag{8.5.15}$$

the upwind scheme is dissipative in the sense of Kreiss. The order of the scheme being one, the conditions of Kreiss's theorem are satisfied and the upwind scheme will be stable for functions $a(x)$ such that

$$0 < \frac{a(x)\Delta t}{\Delta x} < 1$$

for all values of x in the computational domain.

Lax–Wendroff scheme

The dissipation of the scheme is of fourth order, since for small ϕ, from equation (E8.3.13)

$$|G| \approx 1 - \frac{\sigma^2}{8}(1-\sigma^2)\phi^4 + 0(\phi^6) \tag{8.5.16}$$

showing that the Lax–Wendroff scheme is dissipative to the fourth order. Since $G(\pi) = 1 - 2\sigma^2$ the Lax–Wendroff scheme is dissipative in the sense of Kreiss for non-zero values of σ.

8.5.3 Non-linear problems

Very little information is available on the stability of general non-linear discretized schemes. Within the framework of the Von Neumann method it can be said that the stability of the linearized equations, with frozen coefficients, is necessary for the stability of the non-linear form but that it is certainly not sufficient. Products of the form $u(\partial u/\partial x)$ will generate high-frequency waves which, through a combination of the Fourier modes on a finite mesh, will reappear as low-frequency waves and could deteriorate the solutions. Indeed, a discretization of the form

$$\left(u\frac{\partial u}{\partial x}\right)_i = u_i\left(\frac{u_{i+1} - u_{i-1}}{2\Delta x}\right) \tag{8.5.17}$$

becomes, when the Fourier expansion (8.2.1) is introduced,

$$\begin{aligned}\left(u\frac{\partial u}{\partial x}\right)_i &= \sum_{k_1}\left(\sum_{k_2} v(k_2)e^{Ik_2 i\Delta x}\right)v(k_1)\frac{e^{Ik_1 i\Delta x}}{2\Delta x}\left(e^{Ik_1\Delta x} - e^{-Ik_1\Delta x}\right) \\ &= \frac{I}{\Delta x}\sum_{k_1}\sum_{k_2} v(k_1)v(k_2)\, e^{I(k_1+k_2)\Delta x}\sin k_1\Delta x\end{aligned} \tag{8.5.18}$$

The sum $(k_1 + k_2)\Delta x$ can become larger than the maximum value π associated with the $(2\Delta x)$ wavelength. In this case the corresponding harmonic will behave as a frequency $[2\pi - (k_1 + k_2)\Delta x]$ and will therefore appear as a low-frequency contribution. This non-linear phenomenon is called *aliasing*, and is to be avoided by providing enough disipation in the scheme to damp the high frequencies.

For non-linear problems we also observe that the coefficient of a single harmonic k_1 is a function of the amplitude of the signal through the factor $v(k_2)$ in the above development of the non-linear term uu_x. Hence for small amplitudes the non-linear version of a linearly stable scheme could remain stable, while an unstable behaviour could appear for larger amplitudes of the solution. In this case the scheme could be generally stabilized by adding additional dissipation to the scheme without affecting the order of accuracy.

A typical example is the leapfrog scheme, which is neutrally stable, $|G| = 1$ for all $|\sigma| < 1$. Hence this scheme is not dissipative in the sense of Kreiss, and when applied to the inviscid Burger's equation $u_t + uu_x = 0$, the computations become unstable in certain circumstances, as can be seen from Figure 8.5.1. This figure shows the computed solutions of Burger's equation for a stationary shock, after 10, 20 and 30 time steps at a Courant number of 0.8 and a mesh size of $\Delta x = 0.05$. The open squares indicate the exact solution. The amplitude of the errors increases continuously, and the solution is completely destroyed after 50 time steps. The instability is entirely due to the non-linearity of the equation, since the same scheme applied to the linear convection equation does not diverge, although strong oscillations are generated, as shown in Figure 8.3.6(d).

In the present case the high-frequency errors are generated by the fact that the shock is located on a mesh point. This point has zero velocity and, with an initial solution passing through this point, a computed shock structure is enforced with this internal point fixed, creating high-frequency errors at the two adjacent points. This is clearly seen in Figure 8.5.1, looking at the evolution of the computed solutions, and also by comparing it with Figure 8.5.2, which displays the results of an identical computation for a stationary shock located between two mesh points. This computation does not become unstable, since the shock structure is not constrained by an internal point. Observe also the propagation of the generated high-frequency errors away

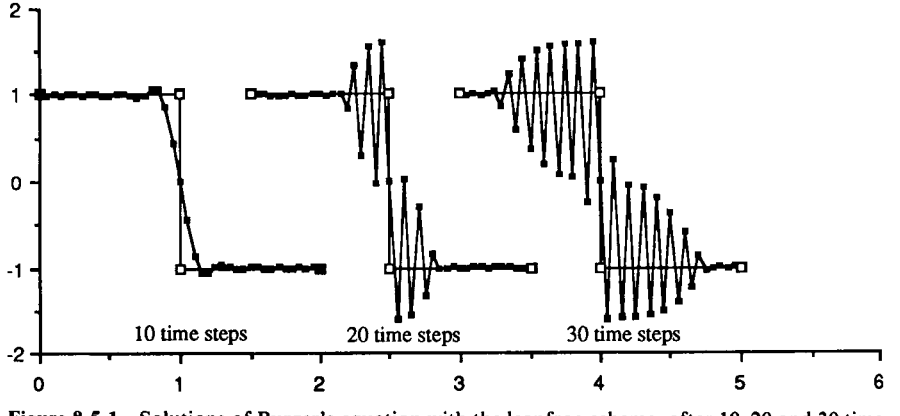

Figure 8.5.1 Solutions of Burger's equation with the leapfrog scheme, after 10, 20 and 30 time steps, for a stationary shock located on a mesh point

