Sudan University of Science and Technology

College of Graduate Studies

Numerical Schemes for Hyperbolic Equation in One Space Dimension

الطرق العددية للمعادلات الزائدية في بعد مكانى واحد

A thesis Submitted in Partial Fulfillment for the Degree of M. Sc in Mathematics

By: Rayan Adil Mohamed Ahmed

Supervised By: Dr. Mohamed Hassan Mohamed Khabir

2015

Dedication

This research is dedication:-

To my parents

To my sisters

To my aunt

To all my family member

To my friends

To someone who has a lot in my Deep down...

Acknowledgement

I would like to express my special thanks to:-

Dr: Mohamed Hassan Mohamed Khabir

My father: Adil Mohamed Ahmed

My mother : Hasnat seed

My special thanks extended to all member of faculty of sciences ,to my family , to my friend and to every body who help me .

Abstract

We find the approximate solution for hyperbolic equation in one space dimension using two finite different schemes: Lax- Wendroff and upwind schemes Then, we study Fourier analysis of these two schemes. we also approximate the numerical solution of system of hyperbolic equations by using finite volume scheme and leap-frog schemes. As well, we study the Fourier analysis of these two schemes. Finally, we study the consistency, convergence and stability for hyperbolic equation in one space dimension and we state and prove the main part of the key lax Equivalence theorem.

.

الخلاصة

تناولنا في هذا البحث إيجاد الحلول التقريبية للمعادلة الزائدية في بعد مكاني واحد باستخدام الفروقات المنتهية وهي: طريقة لاكس وندروف وطريقة اب وند. ثم قمنا طريقتين من طرق بدراسة تحليل فوريير لهاتين الطريقتين. أيضا قربنا الحلول العددية لمجموعة معادلات زائدية باستخدام طريقة الحجم المنتهي وطريقة قفزة الضفدعة وأوجدنا تحليل فورييرلهاتين الطريقتين لمجموعة المعادلات الزائدية. قمنا بدراسة الموائمة والتقارب والاستقرار للمعادلة الزائدية في بعد مكاني واحد وقمنا بكتابة واثبات الجزء الأساسي من نظرية لاكس المكافئة.

The Content

Chapter one

Hyperbolic Equations in One Space Dimension

1.1 Characteristics

The linear advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \qquad (1.1)$$

is an example of the simplest of partial differential equations. Yet to approximate it well on a fixed (x,t)-mesh is a far from trivial problem that is still under active discussion in the numerical analysis literature. Of course, the exact solution is obtained from observing that this is a hyperbolic equation with a single set of characteristics and u is constant along each such characteristic: the characteristics are the solutions of the ordinary differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a(x,t), \qquad (1.2)$$

and along a characteristic curve the solution $u(x,t)$satisfies

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} = 0. \qquad (1.3)$$

Thus from initial data

$$u(x,0) = u^0(x), \qquad (1.4)$$

where $u^0(x)$ is a given function, we can construct an approximate solution by choosing a suitable set of points $x_0, x_1, \ldots$, as in Figure 1.1, and finding the characteristic through $(x_j, 0)$ by a numerical solution of (1.2) with the initial condition $x(0) = x_j$. At all points on this curve we then have $u(x,t) = u^0(x_j)$. This is called the method of characteristics.

Note that for this linear problem in which $a(x,t)$ is a given function, the characteristics cannot cross so long as a is Lipschitz continuous in x and continuous in t.

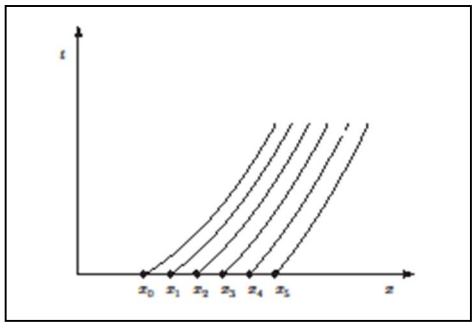

Figure (1.1): Typical characteristics for $u_t + a(x,t)u_x = 0$.

When a is a constant the process is trivial. The characteristics are the parallel straight lines $x - at =$ constant, and the solution is simply

$$u(x,t) = u^0(x - at). \qquad (1.5)$$

Moreover, in the nonlinear problem in which a is a function only of $u, a = a(u)$, the characteristics are also straight lines because u is constant along each, although they are not now parallel. Thus again we are able to write the solution in the form

$$u(x,t) = u^0\big(x - a\big(u(x,t)\big)t\big), \qquad (1.6)$$

until the time when this breaks down because the characteristics can now envelope or cross each other in some other manner.

Consideration of the characteristics of the equation, or system of equations, is essential in any development or study of numerical methods for hyperbolic equations and we shall continually refer to them below. We shall want to consider systems of conservation laws of the form

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{u})}{\partial x} = 0 \qquad (1.7)$$

where $u = u(x,t)$ is a vector of unknown functions and $\mathbf{f}(\mathbf{u})$ a vector of flux functions. For example, if the vector $\mathbf{u}$ has two components u and v,

and $\mathbf{f}$ has two components $f(u,v)$ and $g(u,v)$, we can write out the components of (1.7) as

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u,v) = 0, \tag{1.8}$$

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x} g(u,v) = 0, \tag{1.9}$$

or in matrix form

$$\begin{pmatrix} \frac{\partial u}{\partial t} \\ \frac{\partial v}{\partial t} \end{pmatrix} + \begin{pmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{pmatrix} \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.10}$$

If we define

$$A(\mathbf{u}) := \frac{\partial \mathbf{f}}{\partial \mathbf{u}}, \tag{1.11}$$

the Jacobian matrix formed from the partial derivatives of $\mathbf{f}$, we can write the system as

$$\mathbf{u}_t + A(\mathbf{u})\mathbf{u}_x = 0, \tag{1.12}$$

and the characteristic speeds are the eigenvalues of A. The hyperbolicity of the system is expressed by the fact that we assume A has real eigenvalues and a full set of eigenvectors. Suppose we denote by $\mathbf{\Lambda}$ the diagonal matrix of eigenvalues and by $S = S(\mathbf{u})$ the matrix of left eigenvectors, so that

$$SA = \Lambda S. \tag{1.13}$$

Then premultiplying (1.12) by S gives the characteristic normal form of the equations

$$S\mathbf{u}_t + \Lambda S\mathbf{u}_x = 0. \tag{1.14}$$

If it is possible to define a vector of Riemann invariants $r = r(\mathbf{u})$ such that $\mathbf{r}_t = S\mathbf{u}_t$ and $\mathbf{r}_x = S\mathbf{u}_x$, then we can write

$$\mathbf{r}_t + \Lambda \mathbf{r}_x = 0 \tag{1.15}$$

which is a direct generalisation of the scalar case whose solution we have given in (1.6). However, now each component of $\mathbf{\Lambda}$ will usually depend on all the components of r so that the characteristics will be curved.

Moreover, although these Riemann invariants can always be defined for a system of two equat 1ions, for a larger system this is not always possible.

To apply the method of characteristics to problems like (1.7), where the characteristic speeds depend on the solution, one has to integrate forward simultaneously both the ordinary differential equations for the characteristic paths and the characteristic normal form (1.14) of the differential equations. This is clearly a fairly complicated undertaking, but it will give what is probably the most precise method for approximating this system of equations.

The CFL condition

Courant, Friedrichs and Lewy, in their fundamental 1928 paper1 on difference methods for partial differential equations, formulated a necessary condition now known as the CFL condition for the convergence of a difference approximation in terms of the concept of a domain of dependence. Consider first the simplest model problem (1.1), where a is a positive constant; as we have seen, the solution is $u(x,t) = u^0(x-at)$, where the function u^0 is determined by the initial conditions. The solution at the point (x_j, t_n) is obtained by drawing the characteristic through this point back to where it meets the initial line at $Q \equiv (x_j - at_n, 0)$ – see Figure (1.2).

Now suppose that we compute a finite difference approximation by using the explicit scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0. \qquad (1.16)$$

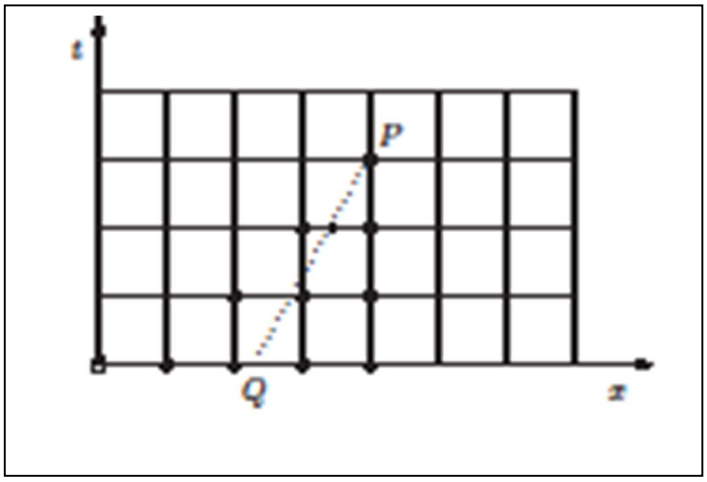

Figure (1.2): Typical domain of dependence.

Then the value on the new time level will be calculated from

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{\Delta x}\left(U_j^n - U_{j-1}^n\right)$$

$$= (1 - v)U_j^n + vU_{j-1}^n, \quad (1.17)$$

where

$$v = \frac{a\Delta t}{\Delta x}. \quad (1.18)$$

The value of U_j^{n+1} depends on the values of U at two points on the previous time level; each of these depends on two points on the time level t_{n-1}, and so on. As illustrated in Figure (1.2), the value of U_j^{n+1} depends on data given in a triangle with vertex (x_j, t_{n+1}), and ultimately on data at the points on the initial line

$$x_{j-n-1}, x_{j-n}, \dots, x_{j-1}, x_j.$$

For an inhomogeneous equation in which a source term h_j^n replaces the zero on the right-hand side of (1.16), U_j^{n+1} depends on data given at all points of the triangle. This triangle is called the domain of dependence of U_j^{n+1}, or of the point (x_j, t_{n+1}), for this particular numerical scheme.

The corresponding domain of dependence of the differential equation is the characteristic path drawn back from (x_j, t_{n+1}) to the initial line, for in the inhomogeneous case $u_t + au_x = h$ data values $h(x,t)$ are picked up along the whole path as well as the initial data at $x = x_j - at_{n+1}$.

The CFL condition then states that for a convergent scheme the domain of dependence of the partial differential equation must lie within the domain of dependence of the numerical scheme.

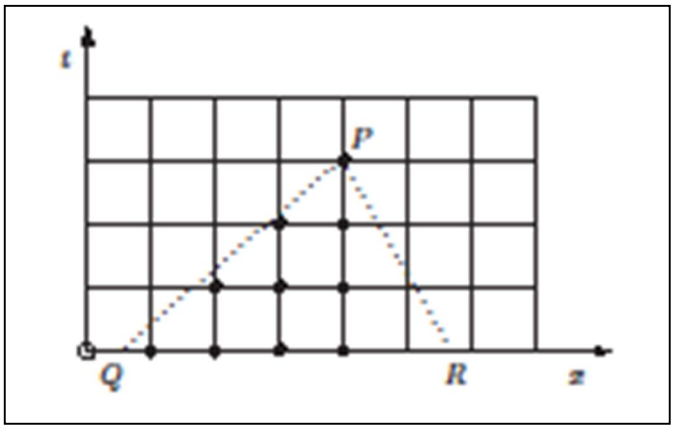

Figure (1.3): Violation of the CFL condition.

Figure (1.3) illustrates two situations in which this condition is violated. Either of the characteristics PQ or PR lies outside the triangle. Suppose that we consider a refinement path on which the ratio $\Delta t/\Delta x$ is constant; then the triangular domain of dependence remains the same. But suppose we alter the given initial conditions in a small region of the initial line $t = 0$ around the point Q. This will then alter the solution of the differential equation at P, since the solution is constant along the characteristic QP. The numerical solution at P, however, remains unaltered, since the numerical data used to construct the solution are unchanged. The numerical solution therefore cannot converge to the required result at P. The same argument of course applies in the same way to the characteristic RP.

The CFL condition shows in this example that the scheme cannot converge for a differential equation for which $a < 0$, since this would give a characteristic like RP. And if $a > 0$ it gives a restriction on the size of the

time step, for the condition that the characteristic must lie within the triangle of dependence requires that $a\,\Delta t/\Delta x \leq 1$.

What we have thus obtained can also be regarded as a necessary condition for the stability of this difference scheme, So far it is only a necessary condition. In general the CFL condition is not sufficient for stability, as we shall show in some examples. Its great merit lies in its simplicity; it enables us to reject a number of difference schemes with a trivial amount of investigation. Those schemes which satisfy the CFL condition may then be considered in more detail, using a test which is sufficient for stability.

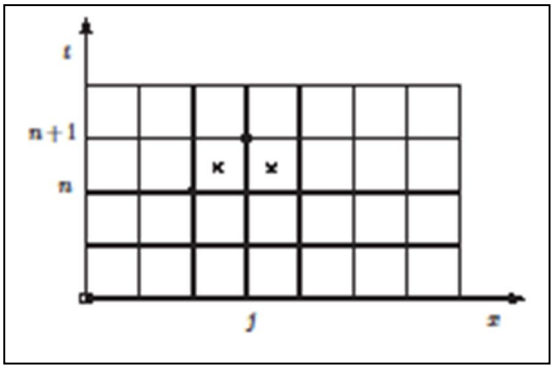

Figure (1.4): General three-point scheme; the points marked × are used for the two-step Lax−Wendroff method.

Now suppose that we approximate the advection equation (1.1) by a more general explicit scheme using just the three symmetrically placed points at the old time level. The CFL condition becomes

$$|a|\Delta t \leq \Delta x, \qquad (1.19)$$

as we see from Figure (1.4); $\nu := |a|\,\Delta t/\Delta x$ is often called the CFL number.

If $a > 0$, the difference scheme must use both U_{j-1}^n and U_j^n to obtain U_j^{n+1}: and if $a < 0$ it must use U_j^n and U_{j+1}^n. To cover both cases we might be tempted to use a central difference in space together with a forward difference in time to obtain

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0. \qquad (1.20)$$

If we satisfy (1.19) the CFL condition holds for either sign of a.

But now in the case where a is constant, and ignoring the effect of the boundary conditions, we can investigate the stability of the scheme by Fourier analysis, The Fourier mode

$$U_j^n = (\lambda)^n \mathrm{e}^{ik(j\Delta x)} \qquad (1.21)$$

satisfies the difference scheme (1.20) provided that the amplification factor λ satisfies

$$\lambda \equiv \lambda(k) = 1 - (a\,\Delta t/\Delta x) i \sin k\Delta x. \qquad (1.22)$$

Thus $|\lambda| > 1$ for all mesh ratios (and almost all modes) and the scheme is unstable for any refinement path along which $a\,\Delta t/\Delta x$ is fixed. Note that this is a case when the highest frequency mode, $k\Delta x = \pi$ or $U_j \propto (-\mathbf{1})^j$, does not grow: but the mode with $k\Delta x = \frac{1}{2}\pi$, or where U_j takes successive values $\ldots, -1,0,1,0,-1,\ldots$, grows in magnitude by $[1 + (a\,\Delta t/\Delta x)^2]^{1/2}$ at each step while shifting to the right. This central difference scheme thus satisfies the CFL condition but is nevertheless always unstable, illustrating the earlier comment that the CFL condition is necessary, but not sufficient, for stability.