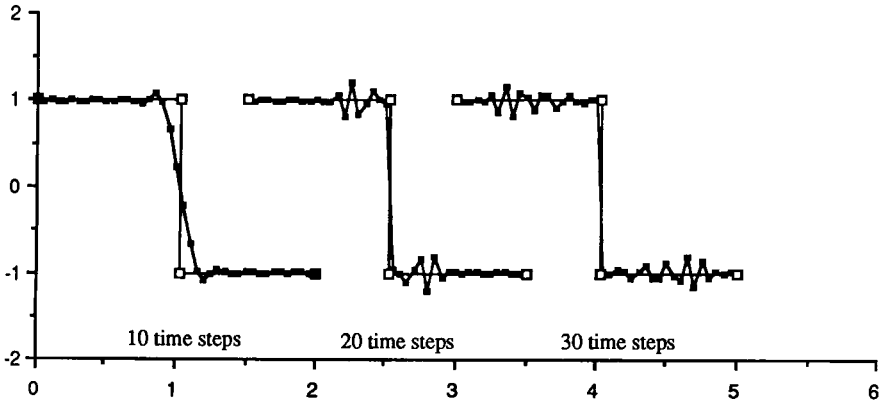

Figure 8.5.2 Solutions of Burger's equation with the leapfrog scheme, after 10, 20 and 30 time steps, for a stationary shock located between mesh points

from the shock position. They propagate at a velocity equal to the numerical group velocity of the scheme, associated with the errors with the shortest $2\Delta x$ wavelength.

Other sources of non-linear instability have been identified for the leapfrog scheme applied to Burger's equation and are described as a 'focusing' mechansim by Briggs *et al.* (1983). The structure of this mechanism has been further investigated by Sloan and Mitchell (1986).

This mechanism is not the classical, finite amplitude instability generated by terms of the form of equation (8.5.18). This instability can be analysed by considering group of modes which are closed under aliasing, that is, modes $k_1, k_2, k_3, \ldots$, such that

$$2\pi(k_1 + k_2)\Delta x = k_3\ \Delta x \tag{8.5.19}$$

For instance, referring to definition (8.1.10) of the discrete wavenumber k_j it is seen that the modes $k_1\ \Delta x = 2\pi/3$, $k_2\ \Delta x = \pi$ and $k_3\ \Delta x = \pi/3$ satisfy equation (8.5.19) for all permutations of the three modes.

By investigating solutions which contain a finite number of closed modes, the non-linear contributions from terms of the form (8.5.18) can lead to exponentially growing amplitudes, for Courant numbers below one, when the amplitudes reach a certain critical threshold function of σ. This is the mechanism which generates the instability of Figure 8.5.1.

The 'focusing' mechanism, described by Briggs *et al.*, is of a different nature. It corresponds to an amplification and a concentration of the initial errors at isolated points in the grid. This generates sharp local peaks as a result of the non-linear interaction between the original stable modes and their immediate neighbours in wavenumber space, even for initial amplitudes below the critical threshold for finite amplitude instabilities. Once the critical amplitude is reached locally it starts growing exponentially. The particular

character of this focusing property lies in its local aspect, while other non-linear instabilities are global in that the breakdown, for a continuous solution, occurs uniformly throughout the grid.

It has to be added that this focusing process can take a long time, several thousand time steps, depending on the initial error level. Figure 8.5.3, from Briggs *et al.* (1983), illustrates this process for an initial solution composed of three modes ($\pi/3$, $2\pi/3$, and π) with amplitudes below critical such that the computed solution should remain stable. The dashed line indicates the critical level above which finite amplitude instability develops. The computed results are shown for a Courant number of 0.9 and $\Delta x = 1/300$ after 400, 1000, 2000, 2200, 2400 and 2680 time steps. Until 1000 time steps the solution still retains its periodic structure; by 2000 time steps the envelope of the initial profile begins to oscillate, and local amplitudes start to concentrate until the critical threshold is reached at a single point after 2680 time steps. From this stage onwards the classical mechanism takes over and the solution diverges rapidly.

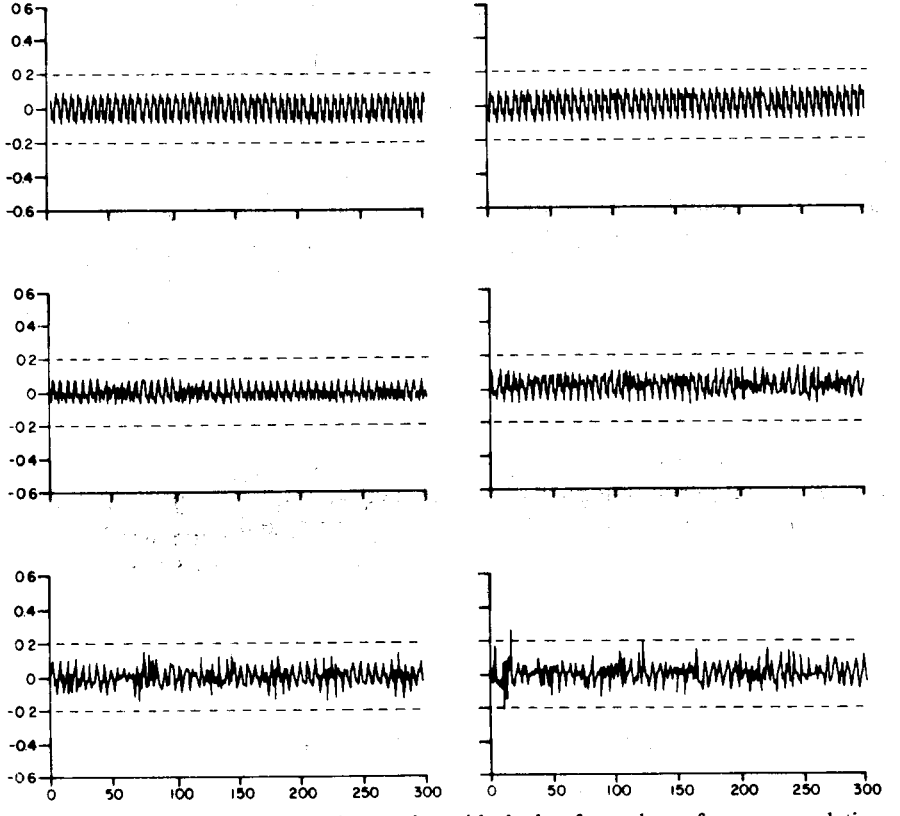

Figure 8.5.3 Solutions of Burger's equation with the leapfrog scheme for a wave solution with three modes, after 400, 1000, 2000, 2200, 2400 and 2680 time steps. (From Briggs *et al.*, 1983)

This mechanism has a strong resemblance to the chaotic behaviour of solutions of non-linear equations and their multiple bifurcations, which are also the basis of descriptions of the generation of turbulence. The reader might refer in this relation to a recent study of McDonough and Bywater (1986) on the chaotic solutions to Burger's equation.

The above examples indicate the degree of complexity involved in the analysis of the stability of non-linear equations and the need for methods which would prevent the development of instabilities for long-term computations. A frequently applied method consists of adding higher-order terms which provide additional dissipation in order to damp the non-linear instabilities without affecting the accuracy. Examples of this approach will be presented in Volume 2, when dealing with the discretization of the Euler equations.

8.6 SOME GENERAL METHODS FOR THE DETERMINATION OF VON NEUMANN STABILITY CONDITIONS

Although simple in principle and in its derivation, the Von Neumann amplification matrix is often very tedious and complicated to analyse in order to obtain the practical stability limits on the parameters of the scheme. If it is straightforward to obtain necessary conditions, it is much more difficult to derive the sufficient conditions for stability. The variety of imprecise conditions found in the literature for relatively simple problems, such as the one-dimensional convection–diffusion equation, testify to these difficulties. The situation is still worse for multi-dimensional problems. For instance, the correct, necessary and sufficient stability limits for the convection–diffusion equation in any number of dimensions had been obtained only recently (Hindmarsh *et al.*, 1984).

Due to the importance of the Von Neumann analysis, we will present a few methods which allow the derivation of precise, necessary and sufficient stability conditions for some limited, but still frequently occurring, discretizations. The first case treats the general two-level, three-point central schemes in one dimension, while the second will present the stability criteria for multi-dimensional, centrally discretized convection–diffusion equations, and, because of its importance, we will reproduce the derivation of Hindmarsh *et al.* (1984). The last case is of a more general nature, and applies to any amplification matrix obtained for an arbitrary system of discretized equation, allowing the reduction of the polynomial (8.2.19) to simpler forms.

8.6.1 One-dimensional, two-level, three-point schemes

We consider here the general scheme

$$b_3 u_{i+1}^{n+1} + b_2 u_i^{n+1} + b_1 u_{i-1}^{n+1} = a_3 u_{i+1}^n + a_2 u_i^n + a_1 u_{i-1}^n \tag{8.6.1}$$

where, for consistency, we should have (u_i = constant should be a solution)

$$b_3 + b_2 + b_1 = a_3 + a_2 + a_1 = 1 \tag{8.6.2}$$

with an arbitrary normalization at one. After elimination of b_2 and a_2, the amplification factor is

$$\begin{aligned} G &= \frac{a_3(e^{I\phi} - 1) + a_1(e^{-I\phi} - 1) + 1}{b_3(e^{I\phi} - 1) + b_1(e^{-I\phi} - 1) + 1} \\ &= \frac{1 - (a_3 + a_1)(1 - \cos\phi) + I(a_3 - a_1)\sin\phi}{1 - (b_3 + b_1)(1 - \cos\phi) + I(b_3 - b_1)\sin\phi} \end{aligned} \tag{8.6.3}$$

Hence

$$|G|^2 = |GG^*| = \frac{A_1\beta^2 + A_2\beta + 1}{B_1\beta^2 + B_2\beta + 1} \tag{8.6.4}$$

where

$$\begin{aligned} &\beta = \sin^2\phi/2 \\ &A_1 = 16a_3a_1 \qquad\qquad B_1 = 16b_3b_1 \\ &A_2 = 4[(a_3 - a_1)^2 - (a_3 + a_1)] \qquad B_2 = 4[(b_3 - b_1)^2 - (b_3 + b_1)] \end{aligned} \tag{8.6.5}$$

Note that the denominator $(B_1\beta^2 + B_2\beta + 1) \geqslant 0$ in the range $0 \leqslant \beta \leqslant 1$, since $(1 + B_1 + B_2) = (1 - 2b_2)^2$ is always non-negative. Hence the condition $|G|^2 \leqslant 1$ leads to

$$(A_1 - B_1)\beta^2 + (A_2 - B_2)\beta \leqslant 0 \tag{8.6.6}$$

and for all values of $0 \leqslant \beta \leqslant 1$ the necessary and sufficient Von Neumann stability conditions are

$$\begin{aligned} (A_2 - B_2) &\leqslant 0 \\ (A_1 - B_1 + A_2 - B_2) &\leqslant 0 \end{aligned} \tag{8.6.7}$$

Example 8.6.1 Diffusion equation

Considering scheme (8.3.16) we have $b_3 = b_1 = 0$, $a_3 = a_1 = \beta$. Hence

$$\begin{aligned} B_1 &= 0 \qquad & B_2 &= 0 \\ A_1 &= 16\beta^2 \qquad & A_2 &= -8\beta \end{aligned} \tag{E8.6.1}$$

and we obtain the necessary and sufficient conditions

$$\begin{aligned} B &> 0 \\ 8\beta(2\beta - 1) &\leqslant 0 \end{aligned} \tag{E8.6.2}$$

leading to the earlier obtained relation

$$0 < \beta \leqslant 1/2$$

8.6.2 Multi-dimensional space-centred, convection–diffusion equation

We consider here a general scalar equation in a space of dimension M of the form

$$\frac{\partial u}{\partial t} + (\vec{a} \cdot \vec{\nabla})u = \vec{\nabla} \cdot (\bar{\bar{\alpha}}\, \vec{\nabla} u) \tag{8.6.8}$$

where $\bar{\bar{\alpha}}$ is a *diagonal* diffusivity tensor and a central discretization of second-order accuracy:

$$\frac{1}{\Delta t}(u_J^{n+1} - u_J) + \sum_{n=1}^{M} a_m \frac{\bar{\delta}_m u_J}{\Delta x_m} = \sum_{m=1}^{M} \alpha_m \frac{\delta_m^2 u_J}{\Delta x_m^2} \tag{8.6.9}$$

where J represents a mesh point index (for instance, in two dimensions $J(i, j)$ and $J(i, j, k)$ in a three-dimensional Cartesian mesh). The operator $\bar{\delta}_m$ is the central difference acting on the variable x_m, that is

$$\frac{\bar{\delta}_m u_J}{\Delta x_m} = \frac{1}{2\Delta x_j}(u_{i,j+1,k} - u_{i,j-1,k}) \qquad \text{if } m = j \tag{8.6.10}$$

and the second derivative operator δ_m^2 is similarly defined:

$$\frac{\delta_m^2 u_J}{\Delta x_m^2} = \frac{1}{\Delta x_j^2}(u_{i,j+1,k} - 2u_{ijk} + u_{i,j-1,k}) \qquad \text{if } m = j \tag{8.6.11}$$