1.2 The Upwind Scheme

We define finite differences in the same way in the two variables t and x; there are three kinds of finite differences:

Forward differences

$$\Delta_{+t} v(x,t) := v(x, t + \Delta t) - v(x,t),$$

$$\Delta_{+x} v(x,t) := v(x + \Delta x, t) - v(x,t);$$

Backward differences

$$\Delta_{-t} v(x,t) := v(x,t) - v(x, t - \Delta t),$$

$$\Delta_{-x}v(x,t) := v(x,t) - v(x-\Delta x,t);$$

Central differences

$$\delta_t v(x,t) := v\left(x, t+\frac{1}{2}\Delta t\right) - v\left(x, t-\frac{1}{2}\Delta t\right),$$

$$\delta_x v(x,t) := v\left(x+\frac{1}{2}\Delta x, t\right) - v\left(x-\frac{1}{2}\Delta x, t\right).$$

When the central difference operator is applied twice we obtain the very useful second order central difference

$$\delta_x^2 v(x,t) := v(x+\Delta x,t) - 2v(x,t) + v(x-\Delta x,t).$$

For first differences it is often convenient to use the double interval central difference

$$\begin{aligned}\Delta_{0x}v(x,t) &:= \frac{1}{2}(\Delta_{+x} + \Delta_{-x})v(x,t)\\ &= \frac{1}{2}[v(x+\Delta x,t) - v(x-\Delta x,t)].\end{aligned}$$

The simplest and most compact stable scheme involving these three points is called an upwind scheme because it uses a backward difference in space if a is positive and a forward difference if a is negative:

$$U_j^{n+1} = \begin{cases} U_j^n - a\dfrac{\Delta t}{\Delta x}\Delta_{+x}U_j^n & \text{if } a < 0, \\ U_j^n - a\dfrac{\Delta t}{\Delta x}\Delta_{-x}U_j^n & \text{if } a > 0. \end{cases} \qquad (1.23)$$

If a is not a constant, but a function of x and t, we must specify which value is used in (1.23). We shall for the moment assume that we use $a(x_j, t_n)$, but still write a without superscript or subscript and $v = a\,\Delta t/\Delta x$ as in (1.18) when this is unambiguous.

This scheme satisfies the CFL condition when (1.19) is satisfied, and a Fourier analysis gives for the constant $a > 0$ case the amplification factor

$$\lambda \equiv \lambda(k) = 1 - (a\,\Delta t/\Delta x)\left(1 - \mathrm{e}^{-ik\Delta x}\right) \equiv 1 - v\left(1 - \mathrm{e}^{-ik\Delta x}\right). \qquad (1.24)$$

This leads to

$$\begin{aligned} |\lambda|^2 &= [(1-v) + v\cos k\Delta x]^2 + [v \sin k\Delta x]^2 \\ &= (1-v)^2 + v^2 + 2v(1-v)\cos k\Delta x \\ &= 1 - 2v(1-v)(1-\cos k\Delta x) \end{aligned}$$

which gives

$$|\lambda|^2 = 1 - 4v(1-v)\sin^2 \frac{1}{2} k\Delta x. \qquad (1.25)$$

It follows that $|\lambda(k)| \leq 1$ for all k provided that $0 \leq v \leq 1$. The same analysis for the case where $a < 0$ shows that the amplification factor $\lambda(k)$ is the same, but with a replaced by $|a|$. Thus in this case the CFL condition gives the correct stability limits.

Error analysis of the upwind scheme

We notice that the scheme (1.23) can be written

$$U_j^{n+1} = \begin{cases} (1+v)U_j^n - vU_{j+1}^n & \text{if } a < 0, \\ (1-v)U_j^n + vU_{j-1}^n & \text{if } a > 0. \end{cases} \qquad (1.26)$$

This can be interpreted as follows. In Figure (1.5) for the case $a > 0$, the characteristic through the point $P = (x_j, t_{n+1})$ meets the previous line $t = t_n$ at the point Q, which by the CFL condition must lie between the points $A = (x_j, t_n)$ and $B = (x_{j-1}, t_n)$. Moreover the exact solution $u(x,t)$ is constant along the characteristic, so that $u(P) = u(Q)$. Knowing an approximate numerical solution at all the points on the line t_n, we can therefore interpolate the value of $U(Q)$ and use this to give the required value U_j^{n+1}. If we use linear interpolation, approximating $u(x, t_n)$ by a linear function of x determined by the approximations at the two points A

and B, we obtain (1.26) exactly when a is constant because $AQ = v\Delta x$ and $QB = (1 - v)\Delta x$; when a varies smoothly this still gives a good approximation.

Notice also that all the coefficients in (1.26) are nonnegative so that a maximum principle applies, provided that $|v| \leq 1$ at all mesh points. We can therefore obtain an error bound for the linear, variable coefficient problem just as we have done for parabolic equations. We must first consider more carefully what domain is given, and what conditions should be specified at the boundaries of the domain: although the physical problem may be given on the whole line, for all values of x, a numerical solution must be confined to a finite region. Suppose, for example, that the region of interest is $0 \leq x \leq X$, so that we have boundaries at $x = 0$ and $x = X$. Since the differential equation is hyperbolic and first order, we will usually have only one boundary conditionwhere we were always given a boundary condition at each end of the domain. The direction of the characteristics shows that we need a boundary condition at $x = 0$ if $a > 0$ there, and at $x = X$ if $a < 0$ there; in the straightforward situation where a has the same sign everywhere, we therefore have just the one boundary condition. The exact solution of the differential equation would then be determined by drawing the characteristic backwards from the point P, until it reaches either the initial line $t = 0$, or a boundary on which a boundary condition is given.

For simplicity we shall first suppose that $a > 0$ on $[0, X] \times [0, t_F]$; we consider the general case later. The truncation error of the scheme is

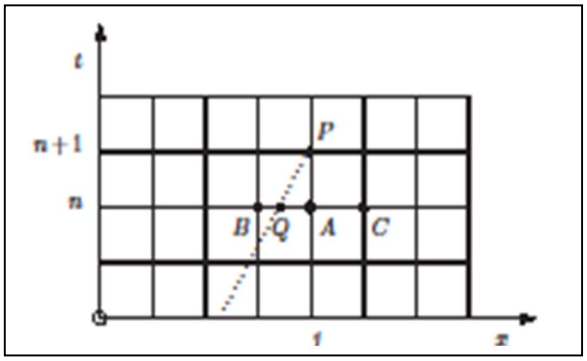

Figure (1.5): Construction of a scheme by linear or quadratic interpolation.

defined as usual and expansion about (x_j, t_n) gives, if u is sufficiently smooth,

$$T_j^n := \frac{u_j^{n+1} - u_j^n}{\Delta t} + a_j^n \frac{u_j^n - u_{j-1}^n}{\Delta x}$$

$$\sim \left[u_t + \frac{1}{2}\Delta t u_{tt} + \cdots\right]_j^n + \left[a\left(u_x - \frac{1}{2}\Delta x u_{xx} + \cdots\right)\right]_j^n$$

$$= \frac{1}{2}(\Delta t u_{tt} - a\Delta u_{xx}) + \cdots. \qquad (1.27)$$

Even if a is constant so that we have $u_{tt} = a^2 u_{xx}$, we still find

$$T_j^n = -\frac{1}{2}(1 - v)a\Delta x u_{xx} + \cdots;$$

hence generally the method is first order accurate. Suppose the difference scheme is applied for $j = 1,2,\ldots,J$, at the points $x_j = j\Delta x$ with $J\Delta x = X$, and the boundary value $U_0^n = u(0, t_n)$ is given. Then for the error $e_j^n = U_j^n - u_j^n$ we have as usual

$$e_j^{n+1} = (1 - v)e_j^n + ve_{j-1}^n - \Delta t e_{j-1}^n - \Delta t T_j^n \qquad (1.28)$$

and $e_0^n = 0$, from which we deduce that if $0 \le v \le 1$ at all points

$$E^{n+1} := \max_j \left|e_j^{n+1}\right| \le E^n + \Delta t \max_j \left|T_j^n\right|.$$

If we suppose that the truncation error is bounded, so that

$$\left|T_j^n\right| \le T \qquad (1.29)$$

for all j and n in the domain, the usual induction argument shows that

$$E^n \le n\Delta t\, T \le t_F T \qquad (1.30)$$

if $U_j^0 = u^0(x_j)$. This result is sufficient to prove first order convergence of the upwind scheme along a refinement path which satisfies the CFL

condition everywhere, provided that the solution has bounded second derivatives.

Now let us consider a completely general set of values

$$\{a_j^n := a(x_j, t_n); \quad j = 0,1, \dots, J\}.$$

It is clear that an equation similar to (1.28) holds at each point: if $a_j^n \geq 0$ and $j > 0$, then (1.28) holds; if $a_j^n \leq 0$ and $j < J$ then a corresponding upwind equation with e_{j-1}^n replacing e_{j+1}^n holds; and the remaining cases, $a_0^n > 0$ or $a_j^n < 0$, correspond to the inflow boundary data being given so that either $e_0^{n+1} = 0$ or $e_j^{n+1} = 0$. The rest of the argument then follows as above.

Fourier analysis of the upwind scheme

Because hyperbolic equations often describe the motion and development of waves, Fourier analysis is of great value in studying the accuracy of methods as well as their stability. The modulus of $\lambda(k)$ describes the damping and the argument describes the dispersion in the scheme, i.e., the extent to which the wave speed varies with the frequency. We must, for the present and for a strict analysis, assume that a is a (positive) constant. The Fourier mode

$$u(x, t) = e^{i(kx+\omega t)} \tag{1.31}$$

is then an exact solution of the differential equation (1.1) provided that ω and k satisfy the dispersion relation

$$\omega = -ak. \tag{1.32}$$

The mode is completely undamped, as its amplitude is constant; in one time step its phase is changed by $-ak\Delta t$. By contrast, the Fourier mode (1.21) satisfies the upwind scheme provided that (1.24) holds. This leads to (1.25), showing that except in the special case $v = 1$ the mode is damped. The phase of the numerical mode is given by

$$\arg\lambda = -\tan^{-1}\left[\frac{v\sin k\Delta x}{(1-v)+v\cos k\Delta x}\right] \qquad (1.33)$$

and we particularly need to evaluate this when $k\Delta x$ is small, as it is such modes that can be well approximated on the mesh. For this, and subsequent schemes, it is useful to have a simple lemma:

Lemma (1.1):

If q has an expansion in powers of p of the form

$$q \sim c_1 p + c_2 p^2 + c_3 p^3 + c_4 p^4 + \cdots$$

as $p \to 0$, then

$$\tan^{-1} q \sim c_1 p + c_2 p^2 + \left(c_3 - \frac{1}{3}c_1^3\right)p^3 + \left(c_4 - \frac{1}{4}c_1^4\right)p^4 + \cdots$$

.

We can now expand (1.33) and apply the lemma, giving

$$\arg\lambda \sim \tan^{-1}\left[v\left(\xi - \frac{1}{6}\xi^3 + \cdots\right)\left(1 - \frac{1}{2}v\xi^2 + \cdots\right)^{-1}\right]$$

$$= -\tan^{-1}\left[v\xi - \frac{1}{6}v(1-3v)\xi^3 + \cdots\right]$$

$$= -v\xi\left[1 - \frac{1}{6}(1-v)(1-2v)\xi^2 + \cdots\right], \qquad (1.34)$$

where we have written

$$\xi = k\Delta x. \qquad (1.35)$$

The case $v = 1$ is obviously very special, as the scheme then gives the exact result. Apart from this, we have found that the upwind scheme always has an amplitude error which, from (1.25), is of order ξ^2 in one time step, corresponding to a global error of order ξ; and from (1.34) it has a relative

phase error of order ξ^2, with the sign depending on the value of v, and vanishing when $v = \frac{1}{2}$.

Some results obtained with the upwind scheme are displayed in Figure (1.6). The problem consists of solving the equation

$$u_t + a(x,t)u_x = 0, \quad x \geq 0, \qquad t \geq 0, \tag{1.36a}$$

where

$$a(x,t) = \frac{1 + x^2}{1 + 2xt + 2x^2 + x^4}, \tag{1.36b}$$

with the initial condition

$$u(x,0) = \begin{cases} 1 & \text{if } 0.2 \leq x \leq 0.4, \\ 0 & \text{otherwise,} \end{cases} \tag{1.37a}$$

and the boundary condition

$$u(0,t) = 0. \tag{1.37b}$$

The exact solution of the problem is

$$u(x,t) = u(x^*,0). \tag{1.38a}$$

where

$$x^* = x - \frac{t}{1 + x^2}. \tag{1.38b}$$

Since $a(x,t) \leq 1$ the calculations use $\Delta t = \Delta x$, and the CFL stability condition is satisfied. The solution represents a square pulse moving to the right. It is clear from the figures how the damping of the high frequency modes has resulted in a substantial smoothing of the edges of the pulse, and a slight reduction of its height. However, the rather small phase error means that the pulse moves with nearly the right speed. The second set of results, with a halving of the mesh size in both co-ordinate directions, shows the expected improvement in accuracy, though the results are still not very satisfactory.

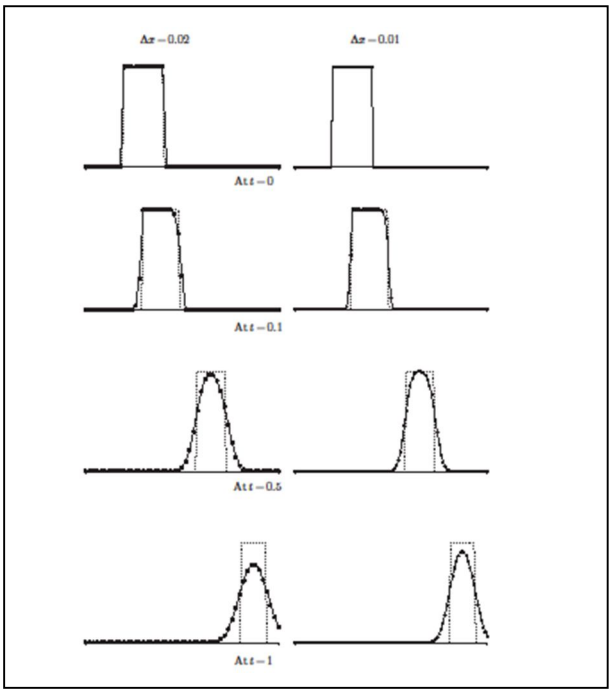

Figure (1.6): Linear advection by the upwind method: problem (1.36), (1.37).