Defining

$$\begin{aligned} \sigma_m &= a_m\, \Delta t / \Delta x_m \\ \beta_m &= \alpha_m\, \Delta t / \Delta x_m^2 \end{aligned} \tag{8.6.12}$$

the above scheme becomes

$$u_J^{n+1} = u_J^n - \sum_{m=1}^{M} (\sigma_m\, \bar{\delta}_m u_J - \beta_m \delta_m^2 u_J) \tag{8.6.13}$$

This discretized equation represents the scheme to be analysed independently of the original equation (8.6.8), used as a starting point. Hence the following results can be applied to a wider range of problems; for instance, the two-dimensional convection equation, discretized with the Lax–Friedrichs scheme (8.4.14), can clearly be written in the above form. As will be seen in Volume 2, many numerical schemes for the inviscid system of Euler equations can also be written in this way.

With representation (8.4.1), and $\phi_m = \varkappa_m\, \Delta x_m$, the amplification factor becomes

$$G = 1 - I \sum_{m=1}^{M} \sigma_m \sin \phi_m - 4 \sum_{m=1}^{M} \beta_m \sin^2 \phi_m / 2 \tag{8.6.14}$$

The modulus squared is given by

$$|G|^2 = \left[1 - 4 \sum_{m=1}^{M} \beta_m \sin^2 \phi_m / 2\right]^2 + \left[\sum_{m=1}^{M} \sigma_m \sin \phi_m\right]^2 \tag{8.6.15}$$

The extreme conditions are obtained when all $\phi_m = \pi$, on the one hand, and when all ϕ_m go to zero on the other. In the first case we obtain the condition

$$\left(1 - 4 \sum_{m=1}^{M} \beta_m\right)^2 \leqslant 1 \tag{8.6.16}$$

Hence this leads to

$$0 \leqslant \sum_{m=1}^{M} \beta_m \leqslant \frac{1}{2} \tag{8.6.17}$$

In the second case, performing a Taylor expansion around $\phi_m = 0$, and neglecting higher-order terms, we obtain

$$\begin{aligned} |G|^2 &= \left[1 - \sum_{m=1}^{M} \beta_m \phi_m^2\right]^2 + \left[\sum_{m=1}^{M} \sigma_m \phi_m\right]^2 + 0(\phi_m^4) \\ &= 1 - 2 \sum_{m=1}^{M} \beta_m \phi_m^2 + \left(\sum_{m=1}^{M} \sigma_m \phi_m\right)^2 + 0(\phi_m^4) \end{aligned} \tag{8.6.18}$$

The right-hand side is a quadratic expression in the ϕ_m. Following Hindmarsh *et al.* (1984), it can be written as follows, introducing the vectors $\vec{\phi} = (\phi_1, \ldots, \phi_M)^T$, $\vec{\sigma} = (\sigma_1, \ldots, \sigma_M)^T$ and the diagonal matrix $\beta = \text{diag}(\beta_1, \ldots, \beta_M)$, neglecting higher-order terms:

$$|G|^2 = 1 - \vec{\phi}^T (2\bar{\bar{\beta}} - \vec{\sigma} \otimes \vec{\sigma}^T) \vec{\phi} \tag{8.6.19}$$

For $|G|^2$ to be lower than one, the symmetric matrix $(2\bar{\bar{\beta}} - \vec{\sigma} x \vec{\sigma}^T)$ must be non-negative definite. In particular, the diagonal elements $(2\beta_m - \sigma_m^2)$ must be non-negative, implying $\sigma_m^2 < 2\beta_m$. If one of the β_m is zero, then σ_m (or a_m) is also zero and the corresponding mth dimension can be dropped from the problem. Therefore we can assume all $\beta_m > 0$ and the equality sign on the lower bound of equation (8.6.17) has to be removed.

Defining the diagonal matrix $\bar{\bar{\gamma}}$ by $\bar{\bar{\gamma}} = \text{diag}((2\beta_1)^{1/2}, \ldots, (2\beta_M)^{1/2})$, we have

$$2\bar{\bar{\beta}} - \vec{\sigma} \otimes \vec{\sigma}^T = \bar{\bar{\gamma}} (\bar{\bar{I}} - \bar{\bar{\gamma}}^{-1} \vec{\sigma} \otimes \vec{\sigma}^T \bar{\bar{\gamma}}^{-1}) \bar{\bar{\gamma}} \tag{8.6.20}$$

and the matrix

$$\bar{\bar{A}} \equiv \bar{\bar{I}} - (\bar{\bar{\gamma}}^{-1} \vec{\sigma}) \otimes (\bar{\bar{\gamma}}^{-1} \vec{\sigma})^T \equiv \bar{\bar{I}} - \vec{d} \otimes \vec{d}^T \tag{8.6.21}$$

should also be non-negative definite. Considering the associated quadratic form, for any M-dimensional vector $\vec{x}$,

$$\vec{x}^T \bar{\bar{A}} \vec{x} = \vec{x}^T \cdot \vec{x} - (\vec{d}^T \cdot \vec{x})^2 \tag{8.6.22}$$

the matrix $\bar{\bar{A}}$ is non-negative definite if and only if

$$\vec{d}^T \cdot \vec{d} = \sum_{m=1}^{M} \frac{\sigma_m^2}{2\beta_m} \leqslant 1 \tag{8.6.23}$$

The Von Neumann stability conditions are therefore

$$0 < \sum_{m=1}^{M} \beta_m \leqslant \frac{1}{2} \tag{8.6.24a}$$

and

$$\sum_{m=1}^{M} \frac{\sigma_m^2}{\beta_m} \leqslant 2 \tag{8.6.24b}$$

Assuming all α_m positive, we can easily prove that these conditions are sufficient from the Schwartz inequality applied to the sum

$$\begin{aligned}\left[\sum_{m=1}^{M} \sigma_m \sin\phi_m\right]^2 &\leqslant \left[\sum_{m=1}^{M} \left(\frac{|\sigma_m|}{\sqrt{\beta_m}}\right)\left(\sqrt{(\beta_m)}\,|\sin\phi_m|\right)\right]^2 \\ &\leqslant \sum_{m=1}^{M} \frac{\sigma_m^2}{\beta_m} \cdot \sum_{m=1}^{M} \beta_m \sin^2\phi_m \\ &< \sum_{m=1}^{M} 2\beta_m \sin^2\phi_m \end{aligned} \tag{8.6.25}$$

where the second condition (8.6.24b) has been applied.

If any $\beta_m = 0$ the above condition implies that $\sigma_m = 0$ and the sum (8.6.25) is obtained by summing first only over those m for which $\beta_m > 0$. Inserting this relation into the expression of $|G|^2$, we obtain

$$\begin{aligned}|G|^2 &\leqslant \left[1 - 4\sum_{m=1}^{M} \beta_m \sin^2\phi_m/2\right]^2 + 8\sum_{m=1}^{M} \beta_m \sin^2\phi_m/2 \cdot \cos^2\phi_m/2 \\ &= 1 - 8\sum_{m=1}^{M} \beta_m \sin^4\phi_m/2 + \left[4\sum_{m=1}^{M} \beta_m \sin^2\phi_m/2\right]^2 \end{aligned} \tag{8.6.26}$$

Applying the first stability condition (8.6.24a) with the Schwartz inequality on the last term we obtain

$$\begin{aligned}|G|^2 &\leqslant 1 - 8\sum_{m=1}^{M} \beta_m \sin^4\phi_m/2 + 16\sum_{m=1}^{M} \beta_m \sin^4\phi_m/2 \cdot \sum_{m=1}^{M} \beta_m \\ &< 1 - 8\sum_{m=1}^{M} \beta_m \sin^4\phi_m/2 + 8\sum_{m=1}^{M} \beta_m \sin^4\phi_m/2 \\ &< 1 \end{aligned} \tag{8.6.27}$$

This completes the proof that conditions (8.6.24) are necessary and sufficient for the strict Von Neumann stability of scheme (8.6.13).

Example 8.6.2 The two-dimensional Lax–Friedrichs scheme (8.4.14)

Writing (8.4.14) in the above form we have

$$\sigma_1 = \frac{\Delta t}{\Delta x}\lambda_{\max}^{(A)} \qquad \sigma_2 = \frac{\Delta t}{\Delta y}\lambda_{\max}^{(B)} \tag{E8.6.3}$$

and

$$\beta_1 = \beta_2 = \tfrac{1}{4} \tag{E8.6.4}$$

leading to condition (8.4.19).

Example 8.6.3 Two-dimensional convection–diffusion equation

Consider the energy equation

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} = \alpha\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) \tag{E8.6.5}$$

discretized with central differences and an Euler explicit time integration,

$$T_{ij}^{n+1} - T_{ij}^{n} = -\frac{\sigma_1}{2}(T_{i+1,j}^{n} - T_{i-1,j}^{n}) - \frac{\sigma_2}{2}(T_{i,j+1}^{n} - T_{i,j-1}^{n}) + \beta_1(T_{i+1,j} - 2T_{ij} + T_{i-1,j}) + \beta_2(T_{i,j+1} - 2T_{ij} + T_{i,j-1}) \tag{E8.6.6}$$

where

$$\sigma_1 = \frac{u\Delta t}{\Delta x} \qquad \sigma_2 = \frac{v\Delta t}{\Delta y}$$
$$\beta_1 = \frac{\alpha\Delta t}{\Delta x^2} \qquad \beta_2 = \frac{\alpha\Delta t}{\Delta y^2} \tag{E8.6.7}$$

The necessary and sufficient stability conditions are as follows:

$$(\beta_1 + \beta_2) = \alpha\ \Delta t\left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2}\right) \leqslant \frac{1}{2} \tag{E8.6.8a}$$

$$\frac{\sigma_1^2}{\beta_1} + \frac{\sigma_2^2}{\beta_2} = \frac{\Delta t}{\alpha}(u^2 + v^2) \leqslant 2 \tag{E8.6.8b}$$

Hence the maximum allowable time step is given by

$$\Delta t \leqslant \mathrm{Min}\left(\frac{1}{2\alpha}\frac{\Delta x^2\ \Delta y^2}{\Delta x^2 + \Delta y^2}, \frac{2\alpha}{q^2}\right) \tag{E8.6.9}$$

where $q^2 = u^2 + v^2$ is the square of the velocity $\vec{v}(u, v)$. Observe that the second condition (E8.6.8b) is independent of the mesh sizes Δx, Δy.

Additional remarks: If all β_m are equal, we have the necessary and sufficient condition

$$\sum_{m=1}^{M} \sigma_m^2 \leqslant 2\beta \leqslant \frac{1}{M} \qquad (\beta_m = \beta) \tag{8.6.28}$$

Otherwise for $\beta = \mathrm{Max}_m \beta_m$ the above condition is sufficient.
Introducing the mesh Reynolds, or Peclet, numbers $\mathrm{R}_m = (a_m\ \Delta x_m/\alpha_m)$ the

stability condition (8.6.24b) can be written as

$$0 < \sum_{m=1}^{M} \sigma_m R_m \leqslant 2 \tag{8.6.29}$$

or as

$$\sum_{m=1}^{M} \beta_m R_m^2 \leqslant 2 \tag{8.6.30}$$

When all R_m are lower than two ($R_m < 2$) it is seen that condition (8.6.24a) is more restrictive. The second condition (8.6.24b) will be the more restrictive one when all the R_m are larger than two. Otherwise both conditions have to be satisfied, that is,

$$\Delta t \leqslant \text{Min}\left(\frac{1}{2\Sigma_m(\alpha_m/\Delta x_m^2)}, \frac{2}{\Sigma_m(a_m^2/\alpha_m)}\right) \tag{8.6.31}$$