1.3 The Lax−Wendroff scheme

The phase error of the upwind scheme is actually smaller than that of many higher order schemes: but the damping is very severe and quite unacceptable in most problems. One can generate more accurate explicit schemes by interpolating to higher order. We have seen how the upwind scheme can be derived by using linear interpolation to calculate an approximation to $u(Q)$ in Figure (1.5). A more accurate value may be found by quadratic interpolation, using the values at the three points A, B and C and assuming a straight characteristic with slope v. This gives the Lax−Wendroff scheme, which has turned out to be of central importance in the subject and was first used and studied by those authors in 1960 in their study of hyperbolic conservation laws; it takes the form

$$U_j^{n+1} = \frac{1}{2}v(1+v)U_{j-1}^n + (1-v^2)U_j^n - \frac{1}{2}v(1-v)U_{j+1}^n \qquad (1.39)$$

which may be written

$$U_j^{n+1} = U_j^n - v\Delta_{0x}U_j^n + \frac{1}{2}v^2\delta_x^2 U_j^n. \qquad (1.40)$$

The usual Fourier analysis gives the amplification factor

$$\lambda(k) = 1 - iv\sin k\Delta x - 2v^2\sin^2\frac{1}{2}k\Delta x. \qquad (1.41)$$

Separating the real and imaginary parts we obtain, after a little manipulation,

$$|\lambda|^2 = 1 - 4v^2(1-v^2)\sin^4\frac{1}{2}k\Delta x. \qquad (1.42)$$

Thus we see that the scheme is stable for $|v| \leq 1$, the whole range allowed by the CFL condition. We also find

$$\arg\lambda = -\tan^{-1}\left[\frac{v\sin k\Delta x}{1-2v^2\sin^2\frac{1}{2}\Delta x}\right]$$

$$\sim -v\xi\left[1-\frac{1}{6}(1-v^2)\xi^2 + \cdots\right]. \qquad (1.43)$$

Compared with the upwind scheme we see that there is still some damping, as in general $|\lambda| < 1$, but the amplitude error in one time step is now of order ξ^4 when ξ is small, compared with order ξ^2 for the upwind scheme; this is a substantial improvement. Both the schemes have a relative phase error of order ξ^2, which are equal when $v \sim 0$; but the error is always of one sign (corresponding to a phase lag) for Lax−Wendroff while it goes through a zero at $v = \frac{1}{2}$ for the upwind scheme. However, the much smaller damping of the Lax−Wendroff scheme often outweighs the disadvantage of the larger phase error.

In deriving the Lax−Wendroff scheme above we assumed a was constant. To deal with variable a in the linear equation (1.1) we derive it in a different way, following the original derivation. We first expand in a Taylor series in the variable t, giving

$$u(x,t+\Delta t) = u(x,t) + \Delta t u_t(x,t) + \frac{1}{2}(\Delta t)^2 u_{tt}(x,t) + O((\Delta t)^3). \qquad (1.44)$$

Then we convert the t-derivatives into x-derivatives by using the differential equation, so that

$$u_t = -au_x, \qquad (1.45)$$

$$u_{tt} = -a_t u_x - au_{xt}, \qquad (1.46)$$

$$u_{xt} = u_{tx} = -(au_x)_x, \qquad (1.47)$$

which give

$$u_{tt} = -a_t u_x + a(au_x)_x. \qquad (1.48)$$

Approximating each of these x-derivatives by central differences gives the scheme

$$U_j^{n+1} = U_j^n - a_j^n \Delta t \frac{\Delta_{0x} U_j^n}{\Delta x} + \frac{1}{2}(\Delta t)^2 \left[-(a_t)_j^n \frac{\Delta_{0x} U_j^n}{\Delta x} + a_j^n \frac{\delta_x \left(a_j^n \delta_x U_j^n \right)}{(\Delta x)^2} \right]. \qquad (1.49)$$

This scheme involves evaluating the function $a(x,t)$ at the points $x = x_j \pm \frac{1}{2}\Delta x$ as well as a and at $a_t at\ x_j$. Note, however, that the scheme can be simplified by replacing an $a_j^n + \frac{1}{2}\Delta t (a_t)_j^n$ by $a_j^{n+1/2}$ in the coefficient of $\Delta_{0x} U_j^n$; see also the next section for conservation laws with $au_x \equiv f_x$, and also the following section on finite volume schemes.

The results in Figure (1.7) are obtained by applying this scheme to the same problem (1.36), (1.37) used to test the upwind scheme, with the same mesh sizes[6].Comparing the results of Figure (1.6) and Figure (1.7) we see that the Lax−Wendroff scheme maintains the height and width of the pulse rather better than the upwind scheme, which spreads it out much more. On the other hand, the Lax−Wendroff scheme produces oscillations which follow behind the two discontinuities as the pulse moves to the right. Notice also that the reduction in the mesh size Δx does

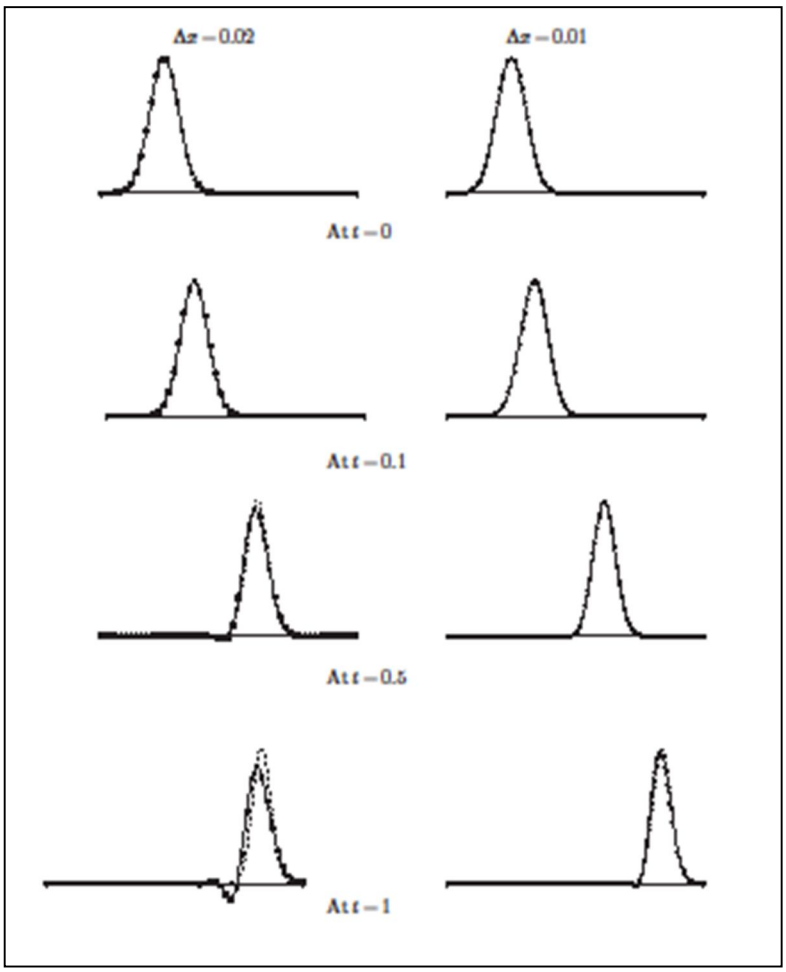

Figure (1.7): Linear advection by the Lax−Wendroff method: problem (1.36), (1.37).

improve the accuracy of the result, but not by anything like the factor of 4 which would be expected of a scheme for which the error is $O((\Delta x)^2)$. The analysis of truncation error is only valid for solutions which are sufficiently smooth, while this problem has a discontinuous solution. In fact the maximum error in this problem is $O\left((\Delta x)^{1/2}\right)$ for the upwind scheme and $O\left((\Delta x)^{2/3}\right)$ for the Lax−Wendroff scheme. The error therefore tends to zero rather slowly as the mesh size is reduced.

The oscillations in Figure (1.7) arise because the Lax−Wendroff scheme does not satisfy a maximum principle. We see from (1.39) that with $v > 0$ the coefficient of U_{j+1}^n is negative, since we require that $v \leq 1$ for stability. Hence U_j^{n+1} is given as a weighted mean of three values on the previous

time level, but two of the weights are positive and one is negative. It is therefore possible for the numerical solution to have oscillations with internal maxima and minima.

As an example of a problem with a smooth solution, we consider the same equation as before, (1.36a,b), but replace the initial condition (1.37) by

$$u(x,0) = \exp[-10(4x-1)^2]. \qquad (1.50)$$

The results are illustrated in Figure (1.8). As before, the solution consists of a pulse moving to the right, but now the pulse has a smooth Gaussian shape, instead of a discontinuous square wave. Using the same mesh sizes as before, the results are considerably more accurate. There is still some sign of an oscillation to the left of the pulse by the time that $t = 1$, but it is a good deal smaller than in the discontinuous case. Moreover, the use of the smaller mesh size has reduced the size of the errors and this oscillation becomes nearly invisible.

The Lax−Wendroff method for conservation laws

In practical situations a hyperbolic equation often appears in the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \qquad (1.51)$$

which may be written in the form we have considered above,

$$u_t + a u_x = 0, \qquad (1.52)$$

where $a = a(u) = \partial f/\partial u$. It is then convenient to derive the Lax−Wendroff scheme directly for the conservation form (1.51). The function f does not involve x or t explicitly but is a function of u only. The t-derivatives required in the Taylor series expansion (1.44) can now be written

$$u_t = -\big(f(u)\big)_x \qquad (1.53)$$

and

$$u_{tt} = -f_{xt} = -f_{tx} = -(au_t)_x = (af_x)_x. \qquad (1.54)$$

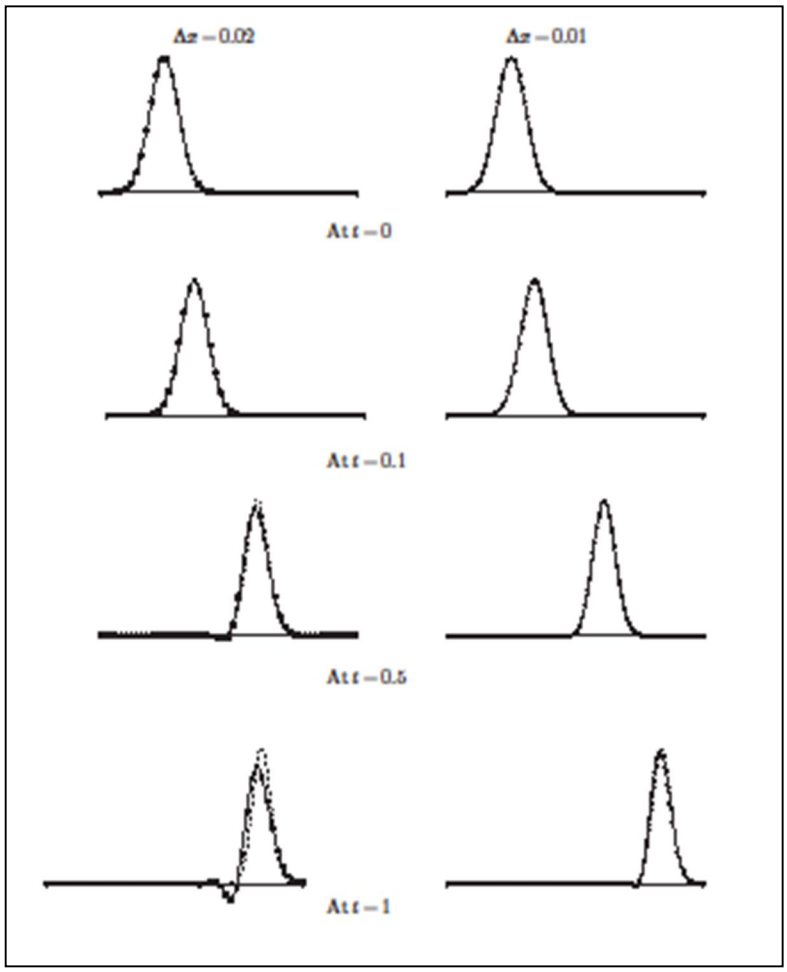

Figure (1.8): Linear advection by the Lax−Wendroff method: (1.36) with the data (1.50).

Replacing the x-derivatives by central differences as before we now obtain

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}\Delta_{0x} f\big(U_j^n\big) + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 \delta_x\big[a\big(U_j^n\big)\delta_x f\big(U_j^n\big)\big]. \qquad (1.55)$$

It is clear that this reduces to (1.40) when $f(u) = au$ where a is constant. If we expand the last term in (1.55) we see that it involves the values of

$a\left(U^n_{j-1/2}\right)$ and $a\left(U^n_{j+1/2}\right)$; in evaluating these we could set $U^n_{j\pm1/2} := \frac{1}{2}\left(U^n_j + U^n_{j\pm1}\right)$, but a commonly used alternative is to replace them by $\Delta_{\pm x} f\left(U^n_j\right)/\Delta_{\pm x} U^n_j$. Then writing F^n_j for $f\left(U^n_j\right)$ and $A^n_{j\pm1/2}$ for either choice of the characteristic speeds, the scheme becomes

$$U^{n+1}_j = U^n_j - \frac{1}{2}\frac{\Delta t}{\Delta x}\left\{\left[1 - A^n_{j+1/2}\frac{\Delta t}{\Delta x}\right]\Delta_{+x}F^n_j + \left[1 + A^n_{j-1/2}\frac{\Delta t}{\Delta x}\right]\Delta_{-x}F^n_j\right\}. \qquad (1.56)$$

As an example of the use of this scheme we consider the limiting case of Burgers' equation, for inviscid flow,

$$u_t + uu_x = 0, \qquad (1.57)$$

or in conservation form

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0. \qquad (1.58)$$

The general solution when it is smooth is easily obtained by the method of characteristics, or it is sufficient to verify that the solution is given implicitly by

$$u \equiv u(x,t) = u^0\left(x - tu(x,t)\right). \qquad (1.59)$$

The characteristics are straight lines, and the solution $u(x,t)$ is constant along each of them. Given the initial condition $u(x,0) = u^0(x)$, they are obtained by drawing the straight line with slope $\mathrm{d}t/\mathrm{d}x = 1/u^0(x_0)$ through the point $(x_0, 0)$, for each value of x_0. The approximation obtained with the upwind scheme, which we write in the form

$$U^{n+1}_j = U^n_j - \frac{1}{2}\frac{\Delta t}{\Delta x}\left\{\left[1 - \operatorname{sgn} A^n_{j+\frac{1}{2}}\right]\Delta_{+x}F^n_j + \left[1 + \operatorname{sgn} A^n_{j-\frac{1}{2}}\right]\Delta_{-x}F^n_j\right\} \qquad (1.60)$$

where the preferred choice is $A^n_{j\pm\frac{1}{2}} := \Delta_{\pm x}F^n_j/\Delta_{\pm x}U^n_j$, reducing to $a\left(U^n_j\right)$ when $U^n_j = U^n_{j\pm1}$; this form clearly generalises (1.23) and is directly comparable with (1.56).

Chapter Two

Finite Volume Schemes

2. 1 Introduction

Many of the methods that are used for practical computation with conservation laws are classed as finite volume method . Suppose we take the system of equations $\mathbf{u}_t + \mathbf{f}_x = 0$ in conservation law form and integrate over a region Ω in (x,t)-space; using the Gauss divergence theorem this becomes a line integral,

$$\int\int_{\Omega} (\boldsymbol{u}_t + \mathbf{f}_x)\, \mathrm{d}x\, \mathrm{d}t \equiv \int\int_{\Omega} \mathrm{div}(\mathbf{f}, \mathbf{u})\, \mathrm{d}x\, \mathrm{d}t$$

$$= \oint_{\partial\Omega} [\mathbf{f}\, \mathrm{d}t - \mathbf{u}\, \mathrm{d}x]. \qquad (2.1)$$

In particular, if we take the region to be a rectangle of width Δx and height Δt and introduce averages along the sides, such as $\mathbf{u}_{\mathrm{top}}$ etc., we obtain

$$(\mathbf{u}_{\mathrm{top}} - \mathbf{u}_{\mathrm{bottom}})\Delta x + (\mathbf{f}_{\mathrm{right}} - \mathbf{f}_{\mathrm{left}})\Delta t = 0. \qquad (2.2)$$

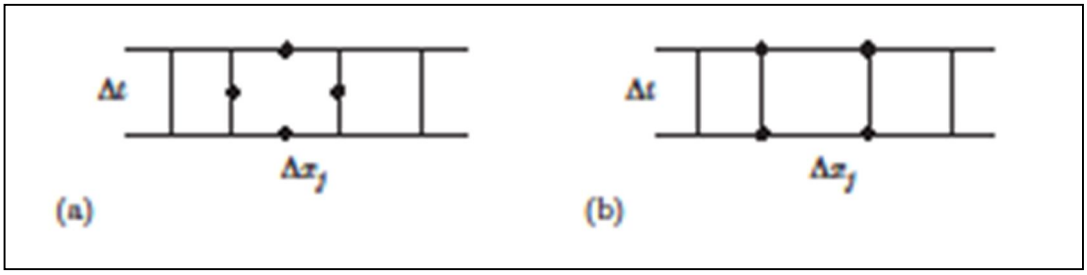

Figure (2.1): Two finite volume schemes: (a) with mid-point quadrature; (b) with trapezoidal quadrature.