8.6.3 General multi-level, multi-dimensional schemes

In the general case (discussed in Section 8.2) the strict Von Neumann stability condition is expressed by requirements on the eigenvalues of the matrix G, obtained as a solution of $\det | G - \lambda I | = 0$. These eigenvalues are the zeros of the polynomial of degree p, when G is a $p \times p$ matrix,

$$P(\lambda) = \det | G - \lambda I | = 0 \tag{8.6.32}$$

The stability condition (8.2.24) requires that all the eigenvalues should be lower than or equal to one, and the eigenvalues $\lambda_j = 1$ should be single. Hence this condition has to be satisfied by the zeros of the polynomial $P(\lambda)$. A polynomial satisfying this condition is called a *Von Neumann polynomial.*

The following remarkable theorem, based on the Schur theory of the zeros of a polynomial, can be found in Miller (1971). Let $\bar{P}(\lambda)$ be the associated polynomial of

$$P(\lambda) = \sum_{k=0}^{p} a_k \lambda^k \tag{8.6.33}$$

$$\bar{P}(\lambda) = \sum_{k=0}^{p} a_{p-k}^* \lambda^k \tag{8.6.34}$$

where a_k^* is the complex conjugate of a_k, and define a reduced polynomial of degree not higher than $(p-1)$:

$$P_1(\lambda) = \frac{1}{\lambda} [\bar{P}(0)P(\lambda) - P(0)\bar{P}(\lambda)] \tag{8.6.35}$$

Then the zeros of $P(\lambda)$ satisfy the stability conditions ($P(\lambda)$ is a Von Neumann polynomial) if and only if

(1) $| \bar{P}(0) | > | P(0) |$ and $P_1(\lambda)$ is a Von Neumann polynomial; or
(2) $P_1(\lambda) \equiv 0$ and the zeros of $dP/d\lambda = 0$ are such that $| \lambda | \leqslant 1$.

Hence applying this theorem reduces the analysis to the investigation of the properties of a polynomial of a lower degree (at least $p-1$). Repeating the application of this theorem to P_1 or to $\mathrm{d}P/\mathrm{d}\lambda$, the degree of the resulting polynomials is further reduced until a polynomial of degree one, which can be more easily analysed, is obtained.

Examples of applications of this technique to various schemes for the convection–diffusion equation can be found in Chan (1984), where some stability conditions for higher-order schemes are obtained for the first time.

Example 8.6.4 Leapfrog scheme applied to the convection–diffusion equation

The equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = \alpha \frac{\partial^2 u}{\partial x^2} \tag{E8.6.10}$$

is discretized with central differences in space and time, leading to a leapfrog scheme with the diffusion terms discretized at level $(n-1)$:

$$u_i^{n+1} - u_i^{n-1} = -\sigma(u_{i+1}^n - u_{i-1}^n) + 2\beta(u_{i+1}^{n-1} - 2u_i^{n-1} + u_{i-1}^{n-1}) \tag{E8.6.11}$$

The amplification factors or eigenvalues are solutions of the second-order polynomial (applying the method of Section 8.3.3)

$$P(\lambda) \equiv \lambda^2 + 2\lambda\sigma I \sin\phi - 1 - 4\beta(\cos\phi - 1) = 0 \tag{E8.6.12}$$

We obtain

$$\bar{P}(\lambda) = 1 - 2\lambda\sigma I \sin\phi - [1 + 4\beta(\cos\phi - 1)]\lambda^2 \tag{E8.6.13}$$

The condition $|\bar{P}(0)| > |P(0)|$ leads to

$$|1 - 4\beta(1 - \cos\phi)| < 1$$

or

$$4\beta < 1 \tag{E8.6.14}$$

which is a necessary condition for stability.
Constructing $P_1(\lambda)$, we obtain, with $\gamma = 1 - 4\beta(1 - \cos\phi)$,

$$P_1(\lambda) = \lambda(1 - \gamma^2) + 2\sigma I(1 - \gamma)\sin\phi \tag{E8.6.15}$$

and the stability condition becomes $|\bar{P}_1(0)| > |P_1(0)|$, or

$$|\sigma \sin\phi| \leqslant 1 - 2\beta(1 - \cos\phi) \tag{E8.6.16}$$

Following Chan (1984), this leads to the necessary and sufficient condition, for all ϕ,

$$\sigma^2 + 4\beta \leqslant 1 \tag{E8.6.17}$$

Summary

The Von Neumann stability method, based on a Fourier analysis in the space domain, has been developed for linear, one- and multi-dimensional problems. This method is the most widely applied technique and the amplification factor is easily obtained. Although the stability conditions cannot always be derived analytically, we could, if necessary, analyse the properties of the amplification matrix numerically. These properties also contain information on the dispersion and diffusion errors of a numerical scheme, allowing the selection of a scheme as a function of the desired properties. For non-linear problems it has been shown that a local, linearized stability analysis will lead to necessary conditions.

References

Book, D. L. (Ed.) (1981). *Finite Difference Techniques for Vectorized Fluid Dynamics Calculations*, New York: Springer Verlag.

Briggs, W. L., Newell, A. C., and Sarie, T. (1983). 'Focusing: a mechanism for instability of nonlinear finite difference equations.' *Journal of Computational Physics*, **51**, 83–106.

Cathers, B., and O'Connor, B. A. (1985). 'The group velocity of some numerical schemes.' *Int. Journal for Numerical Methods in Fluids*, **5**, 201–24

Chan, T. F. (1984). 'Stability analysis of finite difference schemes for the advection–diffusion equation.' *SIAM Journal of Numerical Analysis*, **21**, 272–83.

Charney, J. G., Fjortoft, R., and Von Neumann, J. (1950). 'Numerical integration of the barotropic vorticity equation.' *Tellus*, **2**, 237–54.

Courant, R., Friedrichs, K. O., and Lewy, H. (1928). 'Uber die partiellen differenzgleichungen der mathematischen Physik.' *Mathematische Annalen*, **100**, 32–74. English translation in *IBM Journal* (1967), 215–34.

Courant, R., and Hilbert, D. (1962). *Methods of Mathematical Physics*, Vols I and II, New York: John Wiley Interscience.

Crank, J., and Nicholson, P. (1947). 'A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type.' *Proceedings of the Cambridge Philosophical Society*, **43**, 50–67.

Du Fort, E. C., and Frankel, S. P. (1953). 'Stability conditions in the numerical treatment of parabolic equations.' *Math. Tables and Other Aids to Computation*, **7**, 135–52.

Hindmarsh, A. C., Gresho, P. M., and Griffiths, D. F. (1984). 'The stability of explicit Euler time integration for certain finite difference approximations of the multidimensional advection-diffusion equation.' *Int. Journal for Numerical Methods in Fluids*, **4**, 853–97.

Hirt, C. W. (1968). 'Heuristic stability theory for finite difference equations.' *Journal of Computational Physics*, **2**, 339–55.

Kreiss, H. O. (1964). 'On difference approximations of the dissipative type for hyperbolic differential equations.' *Comm. Pure and Applied Mathematics*, **17**, 335–53.

Lambert, J. D. (1973). *Computational Methods in Ordinary Differential Equations*, Chichester: John Wiley.

Lax, P. D. (1954). 'Weak solutions of nonlinear hyperbolic equations and their numerical computation.' *Comm. Pure and Applied Mathematics*, **7**, 159–93.

Lax, P. D., and Wendroff, B. (1960). 'Systems of conservation laws.' *Comm. Pure and Applied Mathematics*, **13**, 217–37.
McAveney, B. J., and Leslie, L. M. (1972). 'Comments on a direct solution of Poisson's equation by generalized sweep-out method.' *Journal Meteor, Society of Japan*, **50**, 136.
McDonough, J. M., and Bywater, R. J. (1986). 'Large-scale effects on local small-scale chaotic solutions to Burger's equation.' *AIAA Journal*, **24**, 1924–30.
Miller, J. J. H. (1971). 'On the location of zeros of certain classes of polynomials with applications to numerical analysis.' *Journal Inst. Math. Applic.*, **8**, 397–406.
Peyret, R., and Taylor, T. D. (1983). *Computational Methods for Fluid Flow*, New York: Springer Verlag.
Richtmyer, R. D., and Morton, K. W. (1967). *Difference Methods for Initial Value Problems*, 2nd edn, Chichester: John Wiley/Interscience.
Roache, P. J. (1971). 'A new direct method for the discretized Poisson equation.' *Second Int. Conf. on Num. Methods in Fluid Dynamics*, New York: Springer Verlag.
Sloan, D. M., and Mitchell, A. R. (1986). 'On nonlinear instabilities in leap-frog finite difference schemes.' *Journal of Computational Physics*, **67**, 372–95.
Trefethen, L. N. (1982). 'Group velocity in finite difference schemes.' *SIAM Review*, **24**, 113–36
Trefethen, L. N. (1983). 'Group velocity interpretation of the stability theory of Gustafsson, Kreiss and Sundstrom.' *Journal of Computational Physics*, **49**, 199–217.
Trefethen, L. N. (1984). 'Instability of difference models for hyperbolic initial boundary value problems.' *Comm. Pure and Applied Mathematics*, **37**, 329–67.
Varga, R. S. (1962). *Matrix Iterative Analysis*, Englewood Cliffs, NJ: Prentice-Hall.
Vichnevetsky, R., and Bowles, J. B (1982). *Fourier Analysis of Numerical Approximations of Hyberbolic Equations*, Philadelphia: SIAM Publications.

PROBLEMS

Problem 8.1

Derive the succession of operators for the various examples of Table 8.1.

Problem 8.2

Apply a forward space differencing with a forward time difference (Euler method) to the convective equation $u_t + au_x = 0$. Analyse the stability with the Von Neumann method and show that the scheme is unconditionally unstable for $a > 0$ and conditionally stable for $a < 0$. Derive also the equivalent differential equation and show why this scheme is unstable when $a > 0$.

Problem 8.3

Show that the forward/backward scheme for the second-order wave equation (E8.2.12) is

$$w_i^{n+1} - 2w_i^n + w_{i-1}^{n-1} = \left(\frac{a\Delta t}{\Delta x}\right)^2 (w_{i+1}^n - 2w_i^n + w_{i-1}^n)$$

referring to Table 8.1. Obtain the explicit form of the operators and matrices S, C and G.

Problem 8.4

Consider the same space discretization of the second-order wave equation as in the previous problem but apply a full Euler scheme (forward in time):

$$v_i^{n+1} - v_i^n = \sigma(w_{i+1}^n - w_i^n)$$
$$w_i^{n+1} - w_i^n = \sigma(v_i^n - v_{i-1}^n)$$

Calculate for this scheme the operators and corresponding matrices C and G.

Problem 8.5

Solve the one-dimensional heat conduction equation $u_t = \alpha u_{xx}$ for the following conditions, with k an integer:

$$u(x, 0) = \sin k\pi x \qquad 0 \leqslant x \leqslant 1$$
$$u(0, t) = 0$$
$$u(1, t) = 0$$

applying the explicit central scheme (8.3.16). Compare with the exact solution for different values of β, in particular $\beta = 1/3$ and $\beta = 1/6$ (which is the optimal value). Consider initial functions with different wavenumbers k, namely $k = 1, 5, 10$.

The exact solution is $u = e^{-\alpha k^2 \pi^2 t} \sin k\pi x$. Compute with $x_i = i\Delta x$ and i ranging from 0 to 30. Make plots of the computed solution as a function of x and of the exact solution. Perform the calculations for five and ten time steps and control the error by comparing with equation (8.3.25) for ε_D in the case of $\beta = 1/3$. Calculate the higher-order terms in ε_D for $\beta = 1/6$ by taking more terms in the expansion.

Problem 8.6

Calculate the amplitude and phase errors for the Lax–Friedrichs scheme (E8.3.1) after ten time steps for an initial wave of the form

$$u(x, 0) = \sin k\pi x \qquad 0 \leqslant x \leqslant 1$$

for $k = 1, 10$. Consider $\Delta x = 0.02$ and a velocity $a = 1$. Perform the calculations for $\sigma = 0.25$ and $\sigma = 0.75$. Plot the computed and exact solutions for these various cases and compare and comment on the results.