Then to obtain a specific numerical scheme these averages need to be approximated by some form of quadrature. For instance, we can use mid-point quadrature on all four sides – see Figure (2.1) (a): if we denote by U_j^n the approximate solution at time level n at the centre of cell j of width Δx_j, and by $\mathbf{F}_{j+1/2}^{n+1/2}$ the flux value halfway up a cell side, we obtain the scheme

$$\mathbf{U}_j^{n+1} = U_j^n - (\Delta t/\Delta x)\left(\mathbf{F}_{j+1/2}^{n+1/2} - \mathbf{F}_{j-1/2}^{n+1/2}\right). \qquad (2.3)$$

It remains to calculate the fluxes from the set of U_j^n values.

Note, however, that in (2.3) we have allowed for the cell widths to be quite arbitrary. This is a great advantage of this formulation, and is very useful in practical calculations – even more so in more space dimensions. Thus, for instance, we can sum the integrals over a set of contiguous cells to obtain from (2.3)

$$\sum_{j=k}^{l} \Delta x_j \left(\mathbf{U}_j^{n+1} - U_j^n\right) + \Delta t\left(\mathbf{F}_{l+1/2}^{n+1/2} - \mathbf{F}_{k-1/2}^{n+1/2}\right) = 0, \qquad (2.4)$$

which exactly mirrors the conservation property of the differential equation. In the case of the Lax–Wendroff scheme, though, if U_j^n is taken to represent the solution at the cell centre then we need to use a Taylor expansion at a cell edge $x_{j+1/2}$ to give, to the required first order accuracy,

$$\mathbf{u}\left(x_{j+1/2}, t_n + \Delta t/2\right) = \mathbf{u}\left(x_{j+1/2}, t_n\right) - \frac{1}{2}\Delta t \mathbf{f}_x\left(x_{j+1/2}, t_n\right) + O((\Delta t)^2);$$

this can be combined with expansions for the cell centre values on either side to give the formula

$$\mathbf{U}_{j+1/2}^{n+1/2} = \frac{\Delta x_{j+1} U_j^n + \Delta x_j \mathbf{U}_{j1}^n - \Delta t\left[\mathbf{f}\left(\mathbf{U}_{j+1}^n\right) - \mathbf{f}\left(\mathbf{U}_j^n\right)\right]}{\Delta x_j + \Delta x_{j+1}}. \qquad (2.5)$$

As we have already noted and demonstrated, a major disadvantage of the Lax–Wendroff method is its proneness to produce oscillatory solutions. The problem has prompted much of the development of finite volume methods, and can be fully analysed for scalar conservation laws. The guiding principle is provided by controlling the total variation of the solution: on a finite domain $[0, X]$ divided into J cells, with U_j^n taking the value U_j^n in cell j at time level n, we can define the total variation as

$$\mathrm{TV}(U^n) := \sum_{j=1}^{J-1} \left| U_{j+1}^n - U_j^n \right| \equiv \sum_{j=1}^{J-1} \left| \Delta_{+x} U_j^n \right|. \qquad (2.6)$$

More generally, for the exact solution $u(x,t)$, $\mathrm{TV}\big(u(\cdot,t)\big)$ can be defined by taking the supremum, over all subdivisions of the $[0,X]$ interval such as $0 = \xi_0 < \xi_1 < \cdots < \xi_K = X$, of the sum of the corresponding differences $\left|u(\xi_{j+1},t) - u(\xi_j,t)\right|$. Clearly, these are consistent definitions when U^n is regarded as a piecewise constant approximation to $u(\cdot,t_n)$. To simplify the subsequent discussion, however, by leaving side the specification of boundary conditions, we will assume that both $u(\cdot,t)$ and U^n are extended by constant values to the left and right so that the range of the summation over j will not be specified.

2.2 Harten concept:

A key property of the solution of a conservation law such as (1.51) is that $\mathrm{TV}\big(u(\cdot,t)\big)$ is a nonincreasing function of $t-$ which can be deduced informally from the constancy of the solution along the characteristics described by (1.6). Thus we define TVD (total variation diminishing) schemes as those for which we have $\mathrm{TV}(U^{n+1}) \leq \mathrm{TV}(U^n)$. This concept is due to Harten who established the following useful result:

Theorem (1.1): (Harten)

A scheme is TVD if it can be written in the form

$$U_j^{n+1} = U_j^n - C_{j-1}\Delta_{-x}U_j^n + D_j\Delta_{+x}U_j^n, \qquad (2.7)$$

where the coefficients C_j and D_j, which may be any functions of the solution variables $\{U_j^n\}$, satisfy the conditions

$$C_j \geq 0, \quad D_j \geq 0 \quad \text{and} \quad C_j + D_j \leq 1 \ \ \forall j. \qquad (2.8)$$

Proof:

Taking the forward difference of (2.7), and freely using the identity $\Delta_{+x}U_j \equiv \Delta_{-x}U_{j+1}$, we get

$$U_{j+1}^{n+1} - U_j^{n+1} = \Delta_{+x}U_j^n - C_j\Delta_{+x}U_j^n + C_{j-1}\Delta_{-x}U_j^n + D_{j+1}\Delta_{+x}U_{j+1}^n - D_j\Delta_{+x}U_j^n$$

$$= (1 - C_j - D_j)\Delta_{+x}U_j^n + C_{j-1}\Delta_{-x}U_j^n + D_{j+1}\Delta_{+x}U_{j+1}^n.$$

By the hypotheses of (2.8), all the coefficients on the right of this last expression are nonnegative. So we can take absolute values to obtain

$$\left|\Delta_{+x}U_j^{n+1}\right| \le (1 - C_j - D_j)\left|\Delta_{+x}U_j^n\right| + C_{j-1}\left|\Delta_{-x}U_j^n\right| + D_{j+1}\left|\Delta_{+x}U_{j+1}^n\right|,$$

then summing over j leads to cancellation and hence the result $\mathrm{TV}(U^{n+1}) \le \mathrm{TV}(U^n)$.

Suppose we attempt to apply this theorem to both the Lax−Wendroff method and the upwind method. We consider the latter first, in the form given in (1.60) with An $A_{j\pm1/2}^n := \Delta_{\pm x} F_j^n / \Delta_{\pm x} U_j^n$. This corresponds to the scalar case of the scheme is best considered as a finite volume scheme in which the fluxes of (2.3) are given by

$$F_{j+1/2}^{n+1/2} = \begin{cases} f(U_j^n) & \text{when} \quad A_{j+1/2}^n \ge 0, \\ f(U_{j+1}^n) & \text{when} \quad A_{j+1/2}^n < 0; \end{cases} \qquad (2.9)$$

or, equivalently,

$$F_{j+1/2}^{n+1/2} = \frac{1}{2}\left[(1 + \operatorname{sgn} A_{j+1/2}^n)F_j^n + (1 - \operatorname{sgn} A_{j+1/2}^n)F_{j+1}^n\right]. \qquad (2.10)$$

Then, comparing (1.60) with (2.7) after replacing the flux difference $\Delta_{-x}F_j^n$ by $A_{j-1/2}^n\Delta_{-x}U_j^n$, we are led to setting

$$C_{j-1} = \frac{1}{2}\frac{\Delta t}{\Delta x}(1 + \operatorname{sgn} A_{j+1/2}^n)\, A_{j-1/2}^n.$$

This is clearly always nonnegative, thus satisfying the first condition of (2.8). Similarly, we set

$$D_j = \frac{1}{2}\frac{\Delta t}{\Delta x}\left(1 - \operatorname{sgn} A^n_{j+1/2}\right) A^n_{j+1/2},$$

which is also nonnegative. Moreover, adding the two together and remembering the shift of subscript in the former, we get

$$C_j + D_j = \frac{1}{2}\frac{\Delta t}{\Delta x}\left[\left(1 + \operatorname{sgn} A^n_{j+1/2}\right) A^n_{j-1/2} + \frac{\Delta t}{\Delta x}\left(1 - \operatorname{sgn} A^n_{j+1/2}\right) A^n_{j+1/2}\right]$$

$$\equiv \left|A^n_{j+1/2}\right| \frac{\Delta t}{\Delta x},$$

which is just the CFL number. Hence the last condition of (2.8) corresponds to the CFL stability condition; the Roe first order upwind scheme is TVD when Δt is chosen so that it is stable[7].

On the other hand, if we attempt to follow similar arguments with the Lax–Wendroff scheme in the corresponding form of (1.56) and write $U^n_{j\pm1/2}$ for $A^n_{j\pm1/2}\,\Delta t/\Delta x$, we are led to setting

$$C_j = \frac{1}{2} v^n_{j+1/2}\left(1 + v^n_{j+1/2}\right), \quad \text{and} \quad D_j = -\frac{1}{2} v^n_{j+1/2}\left(1 - v^n_{j+1/2}\right), \qquad (2.11)$$

both of which have to be nonnegative. Then the third condition of (1.8) requires that the CFL condition $\left(v^n_{j+1/2}\right)^2 \le 1$ be satisfied, and the only values that $v^n_{j+1/2}$ can take to satisfy all three conditions are $-1, 0$ and $+1$; this is clearly impractical for anything other than very special cases.

The TVD property of the Roe upwind scheme has made it a very important building block in the development of more sophisticated finite volume methods.

However, these two schemes are only first order accurate and it is no easy matter to devise TVD schemes that are second order accurate. To consider why this is so let us consider an explicit TVD three-point scheme in the form (2.7) and satisfying the conditions (2.8). For the linear advection equation $u_t + au_x = 0$ we suppose that C and D are constants. Then it is easy to see, following the argument that led to the Lax–Wendroff method

in (1.39), that second order accuracy leads directly to these coefficients, as in (2.11), and hence the violation of the TVD conditions except in very special cases. From another viewpoint, in the two successful TVD schemes we have constructed the fluxes from just the cell average values U_j^n in each cell, and we cannot expect to approximate the solution to second order accuracy with a piecewise constant approximation.

This observation points the way to resolving the situation: an intermediate stage, variously called recovery or reconstruction, is introduced to generate a higher order approximation. $\tilde{U}^n(\cdot)$ to $u(\cdot, t_n)$ from the cell averages $\{U_j^n\}$. Probably the best known approach is that used by van Leer[4]to produce his MUSCL schemes (Monotone Upstream-centred Schemes for Conservation Laws). This uses discontinuous piecewise linear approximations to generate second order approximations. Another well-established procedure leads to the Piecewise Parabolic Method (PPM) scheme of Colella and Woodward, [2]which can be third order accurate. In all cases the recovery is designed to preserve the cell averages. So for the recovery procedure used in the MUSCL schemes, for each cell we need only calculate a slope to give a straight line through the cell average value at the centre of the cell, and this is done from the averages in neighbouring cells. The PPM, however, uses a continuous approximation based on cell-interface values derived from neighbouring cell averages: so a parabola is generated in each cell from two interface values and the cell average.

2. 3 The box scheme

To give some indication of the range of schemes that are used in practice we will describe two other very important schemes. The box scheme is a very compact implicit scheme often associated with the names of

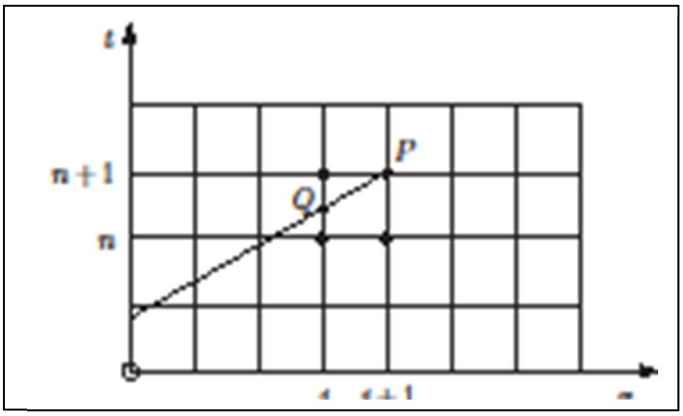

Figure (2.2): The box scheme.

Thomee and, in a different context, Keller: for the simplest model problem $u_t + au_x = 0$ with constant a it takes the form

$$\frac{\delta_t\left(U_j^{n+1/2} + U_{j+1}^{n+1/2}\right)}{2\Delta t} + \frac{a\delta_x\left(U_{j+1/2}^{n} + U_{j+1/2}^{n+1}\right)}{2\Delta x} = 0. \qquad (2.12)$$

By introducing the averaging operator

$$\mu_x U_{j+1/2} := \frac{1}{2}\left(U_j + U_{j+1}\right), \qquad (2.13)$$

and similarly μ_t we can write the scheme in the very compact form

$$\left(\mu_x\delta_t + v\mu_t\delta_x\right)U_{j+1/2}^{n+1/2} = 0, \qquad (2.14)$$

where $v = a\,\Delta t/\Delta x$ is the CFL number.

If we expand all the terms in Taylor series about the central point $\left(x_{j+1/2}, t_{j+1/2}\right)$ as origin, it is easy to see that the symmetry of the averaged differences will give an expansion in even powers of Δx or Δt, so that the scheme is second order accurate. When the coefficient a is a function of x and t, it is sensible to replace it by $a_{j+1/2}^{n+1/2} := a\left(x_{j+1/2}, t_{n+1/2}\right)$ in (2.12); this will leave the Taylor series unaltered, so that the truncation error remains second order.

When applied to the nonlinear problem in conservation form (1.51), it is written

$$\frac{\delta_t\left(U_j^{n+1/2} + U_{j+1}^{n+1/2}\right)}{2\Delta t} + \frac{\delta_x\left(F_{j+1/2}^{n} + F_{j+1/2}^{n+1}\right)}{2\Delta x} = 0 \qquad (2.15)$$

where $F_j^n := f\left(U_j^n\right)$. In this form it is clear that it corresponds to a finite volume scheme, using a box formed from four neighbouring mesh nodes in a square, and applying the trapezoidal rule to evaluate the integral along each edge – see Figures (2.2) and (2.1) (b); and in common with other finite volume schemes it can be readily applied on a nonuniform mesh.

The scheme is implicit as it involves two points on the new time level, but for the simplest model problem this requires no extra computation; and when used properly it is unconditionally stable. We can write (2.12) in the form

$$U_{j+1}^{n+1} = U_j^n + \left(1 + v_{j+1/2}^{n+1/2}\right)^{-1}\left(1 - v_{j+1/2}^{n+1/2}\right)\left(U_{j+1}^n - U_j^{n+1}\right) \qquad (2.16)$$

where

$$v_{j+1/2}^{n+1/2} = a_{j+1/2}^{n+1/2}\frac{\Delta t}{\Delta x}.$$

When $a(x,t)$ is positive, so that the characteristic speed is positive, we must be given a boundary condition on the left of the region. This will define U_0^{n+1}, the first value of U on the new time level, and (2.16) will give directly the values of U in succession from left to right. If the speed is negative, and we are given a boundary condition on the right, a similar formula is used in the direction from right to left. When the equation is nonlinear and in conservation form the scheme is not quite so easy to use, as (2.15) now represents a nonlinear equation which must be solved for U_{j+1}^{n+1}.