Hint: The exact solution is $\tilde{u} = \sin \pi k(x - t)$. The exact numerical solution is $\bar{u}_i^n = |G|^n \sin \pi k(x_i - \bar{a}n\ \Delta t)$ where $\bar{a}$ is the numerical speed of propagation and is equal to a ε_ϕ (equation (E8.3.5)). Show that we can write $\bar{u}_i^n = |G|^n \sin[\pi k(x_i - n\Delta t) + n(\tilde{\phi} - \phi)]$.

Problem 8.7

Repeat Problem 8.6 with the upwind scheme (7.2.8).

Problem 8.8

Repeat Problem 8.6 with the leapfrog scheme (8.3.33).

Problem 8.9

Apply the central difference in time (leapfrog scheme) to the heat-conduction equation with the space differences of second-order accuracy:

$$u_i^{n+1} - u_i^{n-1} = 2\frac{\alpha\Delta t}{\Delta x^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n)$$

Calculate the amplification matrix and show that the scheme is unconditionally unstable.

Problem 8.10

Analyse the leapfrog scheme with the upwind space discretization of the convection equation $u_t + au_x = 0$. This is the scheme

$$u_i^{n+1} - u_i^{n-1} = -2\sigma(u_i^n - u_{i-1}^n)$$

Calculate the amplification matrix and show that the scheme is unstable.

Problem 8.11

Consider the implicit upwind scheme (7.2.9) and analyse its stability. Show that the scheme is unconditionally stable for $a > 0$ and unstable for $a < 0$.

Problem 8.12

Write a program to solve the linear convection equation and obtain Figure 8.3.6.

Problem 8.13

Write a program to solve the linear convection equation and obtain Figures 8.3.7 and 8.3.8 for the wave packet problem. Compare with similar calculations at $CFL = 0.2$

Problem 8.14

Apply an upwind (backward in space) discretization to the two-dimensional convection equation

$$u_t + au_x + bu_y = 0$$

with an explicit Euler time integration. Perform a Von Neumann stability analysis and show that we obtain the CFL conditions in both directions.

Problem 8.15

Apply the Dufort–Frankel scheme to the leapfrog convection-diffusion equation

$$u_i^{n+1} - u_i^{n-1} = -\sigma(u_{i+1}^n - u_{i-1}^n) + 2\beta(u_{i+1}^n - u_i^{n+1} - u_i^{n-1} + u_{i-1}^n)$$

and show, by application of Section 8.6.3, that the stability condition is $|\sigma| < 1$, independently of the diffusion related coefficient β.
Hint: Write the scheme as

$$u_i^{n+1} - u_i^{n-1} = -\sigma(u_{i+1}^n - u_{i-1}^n) + 2\beta(u_{i+1}^n - 2u_i^n + u_{i-1}^n) - 2\beta(u_i^{n+1} - 2u_i^n + u_i^{n-1})$$

Problem 8.16

Write a program to solve Burger's equation for the stationary shock and obtain the results of Figures 8.5.1 and 8.5.2.

Problem 8.17

Obtain the results of Example 8.3.3 for the Lax-Wendroff scheme.

Problem 8.18

Obtain the results of Example 8.3.4 for the leapfrog scheme and derive the expansion of the dispersion error in powers of the phase angle.

Problem 8.19

Obtain the relations of Example 8.3.5.

Problem 8.20

Apply a central time integration (leapfrog) method to the finite element scheme of Problem 5.14, considering the linear equation $f = au$. The scheme is

$$(u_{l-1}^{n+1} - u_{l-1}^{n-1}) + 4(u_l^{n+1} - u_l^{n-1}) + (u_{l+1}^{n+1} - u_{l+1}^{n-1}) + 6\sigma(u_{l+1}^n - u_{l-1}^n) = 0$$

Determine the amplification factor and obtain the stability condition $\sigma \leqslant 1/\sqrt{3}$. Determine the dispersion and diffusion errors and obtain the numerical group velocity. Compare with the leapfrog scheme (8.3.33).
Hint: The amplification factor is

$$G = -Ib \pm \sqrt{1 - b^2} \qquad b = \frac{3\sigma \sin \phi}{2 + \cos \phi}$$

Problem 8.21

Apply a generalized trapezium formula in time as defined by Problem 7.5, to the finite element discretization of Problem 5.14. Obtain the scheme

$$(u_{l-1}^{n+1} - u_{l-1}^n) + 4(u_l^{n+1} - u_l^n) + (u_{l+1}^{n+1} - u_{l+1}^n) + 3\theta\sigma(u_{l+1}^{n+1} - u_{l-1}^{n+1}) + 3(1-\theta)\sigma(u_{l+1}^n - u_{l-1}^n) = 0$$

and derive the amplification factor

$$G = \frac{2 + \cos \phi - 3(1-\theta)I\sigma \sin \theta}{2 + \cos \phi + 3\theta I\sigma \sin \phi}$$

Show that the scheme is unconditionally stable for $\theta \geqslant 1/2$.

Show that for $\theta = 1/2$ (the trapezium formula), there is no dissipation error, and obtain the numerical dispersion relation as $\tan(\omega \Delta t/2) = 3\sigma \sin \phi / [2(2 + \cos \phi)]$.

Problem 8.22

Find the numerical group velocities for the upwind, Lax-Friedrichs and Lax-Wendroff schemes for the linear convection equation, applying the relation (8.3.41) to the real part of the numerical dispersion relation. Plot the ratios v_G/a in function of ϕ and observe the deviations from the exact value of 1.
Hint: Obtain

First order upwind scheme $v_G = a[(1-\sigma)\cos \phi + \sigma] / [(1 + \sigma(\cos \phi - 1))^2 + \sigma^2 \sin^2 \phi]$

Lax-Friedrichs scheme $v_G = a/(\cos^2 \phi + \sigma^2 \sin^2 \phi)$

Lax-Wendroff scheme $v_G = a[(1 - 2\sigma^2 \sin^2 \varphi/2)\cos \varphi + \sigma^2 \sin^2 \varphi] / [(1 - 2\sigma^2 \sin^2 \varphi/2)^2 + \sigma^2 \sin^2 \varphi]$