One of the serious difficulties in using the scheme is the possibility of a chequerboard mode of the form $(-1)^{j+n}$ contaminating the solution. Because of the averaging in both space and time in equation (2.12) this is a spurious solution mode of this equation. It is only the boundary condition

and the initial condition that control its presence; as we see in Figure (2.3), it is much more evident with square pulse data than with smooth data.

For a system of equations, such as (1.5), the calculation becomes more elaborate, as the scheme is now truly implicit. We will have a set of simultaneous equations to solve on the new time level, and in general it will not be possible to solve them by a simple sweep in one direction, as it was for a single equation, since there will normally be some boundary conditions given at each end of the range of x. The matrix of the system will be typically block tridiagonal, with the block dimension equal to the order of

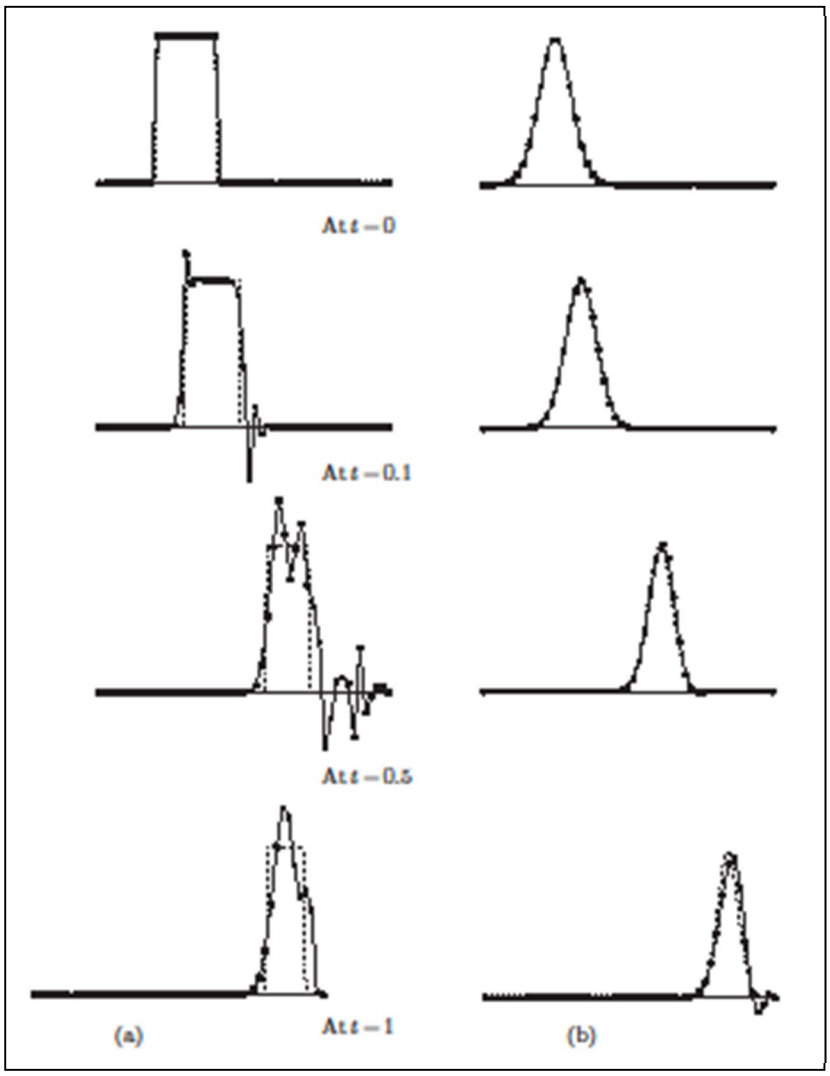

Figure (2.3): Linear advection by the box scheme with $\Delta t = \Delta x = 0.02$ for (a) a square pulse and (b) Gaussian initial data.

the system, but the detailed structure will depend on the number of boundary conditions imposed at each end.

For the scalar problem the CFL condition is satisfied for any value of the ratio $\Delta t/\Delta x$ if we use the scheme in the correct direction, that is, in the form (2.16) when a is positive; as we see from Figure (2.2), the characteristic always passes among the three points used in the construction of U_{j+1}^{n+1}. However, the three coefficients are not all positive, so there is no maximum principle; nor does the scheme have any natural TVD properties, being prone as we have seen to oscillatory solutions. Because we have neither the whole real line as our domain nor periodic boundary conditions, a rigorous Fourier analysis is not straightforward even for constant a. However, we can substitute a Fourier mode to consider its possible damping and its phase accuracy. We easily find

$$\lambda(k) = \frac{\cos\frac{1}{2}k\Delta x - iv\sin\frac{1}{2}k\Delta x}{\cos\frac{1}{2}k\Delta x + iv\sin\frac{1}{2}k\Delta x}, \qquad (2.17)$$

from which we deduce

$$|\lambda(k)| = 1 \qquad (2.18)$$

for any value of v, and

$$\arg\lambda = -2\tan^{-1}\left(v\tan\frac{1}{2}k\Delta x\right)$$

$$\sim -v\xi\left[1 + \frac{1}{12}(1-v^2)\xi^2 + \cdots\right]. \qquad (2.19)$$

Thus the scheme has no damping of modes and, comparing (2.19) with (1.43), we see that it has the same second order accuracy and that its phase error is asymptotically half that of the Lax−Wendroff scheme .

2. 4The leap-frog scheme

The second important scheme is called the leap-frog scheme because it uses two time intervals to get a central time difference and spreads its ' legs '

to pick up the space difference at the intermediate time level; the values used are shown in Figure (2.4). For (1.51) or (1.52) it has the form

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} + \frac{f\left(U_{j+1}^n\right) - f\left(U_{j-1}^n\right)}{2\Delta x} = 0, \qquad (2.20)$$

or

$$U_j^{n+1} = U_j^{n-1} - (a\,\Delta t/\Delta x)\left[U_{j+1}^n - U_{j-1}^n\right]. \qquad (2.21)$$

Thus it is an explicit scheme that needs a special technique to get it started. The initial condition will usually determine the values of U^0, but a special procedure is needed to give U^1. Then the leap-frog scheme can be used to give $U^2, U^3, \ldots$ in succession. The additional starting values U^1 can be obtained by any convenient one-step scheme, such as Lax−Wendroff.

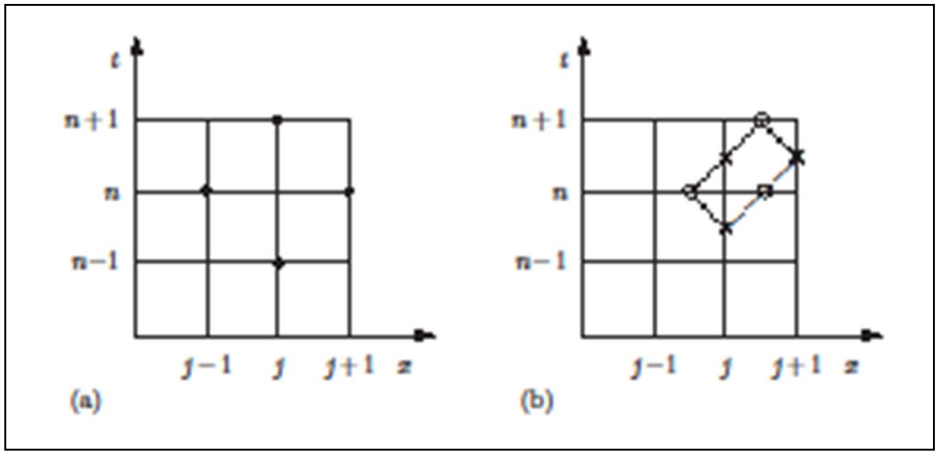

Figure (2.4): The leap-frog scheme: (a) unstaggered; (b) staggered, × = V and o = W.

It is clear from Figure (2.4) (a) that the CFL condition requires that $|v| \leq 1$, as for the Lax−Wendroff scheme. When $f = au$ with constant a the usual Fourier analysis leads to a quadratic for $\lambda(k)$:

$$\lambda^2 - 1 + 2iv\lambda \sin k\,\Delta x = 0 \qquad (2.22)$$

with solutions

$$\lambda(k) = -iv \sin k\Delta x \pm [1 - v^2 \sin^2 k\Delta x]^{1/2}. \qquad (2.23)$$

Since the product of these roots is -1, we must require both roots to have modulus 1 for the scheme to be stable. It is easy to verify that the roots are complex and equal in modulus for all k if and only if $|v| \leq 1$: so for this scheme the Fourier analysis gives the same result as the CFL condition; and when the stability condition is satisfied there is no damping.

The result of the Fourier analysis leading to two values of $\lambda(k)$ is a serious problem for this scheme, as it means that it has a spurious solution mode. It arises from the fact that the scheme involves three time levels and so needs extra initial data, and it is this that determines the strength of this mode. Taking the positive root in (2.23) we obtain a mode that provides a good approximation to the differential equation, namely the ' true ' mode λ_T given by

$$\arg \lambda_T = -\sin^{-1}(v \sin k\,\Delta x)$$

$$\sim -v\xi\left[1 - \frac{1}{6}(1 - v^2)\xi^2 + \cdots\right]. \qquad (2.24)$$

Note that the phase error here has the same leading term as the Lax–Wendroff scheme – see (1.43). On the other hand, taking the negative root gives the spurious mode

$$\lambda_S \sim (-1)\left[1 + iv\xi - \frac{1}{2}v^2\xi^2 + \cdots\right], \qquad (2.25)$$

which gives a mode oscillating from time step to time step and travelling in the wrong direction. In practical applications, then, great care has to be taken not to stimulate this mode, or in some circumstances it may have to be filtered out.

The results displayed in Figure (2.5) illustrate the application of the leapfrog method for a square pulse and for Gaussian initial data; the first time step used the Lax–Wendroff scheme. The results clearly show the oscillating wave moving to the left. In some respects the results are similar to those for the box scheme; but the oscillations move at a speed independent of the mesh and cannot be damped, so in this case they have to be countered by some form of filtering.

The real advantage of the leap-frog method occurs when it is applied to a pair of first order equations such as those derived from the familiar second order wave equation

$$u_{tt} = a^2 u_{xx}, \tag{2.26}$$

where a is a constant: if we introduce variables $v = u_t$ and $w = -au_x$, it is clear that they satisfy the system

$$\begin{aligned} v_t + aw_x &= 0, \\ w_t + av_x &= 0. \end{aligned} \tag{2.27}$$

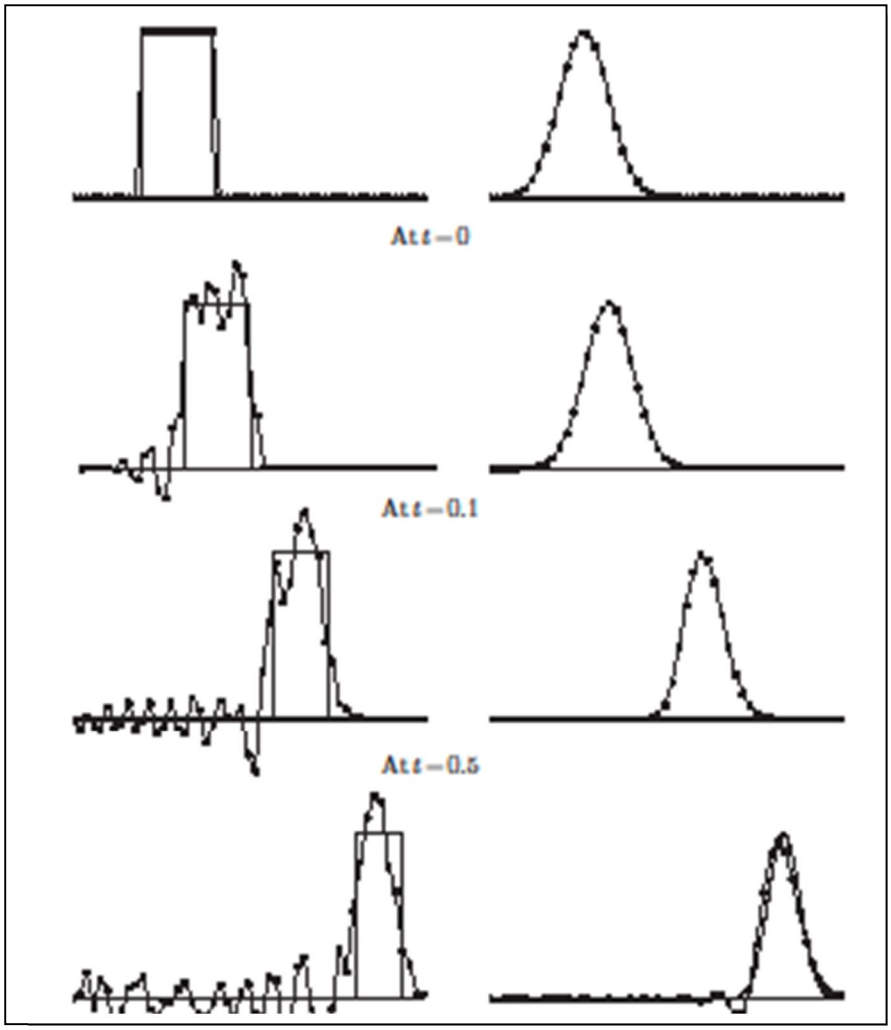

Figure (2.5): Linear advection by the leap-frog scheme with $\Delta t - \Delta x - 0.02$ for (a) a square pulse and (b) Gaussian initial data.

Because of the pattern of differentials here, a staggered form of the leapfrog method can be used that is much more compact than (2.21): as indicated in Figure (2.4) (b) we have V and W at different points and a staggered scheme can be written

$$\frac{V_j^{n+1/2} - V_j^{n-1/2}}{\Delta t} + a\frac{W_{j+1/2}^n - W_{j-1/2}^n}{\Delta x} = 0, \qquad (2.28)$$

$$\frac{W_{j+1/2}^{n+1} - W_{j+1/2}^n}{\Delta t} + a\frac{V_{j+1}^{n+1/2} - V_j^{n+1/2}}{\Delta x} = 0, \qquad (2.29)$$

or

$$\delta_t V + v\delta_x W = 0, \quad \delta_t W + v\delta_x V = 0, \qquad (2.30)$$

where we have taken advantage of the notation to omit the common superscripts and subscripts. With constant a, we can construct a Fourier mode by writing

$$\left(V^{n-1/2}, W^n\right) = \lambda^n \mathrm{e}^{ikx}\left(\hat{V}, \hat{W}\right) \qquad (2.31)$$

where $\hat{V}$ and $\hat{W}$ are constants. These will satisfy the equations (2.28) , (2.29) if

$$\begin{pmatrix} \lambda - 1 & 2iv\sin\frac{1}{2}k\Delta x \\ 2i\lambda v\sin\frac{1}{2}k\,\Delta x & \lambda - 1 \end{pmatrix}\begin{pmatrix} \hat{V} \\ \hat{W} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \qquad (2.32)$$

This requires the matrix in (2.32) to be singular, so that

$$\lambda^2 - 2\left(\mathbf{1} - 2v^2\sin^2\frac{1}{2}k\,\Delta x\right)\lambda + \mathbf{1} = 0 \qquad (2.33)$$

with solutions given by

$$\lambda_{\pm}(k) = 1 - 2v^2s^2 \pm 2ivs[\mathbf{1} - v^2s^2]^{1/2}, \qquad (2.34)$$

where $s = \sin\frac{1}{2}k\,\Delta x$. Again, for the scheme to be stable we need λ_+, λ_- to be a complex conjugate pair so that stability requires $|v| \leq 1$, in which case $|\lambda_\pm| = 1$. The phases are given by

$$\arg\lambda_\pm = \pm\sin^{-1}\left(2vs[1 - v^2s^2]^{1/2}\right)$$

$$\sim \pm v\xi\left[1 - \frac{1}{24}(1 - v^2)\xi^2 + \cdots\right]. \qquad (2.35)$$

Note that the two roots of (2.33) are just the squares of the roots of (2.22) with Δx replaced by $\frac{1}{2}\Delta x$; hence the expansion in (2.35) corresponds to that in (2.24) with ξ replaced by $\frac{1}{2}\xi$. Both modes are now true modes which move to left and right at equal speeds, correctly approximating the behaviour of solutions to the wave equation. Note too that the accuracy is now better than that of the box scheme.

Substituting

$$V_j^{n+1/2} = \left(U_j^{n+1} - U_j^n\right)/\Delta t, \quad W_{j+1/2}^n = -a\left(U_{j+1}^n - U_j^n\right)/\Delta x$$

into the equations (2.28), (2.29) (2.30) gives

$$(\delta_t^2 - v^2\delta_x^2)U_j^n = 0, \qquad (2.36)$$

the simplest central difference representation of the second order wave equation (2.26) for U, together with a consistency relation. Note, too, that if we eliminate either V or W from the equations we find that both satisfy this second order equation. Some of the very attractive properties of this scheme will be derived, and put in a wider context, in the next section.

2.5 Hamiltonian systems and symplectic integration schemes

There are two important structural properties that lie behind the attractive features of the staggered leap-frog scheme applied to the wave equation: first, the wave equation (2.26) is the simplest example of a Hamiltonian PDE; and secondly, the staggered leap-frog scheme is one of the most common examples of a symplectic integration scheme. The importance of

combining these two ideas has been most fully worked out over the last several years in the approximation of ordinary differential equation systems; but as they have recently been introduced into the area of PDEs we shall here outline what is involved, by means of the staggered leap-frog example. We will also show that the box scheme can share some of these properties. In doing so we shall mainly use the terminology and notation of Leimkuhler and Reich (2004)[5].

Hamiltonian systems of ODEs have their origins in Hamilton's 1834 formulation of the equations of motion for a dynamical system, but have since been much generalised and their key properties widely studied. Let $\boldsymbol{q} \in \mathbb{R}^d$ and $\boldsymbol{p} \in \mathbb{R}^d$ be ' position' and ' momentum ' variables, which together we will denote by $\boldsymbol{z}$, and $\mathcal{H}(\boldsymbol{p}, \boldsymbol{q}) \equiv \mathcal{H}(\boldsymbol{z}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, a smooth Hamiltonian function that defines the ODE system

$$\dot{\mathbf{z}} \equiv \begin{pmatrix} \dot{\mathbf{q}} \\ \dot{\mathbf{p}} \end{pmatrix} = \boldsymbol{J} \begin{pmatrix} \mathcal{H}_{\mathrm{q}} \\ \mathcal{H}_{\mathrm{p}} \end{pmatrix} \equiv \boldsymbol{J} \nabla_{\mathbf{z}} \mathcal{H}, \qquad (2.37)$$

where the canonical structure matrix **J** has the form

$$\boldsymbol{J} = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix} \text{ with } \boldsymbol{J}^{-1} = \begin{pmatrix} 0 & -I_d \\ I_d & 0 \end{pmatrix},$$

in which I_d is the d-dimensional identity matrix. It is clear that $\mathcal{H}$ is constant along any trajectory; its value represents the energy of the system, which we shall sometimes denote by $E(\mathbf{z})$. Indeed, consider an arbitrary function $\mathcal{G} : \mathbb{R}^{2d} \to \mathbb{R}$, for which we will have, along any trajectory

$$\frac{\mathrm{d}\mathcal{G}(\mathbf{z})}{\mathrm{d}t} = (\nabla_{\mathbf{z}}\mathcal{G})^T \dot{\mathbf{z}} = (\nabla_{\mathbf{z}}\mathcal{G})^T \boldsymbol{J} \nabla_{\mathbf{z}} \mathcal{H} =: \{\mathcal{G}, \mathcal{H}\}. \qquad (2.38)$$

The expression $\{\mathcal{G}, \mathcal{H}\}$ is called the Poisson bracket of $\mathcal{G}$ and $\mathcal{H}$. It is clearly antisymmetric, and hence zero when $\mathcal{G} = \mathcal{H}$: and whenever it is identically zero the quantity $\mathcal{G}(\boldsymbol{q}, \boldsymbol{p})$ is constant along the trajectory.

Then $\mathcal{G}$ is called a constant of the motion, with the energy being such a constant for any Hamiltonian system. The best-known example of a Hamiltonian system is the simple plane pendulum, in which $d = 1$ and

$\mathcal{H} = \frac{1}{2}p^2 - (g/L)\cos q$. The trajectories are given by $\dot{q} = p, \dot{p} = -(g/L)\sin q$, and from $\mathcal{H}(q,p) = \text{const.}$ along each, it is easy to deduce that in the (q,p) −phase plane they form the familiar closed curves around centres at $p = 0,\ q = 2m\pi$ separated by saddle points at $p = 0,\ q = (2m+1)\pi$.

Of even greater significance than the existence of constants of the motion are the structural properties of the flow map formed from a set of trajectories of a Hamiltonian system: for example, in the scalar $d = 1$ case it is area-preserving; more generally it is said to be symplectic. To see what is involved in these ideas we need a few more definitions. A general mapping $\Psi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is said to be symplectic, with respect to the canonical structure matrix $\boldsymbol{J}$, if its Jacobian $\Psi_{\mathbf{z}}$ is such that

$$\Psi_{\mathbf{z}}^T \boldsymbol{J}^{-1} \Psi_z = \boldsymbol{J}^{-1} \tag{2.39}$$

In the scalar case it is then easy to calculate that

$$\text{if } \Psi_{\mathbf{z}} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ then } \Psi_{\mathbf{z}}^T \boldsymbol{J}^{-1} \Psi_{\mathbf{z}} = \begin{pmatrix} 0 & -ad + bc \\ ad - bc & 0 \end{pmatrix},$$

so that Ψ is simplectic iff $\det \Psi_{\mathbf{z}} \equiv ad - bc = 1$. Hence if this holds, and if $\mathbf{z} \in \Omega \subset \mathbb{R}^2$ is mapped into $\hat{\mathbf{z}} = \Psi(\mathbf{z}) \in \widehat{\Omega} \subset \mathbb{R}^2$, we have

$$\int_{\widehat{\Omega}} \mathrm{d}\hat{\mathbf{z}} = \int_{\Omega} \det \Psi_{\mathbf{z}}\, \mathrm{d}\mathbf{z} = \int_{\Omega} \mathrm{d}\mathbf{z},$$

i.e., the mapping is area-preserving. So the symplectic property generalizes the area preserving property to $d > 1$.

To apply this concept to the mapping produced by integrating a differential equation we define, in the language of differential geometry, the differential one-form of a function $f : \mathbb{R}^{2d} \to \mathbb{R}$, in the direction $\xi \in \mathbb{R}^{2d}$,

$$\mathrm{d}f(\boldsymbol{\xi}) := \nabla_{\mathbf{z}} f \cdot \boldsymbol{\xi} \equiv \sum_{i=1}^{2d} \frac{\partial f}{\partial z_i} \xi_i. \tag{2.40}$$

Then for two such functions, f and g, we can define a differential two form, called the wedge product, as

$$(\mathrm{d}f \wedge \mathrm{d}g)(\boldsymbol{\xi}, \boldsymbol{\eta}) := \mathrm{d}g(\boldsymbol{\xi})\,\mathrm{d}f(\boldsymbol{\eta}) - \mathrm{d}f(\boldsymbol{\xi})\,\mathrm{d}g(\boldsymbol{\eta}). \qquad (2.41)$$

In particular, we can apply (2.40) to the components z_i of $\mathbf{z} \equiv (\mathbf{q}, \mathbf{p})$ to obtain $\mathrm{d}z_i(\boldsymbol{\xi}) = \xi_i$ and write these as a vector $\mathrm{d}\mathbf{z} \equiv (\mathrm{d}\mathbf{q}, \mathrm{d}\mathbf{p})^T = (\mathrm{d}z_1, \mathrm{d}z_2, \dots, \mathrm{d}z_{2d})^T$. It is also easy to see that if we apply (2.40) to the components of the transformed variable $\hat{\mathbf{z}} = \Psi(\mathbf{z})$ we obtain

$$\mathrm{d}\hat{\mathbf{z}}(\boldsymbol{\xi}) = \Psi_{\mathbf{z}}\mathrm{d}\mathbf{z}(\boldsymbol{\xi}) \equiv \Psi_{\mathbf{z}}\boldsymbol{\xi}. \qquad (2.42)$$

Furthermore, we can apply (2.41) to these components and then define the wedge product

$$\mathrm{d}\mathbf{q} \wedge \mathrm{d}\mathbf{p} := \sum_{i=1}^{d} \mathrm{d}q_i \wedge \mathrm{d}p_i\,. \qquad (2.43)$$

It is the conservation of this quantity that turns out to be the key characterization of Hamiltonian systems.

First of all, with a calculation as in the scalar case, we see that

$$\begin{aligned}\boldsymbol{\xi}^T J^{-1}\boldsymbol{\eta} &= \big(\mathrm{d}\mathbf{q}^T(\boldsymbol{\xi}), \mathrm{d}\mathbf{p}^T(\boldsymbol{\xi})\big) J^{-1} \big(\mathrm{d}\mathbf{q}(\boldsymbol{\eta}), \mathrm{d}\mathbf{p}(\boldsymbol{\xi})\big)^T \\ &= \sum_{i=1}^{d} [\mathrm{d}p_i(\boldsymbol{\xi})\mathrm{d}q_i(\boldsymbol{\eta}) - \mathrm{d}q_i(\boldsymbol{\xi})\mathrm{d}p_i(\boldsymbol{\eta})] \\ &= \sum_{i=1}^{d} \mathrm{d}q_i \wedge \mathrm{d}p_i \equiv \mathrm{d}\mathbf{q} \wedge \mathrm{d}\mathbf{p}. \qquad (2.44)\end{aligned}$$

Then if we premultiply (2.39) by $\boldsymbol{\xi}^T$ and postmultiply by $\boldsymbol{\eta}$, and compare the result with the combination of (2.44) with (2.42), we deduce immediately that a mapping from $(\boldsymbol{q}, \boldsymbol{p})$ to $(\hat{\boldsymbol{q}}, \hat{\boldsymbol{p}})$ is symplectic iff

$$\mathrm{d}\hat{\mathbf{q}} \wedge \mathrm{d}\hat{\mathbf{p}} = \mathrm{d}\mathbf{q} \wedge \mathrm{d}\mathbf{p}. \qquad (2.45)$$

The fundamental result that the flow map of a Hamiltonian system is symplectic can be derived directly from (2.39), but (2.45) is crucially important in characterising the behaviour of the flow.

Numerical methods for approximating ODE systems that retain these properties are called symplectic integration schemes or, more generally, geometric integrators – see Hairer, Lubich and Wanner (2002)[3]. The simplest of these share the staggered structure of the leap-frog scheme. For simplicity we start with the scalar $d = 1$ case, where we alternate between the pair of equations

$$q^{n+1} = q^n + \Delta t \mathcal{H}_p\left(q^n, p^{n+1/2}\right)$$

$$p^{n+1/2} = p^{n-1/2} - \Delta t \mathcal{H}_q\left(q^n, p^{n+1/2}\right). \qquad (2.46)$$

If, as in the pendulum case, $\mathcal{H}_p$ depends only on p and $\mathcal{H}_q$ only on q this is an explicit method; more generally, it is implicit. In either case, if we take the differentials of these equations we obtain

$$\mathrm{d}q^{n+1} = \mathrm{d}q^n + \Delta t\left[\mathcal{H}_{pq}\mathrm{d}q^n + \mathcal{H}_{pp}\mathrm{d}p^{n+1/2}\right]$$

$$\mathrm{d}p^{n+1/2} = \mathrm{d}p^{n-1/2} - \Delta t\left[\mathcal{H}_{qq}\mathrm{d}q^n + \mathcal{H}_{qp}\mathrm{d}p^{n+1/2}\right], \qquad (2.47)$$

where we have omitted the arguments from the common Hamiltonian in (2.46). Now when we take the wedge product of these two equations, its antisymmetry implies that terms $\mathrm{d}q^{\boldsymbol{n}} \wedge \mathcal{H}_{qq}\mathrm{d}q^{\boldsymbol{n}}$ and $\mathcal{H}_{pp}\mathrm{d}p^{n+1/2} \wedge \mathrm{d}p^{n+1/2}$ are zero. So we take the wedge product of the first equation with $\mathrm{d}p^{n+1/2}$ and substitute from the second equation in the $\mathrm{d}q^n$ term to get, after omitting these null terms,

$$\mathrm{d}q^{n+1} \wedge \mathrm{d}p^{n+1/2} = \mathrm{d}q^n \wedge \left[\mathrm{d}p^{n-1/2} - \Delta t \mathcal{H}_{qp}\mathrm{d}p^{n+1/2}\right] + \Delta t \mathcal{H}_{pq}\mathrm{d}q^n \wedge \mathrm{d}p^{n+1/2}. \quad (2.48)$$

The two terms in Δt cancel and we have the discrete symplectic property

$$\mathrm{d}q^{n+1} \wedge \mathrm{d}p^{n+1/2} = \mathrm{d}q^n \wedge \mathrm{d}p^{n-1/2}. \qquad (2.49)$$

If the whole procedure is repeated for a system with $d > 1$ the same result is obtained: this is because, from the definitions of (2.41) and (2.43), it is easy to see that for any matrix A we have

$$\mathrm{d}\mathbf{a} \wedge (A\mathrm{d}\mathbf{b}) = (A^T \mathrm{d}\mathbf{a}) \wedge \mathrm{d}\mathbf{b},$$

so that if A is symmetric and $\boldsymbol{a} = \boldsymbol{b}$ the antisymmetry of the wedge product again implies that the result is zero.

In the ODE literature this staggered leap-frog method is usually referred to as the Störmer-Verlet method; and the commonly used Asymmetrical Euler methods differ from it only in their superscript labelling. Their effectiveness in the long time integration of Hamiltonian ODE systems is amply demonstrated in the references already cited.

The transfer of these ideas to PDEs is relatively recent; and there are several alternative approaches. One is to discretise in space so as to obtain a Hamiltonian system of ODEs to which the above ideas can be applied directly: there is increasing interest in mesh-free or particle methods to achieve this step, but as we have hitherto excluded particle methods we shall continue to do so here; alternatively, one may first make a discretisation in space and then apply the 'method of lines ' to integrate in time, but we will not consider this here either. A more fundamental formulation is due to Bridges[1].This leads to a multi-symplectic PDE which generalises the form of (2.37) to

$$\mathbf{K}\mathbf{z}_t + \mathbf{L}\mathbf{z}_x = \nabla_{\mathbf{z}} S(\mathbf{z}), \qquad (2.50)$$

where **K** and **L** are constant skew-symmetric matrices. Unfortunately, these matrices and linear combinations of them are often singular, and the formulation of a given system in this way not very obvious. We will therefore apply a more straightforward approach to a wave equation problem that generalises (2.26) and (2.27).

Suppose we have a Hamiltonian $\mathcal{H}(u, v)$ which is now an integral over the space variable (s) of a function of u, v and their spatial derivatives. Then to derive a Hamiltonian PDE we define a variational derivative of $\mathcal{H}$. For example, consider

$$\mathcal{H}(u,v) = \int E(x,t)\,\mathrm{d}x \equiv \int \left[f(u) + g(u_x) + \frac{1}{2}v^2\right]\mathrm{d}x, \quad (2.51)$$

where we have not specified the interval on which u and v are defined and the equations are to hold; the integrand $E(x,t)$ is called the energy density. The variational derivative of a functional $\mathcal{G}(u)$ is defined by the relation

$$\int \delta_u \mathcal{G}(u)(\delta u)\,\mathrm{d}x = \lim_{\epsilon\to 0} \frac{\mathcal{G}(u+\epsilon\delta u) - \mathcal{G}(u)}{\epsilon};$$

and applying this to (2.51), with boundary conditions that ensure any boundary terms are zero, gives

$$\int \delta_u \mathcal{H}(u,v)(\delta u)\mathrm{d}x = \lim_{\epsilon\to 0} \varepsilon^{-1} \int [f(u+\epsilon\delta u) - f(u) + g((u+\epsilon\delta u)_x) - g(u_x)]\,\mathrm{d}x$$

$$= \int [f'(u)\delta u + g'(u_x)(\delta u)_x]\,\mathrm{d}x$$

$$= \int [f'(u) - \partial_x g'(u_x)]\,\delta u\,\mathrm{d}x. \quad (2.52)$$

Comparing the two sides we deduce that

$$\delta_u \mathcal{H}(u,v) = f'(u) - \partial_x g'(u_x). \quad (2.53)$$

The resulting Hamiltonian PDE is given as

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} = \begin{pmatrix} 0 & +1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \delta_u \mathcal{H} \\ \delta_v \mathcal{H} \end{pmatrix}. \quad (2.54)$$

That is,

$$u_t = v, \quad v_t = \partial_x g'(u_x) - f'(u). \quad (2.55)$$

Moreover, from these equations we can deduce a local energy conservation law of the form $E_t + F_x = 0$: from differentiation of the terms in the energy density of (2.51) and substitution from (2.55) we get, after cancellation and collection of terms,

$$E_t = f'(u)v + g'(u_x)v_x + v[\partial_x g'(u_x) - f'(u)]$$

$$= [v g'(u_x)]_x =: -F_x. \quad (2.56)$$

The quantity $F(x,t) = -vg'(u_x)$ is called the energy flux.

For example, let $f = 0$ and $g(u_x) = \frac{1}{2}(au_x)^2$ with constant a. Then (2.55) becomes

$$u_t = v, \quad v_t - a^2 u_{xx}, \tag{2.57}$$

which is equivalent to the second order wave equation (2.26). If we set $w = -au_x$ we get the first order pair of equations (2.27) to which we applied the staggered leap-frog method in Section 2.4. Furthermore, since $vg'(u_x) = va^2 u_x = -avw$ the local energy conservation law becomes

$$\left[\frac{1}{2}v^2 + \frac{1}{2}w^2\right]_t + [avw]_x = 0, \tag{2.58}$$

which we could deduce directly from (2.27). It is this local property that we shall now show is preserved in a discrete form by the staggered leap-frog scheme. It can be regarded as the simplest consequence of the symplectic character of the method, and corresponds to the energy being a constant of the motion in the ODE case. Consideration of wedge product relations of the form (2.49), which now have to be integrated or summed over the space variables, is beyond the scope of thisthesis.

Chapter Three
Consistency, Convergence and Stability

3.1 Definition of the problems considered

In this chapter we shall gather together and formalise definitions that we have introduced in earlier chapters. This will enable us to state and prove the main part of the key Lax Equivalence Theorem. For simplicity we will not aim at full generality but our definitions and arguments will be consistent with those used in a more general treatment. In the problems which we shall consider, we make the following assumptions:

- The region Ω is a fixed bounded open region in a space which may have one, two, three or more dimensions, with co-ordinates which may be Cartesian $(x, y, \ldots)$, cylindrical polar, spherical polar, etc.;
- The region Ω has boundary $\partial\Omega$;
- The required solution is a function u of the space variables, and of t, defined on $\Omega \times [0, t_F]$; this function may be vector-valued, so that our discussion can be applied to systems of differential equations, as well as to single equations;
- The operator $L(\cdot)$ involves the partial derivatives of u in the space variables; L does not involve t explicitly; for the most part we shall assume that L is a linear operator.
- The boundary conditions will prescribe the values of $g(u)$ on some or all of the boundary Ω, where $g(\cdot)$ is an operator which may involve spatial partial derivatives;
- The initial condition prescribes the value of u for $t = 0$ over the region Ω.

Hence we write the general form of the problems considered as

$$\frac{\partial u}{\partial t} = L(u) \quad \text{in} \quad \Omega \times (0, t_F], \tag{3.1a}$$

$$g(u) = g_0 \quad \text{on} \quad \partial\Omega_1 \subset \partial\Omega, \tag{3.1b}$$

$$u = u^0 \quad \text{on} \quad \Omega \quad \text{when} \quad t = 0. \tag{3.1c}$$

We shall always assume that (3.1) defines a well-posed problem, in a sense which we shall define later; broadly speaking, it means that a solution always exists and depends continuously on the data.

3.2The finite difference mesh and norms

Our finite difference approximation will be defined on a fixed mesh, with the time interval Δt constant both over the mesh and at successive time steps. The region Ω is covered by a mesh which for simplicity we shall normally assume has uniform spacing $\Delta x, \Delta y, \ldots$ in Cartesian co-ordinates, or $\Delta \mathrm{r}, \Delta \theta, \ldots$ in polar co-ordinates. Individual values at mesh points will be denoted by U_j^n; in two or more space dimensions the subscript j will be used to indicate a multi-index, as a condensed notation for $U_{j,k}^n, U_{j,k,l}^n$, etc. We shall assume that a fixed, regular finite difference scheme is applied to a set of points where U_j^n is to be solved for and whose subscripts j lie in a set J_Ω, and it is only these points which will be incorporated in the norms. Usually this will be just the interior points of the mesh; and this means that where made necessary by curved boundaries, derivative boundary conditions etc. The values of U at all such points on time level n will be denoted by U^n:

$$U^n := \{U_j^n, j \in J_\Omega\}. \qquad (3.2)$$

To simplify the notation we will consider schemes which involve only two time levels: for one-step methods this means that each U_j^n, if a vector, has the same dimension as u. However, as we have seen with the leap-frog method in Section 2.4, we can include multi-step methods by extending the dimension of U_j^n compared with u. For example, if a scheme involves three time levels, so that U^{n+1} is given in terms of U^n and U^{n-1}, we can define a new vector $\widetilde{U}^n$ with twice the dimension, whose elements are those of U^n and U^{n-1}.

To compare U with u we need to introduce norms which can be used on either, and in particular on their difference. Thus we first denote by u_j^n mesh values of the function $u(x,t)$ which will usually be the point values

$u(x_j, t_n)$. We hope to show that the mesh values of U converge to these values of u. Then as for the mesh point values U_j^n above we define

$$u^n := \{u_j^n, j \in J_\Omega\}. \qquad (3.3)$$

We shall consider just two norms. Firstly, the maximum norm is given by

$$\|U^n\|_\infty := \max\{|u_j^n|, j \in J_\Omega\}. \qquad (3.4)$$

If we evaluate the maximum norm of u^n the result will approximate the usual supremum norm $\|u\|_\infty$ with u considered as a function of x at fixed time t_n, but will not in general be equal to it. The norms will only be equal if the maximum value of the function $|u(x, t_n)|$ is attained at one of the mesh points.

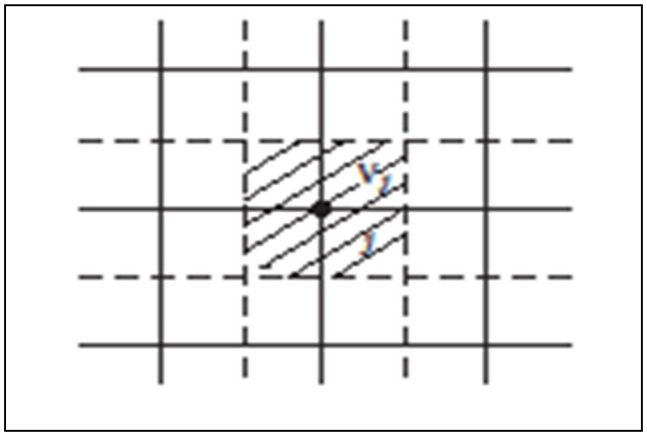

Figure (3.1): Definition of control volume.

Secondly, we shall use a discrete l_2 norm which will approximate the integral L_2 norm. To do so, we introduce a 'control volume' with measure V_j associated with each interior mesh point: these will be non-overlapping elements whose union approximates Ω. Usually, as shown in Figure (3.1), a mesh point x_j will lie at the centre of the control volume – see also Section 2.1 on finite volume methods; but this need not be the case so long as there is a one-to-one correspondence between mesh points and control volumes. In three-dimensional Cartesian geometry, $V_j = \Delta x \Delta y \Delta z$; in three-dimensional cylindrical

geometry, $V_j = r_j \Delta\theta \Delta r \Delta z$, and so on. Then, we define

$$\|u^n\|_2 := \left\{ \sum_{j \in J_\Omega} V_j \left| U_j^n \right|^2 \right\}^{1/2} . \qquad (3.5)$$

For mesh points near the boundary the control volume may or may not be modified to lie wholly in Ω. In either case, the sum in (3.5) clearly approximates an integral so that $\|u^n\|_2$ approximates but does not in general equal the integral L_2 norm

$$\|u(\cdot, t_n)\|_2 := \left[\int_\Omega |u(x, t_n)|^2 \, dV \right]^{1/2} \qquad (3.6)$$

at time t_n. However, if we define u_j^n as the root mean square value of $u(x, t_n)$ averaged over the jth control volume we clearly do have an exact match. For a single differential equation the notation $\left| U_j^n \right|$ is clear; if we are dealing with a system of differential equations, U_j^n is a vector and $\left| U_j^n \right|$ denotes a norm of this vector. The choice of which vector norm to use is immaterial to the subsequent analysis, but of course it must be used consistently throughout.

3.3Finite difference approximations

The general form of difference scheme we shall consider will be written

$$B_1 U^{n+1} = B_0 U^n + F^n. \qquad (3.7)$$

As the notation implies, the difference operators B_0, B_1 are independent of n, corresponding to the assumption that $L(\cdot)$ does not depend explicitly on t; but, although based on fixed difference operators, they may depend on the point where they are applied. Thus at each point $j \in J_\Omega$, a linear difference operator B will be written in the form of a sum over near neighbours also in J_Ω:

$$(BU^n)_j = \sum_{k \in J_\Omega} b_{j,k} U_k^n \quad \forall j \in J_\Omega; \qquad (3.8)$$

We shall always assume that B_1 is linear, of the form (3.8), so that it can be represented by a square matrix. To extend the theory to nonlinear problems it would be necessary for B_0 to be nonlinear but not necessarily B_1; but to cover schemes like the box scheme by such an extension would require B_1 to be nonlinear too.

We shall furthermore assume that B_1 is invertible, i.e. its representing matrix is non-singular. Hence we can write (3.7) as

$$U^{n+1} = B_1^{-1}[B_0 u^n + F^n] \qquad (3.9)$$

We shall also assume that (3.7) is so scaled that formally it represents the differential equation in the limit and hence $B = O(1/\Delta t)$. Thus

$$B_1 u^{n+1} - [B_0 u^n + F^n] \to \frac{\partial u}{\partial t} - L(u) \qquad (3.10)$$

as the mesh intervals $\Delta t, \Delta x, \ldots$ are refined in some manner which may depend on consistency conditions being satisfied.

Moreover, we assume that the matrix B_1 is uniformly well-conditioned in the sense that there is a constant K such that, in whichever norm is being used to carry out the analysis,

$$\|B_1^{-1}\| \leq K_1 \Delta t, \qquad (3.11)$$

even though B_1^{-1} is represented by a matrix of ever-increasing dimension as the limit $\Delta t \to 0$ is approached.

Consistency, order of accuracy and convergence

For brevity we shall characterise the whole of the spatial discretisation by a single parameter h: this may be just the largest of the mesh intervals $\Delta x, \Delta y, \ldots$, though this may need to be scaled by characteristic speeds in each of the co-ordinate directions; or h may be the diameter of the largest control volume around the mesh points. Then taking the limit along some designated refinement path we shall denote by '$\Delta t(h) \to 0$', or sometimes just $\Delta t \to 0$ or $h \to 0$: we shall always need Δt to tend to zero but stability or consistency may require that it does so at a rate determined by h, for example $\Delta t =$

$O(h^2)$ being typical in parabolic problems and $\Delta t = O(h)$ in hyperbolic problems.

The truncation error is defined in terms of the exact solution u as

$$T^n := B_1 u^{n+1} - [B_0 u^n + F^n], \qquad (3.12)$$

and consistency of the difference scheme (3.7) with the problem (3.1a)–(3.1c) as

$$T_j^n \to 0 \quad \text{as} \quad \Delta t(h) \to 0 \quad \forall j \in J_\Omega \qquad (3.13)$$

for all sufficiently smooth solutions u of (3.1a)–(3.1c). Note that this includes consistency of the boundary conditions through the elimination of the boundary values of U in the definition of B_0 and B_1. If p and q are the largest integers for which

$$\left|T_j^n\right| \le C[(\Delta t)^p + h^q] \quad \text{as} \quad \Delta t(h) \to 0 \quad \forall j \in J_\Omega \qquad (3.14)$$

for sufficiently smooth u, the scheme is said to have order of accuracy p in Δt and q in h: or pth order of accuracy in Δt, and qth order of accuracy in h.

Convergence on the other hand is defined in terms of all initial and other data for which (3.1a)–(3.1c) is well-posed, in a sense to be defined in the next section. Thus (3.7) is said to provide a convergent approximation to (3.1a)–(3.1c) in a norm $\|\cdot\|$ if

$$\|U^n - u^n\| \to 0 \quad \text{as} \quad \Delta t(h) \to 0, n\Delta t \to t \in (0, t_F] \qquad (3.15)$$

for every u_0 for which (3.1a)–(3.1c) is well-posed in the norm: here we mean either of the norms (3.4) or (3.5).

3.4 Stability and the Lax Equivalence Theorem

None of the definitions (3.12)–(3.15) in the last section was limited to linear problems: they are quite general. In this section however we are able to consider only linear problems. Suppose two solutions V^n and W^n of (3.7) or (3.9) have the same inhomogeneous terms F^n but start from different

initial data V^0 and W^0: we say the scheme is stable in the norm $\|\cdot\|$ and for a given refinement path if there exists a constant K such that

$$\|V^n - W^n\| \leq K\|V^0 - W^0\|, \quad n\Delta t \leq t_F; \tag{3.16}$$

the constant K has to be independent of V^0, W^0 and of $\Delta t(h)$ on the refinement path, so giving a uniform bound.

Since we are dealing with the linear case (3.16) can be written

$$\|(B_1^{-1}B_0)^n\| \leq K, \quad n\Delta t \leq t_F. \tag{3.17}$$

Notice that for implicit schemes the establishment of (3.11) is an important part of establishing (3.17); consider, for example, the box scheme for linear advection, and the effect of having boundary conditions on one side or the other.

It is now appropriate to formalise our definition of well-posedness. We shall say that the problem (3.1) is well-posed in a given norm $\|\cdot\|$ if, for all sufficiently small h, we can show that (i) a solution exists for all data u^0 for which $\|u^0\|$ is bounded independently of h, and (ii) there exists a constant K' such that for any pair of solutions v and w,

$$\|v^n - w^n\| \leq K'\|v^0 - w^0\|, \quad t_n \leq t_F. \tag{3.18}$$

This differs from the usual definition in that we are using discrete norms; but we have chosen each of these so that it is equivalent to the corresponding function norm as $h \to 0$, if this exists for u, and we define u_j^n appropriately. An important feature of either definition is the following: for u to be a classical solution of (3.1a) it must be sufficiently smooth for the derivatives to exist; but suppose we have a sequence of data sets for which smooth solutions exist and these data sets converge to arbitrary initial data u^0 in the $\|\cdot\|$ norm, uniformly in h; then we can define a generalised solution with this data as the limit at any time t_n of the solutions with the smooth data, because of (3.18). Thus the existence of solutions in establishing well-posedness has only to be proved for a dense set of smooth data (with the definition of denseness again being uniform in h).

There is clearly a very close relationship between the definition of well-posedness for the differential problem and that of stability given by (3.16) for the discrete problem. This definition of stability, first formulated by Lax in 1953, enabled him to deduce the following key theorem:

Theorem (3.1): (Lax Equivalence Theorem)

For a consistent difference approximation to a well-posed linear evolutionary problem, which is uniformly solvable in the sense of (3.11), the stability of the scheme is necessary and sufficient for convergence.

Proof (of sufficiency):

Subtracting (3.12) from (3.7) we have

$$B_1(U^{n+1} - u^{n+1}) = B_0(U^n - u^n) - T^n,$$

i.e.,

$$U^{n+1} - u^{n+1} = (B_1^{-1}B_0)(U^n - u^n) \quad - B_1^{-1}T^n. \tag{3.19}$$

Assuming that we set $U^0 = u^0$, it follows that

$$U^n - u^n = -[B_1^{-1}T^{n-1} + (B_1^{-1}B_0)B_1^{-1}T^{n-2} + \cdots + (B_1^{-1}B_0)^{n-1}B_1^{-1}T^0]. \tag{3.20}$$

Now in applying the theorem, (3.11) and (3.16) are to hold in the same norm, for which we shall also deduce (3.15); we can combine these two to obtain

$$\|(B_1^{-1}B_0)^m B_1^{-1}\| \leq KK_1\Delta t \tag{3.21}$$

from which (3.19) gives

$$\|U^n - u^n\| \leq KK_1\Delta t \sum_{m=0}^{n-1} \|T^m\|.$$

Thus convergence in the sense of (3.15) follows from the consistency of (3.13), if u is sufficiently smooth for the latter to hold. For less smooth solutions, convergence follows from the hypotheses of well-posedness and stability: general initial data can be approximated arbitrarily closely by data

for smooth solutions and the growth of the discrepancy is bounded by the well-posedness of the differential problem and the stability (3.15) of the discrete problem.

Calculating stability conditions

As we are dealing with linear problems, if in (3.16) V^n and W^n are solutions of the difference equations (3.7), then the difference $V^n - W^n$ is a solution of the homogeneous difference equations with homogeneous boundary data. That is, establishing stability is equivalent to establishing the following:

$$B_1 U^{n+1} = B_0 U^n \quad \text{and} \quad n\Delta t \le t_F \quad \Rightarrow \quad \|U^n\| \le K\|U^0\|, \tag{3.22}$$

which is what is meant by (3.17). The constant K will generally depend on the time interval t_F and allows for the sort of exponential growth that might occur with $u_t = u_x + u$, for example. For simple problems one will often find: either $K = 1$, there is no growth and the scheme is stable; or $U^n \sim \lambda^n U^0$, with $|\lambda| > 1$ even as $\Delta t \to 0$ for some mode, so the scheme is unstable.

Thus when establishing a maximum principle we have to establish stability in the maximum norm: strictly speaking, we have also to establish a minimum principle so as to be able to say not only

$$U_j^{n+1} \le \max_k U_k^n \le \|U^n\|_\infty \tag{3.23}$$

but also

$$U_j^{n+1} \ge \min_k U_k^n \ge -\|U^n\|_\infty \tag{3.24}$$

and can then deduce

$$\|U^{n+1}\|_\infty \le \|U^n\|_\infty. \tag{3.25}$$

Amaximum principle is seldom available or even appropriate for hyperbolic problems. As we have noted, the first order scheme (1.23) satisfies a maximum principle whenever $0 \le \nu \le 1$ so that it is then stable in the maximum norm: but we can show that this can never be true of a second

order scheme. For example, consider the Lax–Wendroff method written in the form (1.39). If it were to satisfy a maximum principle, then for any set of non-positive values for U^n one should never have $U^{n+1} > 0$: yet if $0 < \nu < 1$, setting $U^n_{j-1} = U^n_j = 0$ and $U^n_{j+1} = -1$ gives a positive value for U^{n+1}_j. This does not of course demonstrate that the scheme is actually unstable in the maximum norm, merely that we cannot prove such stability by this means.

For this reason, and also because hyperbolic differential equations are much more commonly well-posed in the L_2 norm than in the supremum norm, for hyperbolic problems we have to adopt the more modest target of proving stability in the l_2 norm (3.5). This gives weaker results because we have, recalling that V_j is the measure of the jth control volume,

$$\left[\min_{j\in J_\Omega} V_j\right]^{1/2} \|U\|_\infty \le \|U\|_2 \le \left[\sum_{j\in J_\Omega} V_j\right]^{1/2} \|U\|_\infty; \qquad (3.26)$$

in the bounded region we are working with, the coefficient on the right is a finite constant while that on the left tends to zero as the mesh is refined. It is clear that we would prefer to derive a maximum norm error bound from a stability analysis but, if we have only l_2 stability and so obtain a bound for the l_2 norm of the error $\|E^n\|_2$, then (3.26) gives a poor result for $\|E^n\|_\infty$.

However, it is the l_2 norm which is appropriate for Fourier analysis because of Parseval's relation. Suppose we can assume periodicity on a normalized region $[-\pi, \pi]^d$ which is covered by a uniform (Cartesian) mesh of size $\Delta x_1 = \Delta x_2 = \ldots = \Delta x_d = \pi/J$. Then the Fourier modes that can be distinguished on the mesh correspond to wave numbers, which we denote by the vector $\boldsymbol{k}$, having components given by

$$k = 0, \pm 1, \pm 2, \ldots, \pm J, \qquad (3.27)$$

where the last two with $k\Delta x = \pm\pi$ are actually indistinguishable. Hence we can expand any periodic function on the mesh as

$$U(\mathbf{x}_j) = \frac{1}{(2\pi)^{d/2}} \sum_{(\boldsymbol{k})}' \widehat{U}(\mathbf{k}) \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathrm{x_j}} \qquad (3.28)$$

where the prime on the summation sign means that any term with $k_s = \pm J$ has its weight halved, and we have also used a vector notation $\mathbf{x}_j$ for mesh points. This discrete Fourier expansion has an inverse which is the discrete Fourier transform

$$\widehat{U}(\mathbf{k}) = \frac{1}{(2\pi)^{d/2}} \sum_{(j)}' (\Delta x)^d U(\mathbf{x}_j) \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathrm{x_j}} \qquad (3.29)$$

where each component of j runs from $-J$ to J with the mesh points on the periodic boundaries again having their weights halved so that all the weights are equal to the V_j introduced in (3.5).

Lemma (3.1):

The Fourier modes $(2\pi)^{-d/2}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\boldsymbol{x}_j}$ with components given by (3.27) are orthonormal with respect to the l_2 inner product used in (3.29), namely

$$\langle U, W \rangle_2 := (\Delta x)^d \sum_{(j)}' U_j \bar{W}_j . \qquad (3.30)$$

Proof:

It is sufficient to consider $d = 1$. We first establish the fundamental trigonometric identity

$$\frac{1}{2}\mathrm{e}^{-\mathrm{i}J\theta} + \mathrm{e}^{-\mathrm{i}(J-1)\theta} + \cdots + \mathrm{e}^{\mathrm{i}(J-1)\theta} + \frac{1}{2}\mathrm{e}^{\mathrm{i}J\theta} = \sin J\theta \cos\frac{1}{2}\theta . \qquad (3.31)$$

From the summation

$$1 + \mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{\mathrm{i}2\theta} + \cdots + \mathrm{e}^{\mathrm{i}(J-1)\theta} = \left(\mathrm{e}^{\mathrm{i}J\theta} - 1\right)/\left(\mathrm{e}^{\mathrm{i}\theta} - 1\right)$$

we obtain by adding $\frac{1}{2}\left(\mathrm{e}^{\mathrm{i}J\theta} - 1\right)$

$$\frac{1}{2} + \mathrm{e}^{\mathrm{i}\theta} + \mathrm{e}^{\mathrm{i}2\theta} + \cdots + \mathrm{e}^{\mathrm{i}(J-1)\theta} + \frac{1}{2}\mathrm{e}^{\mathrm{i}J\theta} = \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}J\theta} - 1\right)\frac{\left(\mathrm{e}^{\mathrm{i}\theta} + 1\right)}{\left(\mathrm{e}^{\mathrm{i}\theta} - 1\right)} \qquad (3.32)$$

$$= \frac{1}{2\mathrm{i}}\left(\mathrm{e}^{\mathrm{i}J\theta} - 1\right)\cos\frac{1}{2}\theta. \qquad (3.33)$$

Combining this with a similar sum for $-\theta$ gives (3.31). Now apply this with $\theta = (k_1 - k_2)\Delta x$, so that $J\theta = (k_1 - k_2)\pi$. We obtain

$$\sum_{(j)}{}' \mathrm{e}^{\mathrm{i}k_1 x_j}\mathrm{e}^{-\mathrm{i}k_2 x_j} = \sin(k_1 - k_2)\pi \cot\frac{1}{2}(k_1 - k_2)\Delta x, \quad k_1 \neq k_2,$$

so that

$$\sum_{(j)}{}' \mathrm{e}^{\mathrm{i}k_1 x_j}\mathrm{e}^{-\mathrm{i}k_2 x_j} = (2\pi/\Delta x)\delta_{k_1,k_2}.$$

Hence we have, with V_j the control volume measure,

$$\|U\|_2^2 = \sum_{j\in J_\Omega} V_j\left|U_j\right|^2 \equiv \sum_{(j)}{}' (\Delta x)^d \left|U(\mathrm{x}_j)\right|^2$$

$$= \left(\frac{2\pi}{\Delta x}\right)^d \sum_{(\boldsymbol{k})}{}' \left|\widehat{U}(\mathbf{k})\right|^2 \left(\frac{2\pi}{\Delta x}\right)^d, \qquad (3.34)$$

i.e.,

$$\left\|\widehat{U}\right\|_2^2 = \sum_{(\mathbf{k})}{}' \left|\widehat{U}(\mathbf{k})\right|^2 = \|U\|_2^2, \qquad (3.35)$$

which is the appropriate form of Parseval's relation.

For a rectangular region of general dimensions a simple scaling will reduce the situation to the above case. However, note that not only is Δx then changed but we will also generally have $\Delta k \neq 1$ and that such a coefficient will be needed in the definition of $\left\|\widehat{U}\right\|_2$ for (3.35) to hold. It is also worth noting that when for example we have a problem on $[0,1]$ with $u(0) = u(1) = 0$ we extend this to a periodic problem on $[-1,1]$ by imposing antisymmetry at $x = 0$ and using a sine series. This is why we have taken $[-\pi, \pi]$ as our standard case above.

To establish (3.22) then, for a constant coefficient problem with periodic boundary conditions, we expand arbitrary initial data in the form (3.28) and, from the discrete Fourier transform of (3.22), obtain the same form at successive time levels with the coefficients given by

$$\hat{B}_1(\mathbf{k})\hat{U}^{n+1}(\mathbf{k}) = \hat{B}_0(\mathbf{k})\hat{U}^n(\mathbf{k}), \qquad (3.36)$$

where, if the U^n are p-dimensional vectors, $\hat{B}_0$ and $\hat{B}_1$ are $p \times p$ matrices. The matrix

$$G(\mathbf{k}) = \hat{B}_1^{-1}(\mathbf{k})\hat{B}_0(\mathbf{k}) \qquad (3.37)$$

is called the amplification matrix as it describes the amplification of each mode by the difference scheme. Because we have assumed that $\hat{B}_0$ and $\hat{B}_1$ are independent of t we can write

$$\hat{U}^n = [G(\mathbf{k})]^n \hat{U}^0 \qquad (3.38)$$

and using (3.35) have

$$\sup_{U^0} \frac{\|U^n\|_2}{\|U^0\|_2} = \sup_{\hat{U}^0} \frac{\left[\sum'_{(\mathbf{k})} \left|\hat{U}^n(\mathbf{k})\right|^2\right]^{1/2}}{\left[\sum'_{(\mathbf{k})} \left|\hat{U}^0(\mathbf{k})\right|^2\right]^{1/2}}$$

$$= \sup_{\mathbf{k}} \sup_{\hat{U}^0(\mathbf{k})} \frac{\left|\hat{U}^n(\mathbf{k})\right|}{\left|\hat{U}^0(\mathbf{k})\right|} = \sup_{\mathbf{k}} |[G(\mathbf{k})]^n|. \qquad (3.39)$$

Thus stability in the l_2 norm is equivalent to showing that

$$|[G(\mathbf{k})]^n| \leq \mathrm{K} \quad \forall \mathbf{k}, \quad n\Delta t \leq t_F. \qquad (3.40)$$

Here $|G^n|$ means the $p \times p$ matrix norm subordinate to the vector norm used for U_j^n and. $\hat{U}(\mathbf{k})$.

Then clearly we have the following result.

Theorem (3.2): (von Neumann Condition)

A necessary condition for stability is that there exist a constant K' such that

$$|\lambda(\boldsymbol{k})| \leq 1 + K'\Delta t \quad \forall \mathbf{k}, \quad n\Delta t \leq t_F, \qquad (3.41)$$

for every eigenvalue $\lambda(\mathbf{k})$ of the amplification matrix $G(\mathbf{k})$.

Proof:

By taking any eigenvector of $G(\mathbf{k})$ as $\widehat{U}(\mathbf{k})$ it is obviously necessary that there be a constant K such that $|\lambda^n| \leq K$: then by taking $n\Delta t = t_F$ we have

$$|\lambda| \leq K^{\Delta t/t_F} \leq 1 + (K-1)\,\Delta t/t_F \quad \text{for} \quad \Delta t \leq t_F,$$

the last inequality following from the fact that K^s is a convex function of s.

References:

[1]T.J. Bridges, Multi-symplectic structure and wave propagation, Math. Proc. Comb.Philos. Soc. 121(1997), 147-190.

[2]$P. Colella\ and\ P. R. Woodward, The\ Piecewise\ Parabolic\ method(PPm)$

$for\ gas - dynamical\ simulation, J. of\ Comput. Phys.$ 54(1984), 174 – 201.

[3]E.Hairer, C.Lubich, and G. Wanner, Geometric Numerical Integration, Berlin, Springer- Verlag, 2002.

[4]B.van leer, Towards the ultimate conservation difference scheme. monotonicity and conservation combined in a second order scheme, J. of Comput. Phys. 14(1974), 361-370.

[5]$B. Leimkuhler\ and\ S. Reich, Simulating\ Hamiltonian\ Dynamics, Cambridge,$ ambridge University Press, 2004.

[6]$K. W. Morton\ and\ D. F. Mayer, Numerical\ soluation$

$of\ partial\ Differential$ Equation, An Introduction, Cambridge University Press, 2005.

[7]P.L. ROe, Appoximate Riemann solver, parameter vector, and difference scheme, J. of Comput. Phys. 43(1981), 357-372.

.