MATH2103: Lecture Note on Numerical Solution of Partial Differential Equations

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2025 年 10 月 20 日

Contents

Chapter 1

Hyperbolic equations in one space dimension

1.1 Characteristics

The linear advection equation

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \tag{1.1}$$

must surely be the simplest of all 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

$$\frac{dx}{dt} = a(x,t), \tag{1.2}$$

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

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dx}{dt} = 0. \tag{1.3}$$

Thus from initial data

$$u(x,0) = u^0(x), \tag{1.4}$$

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 Fig. 4.1, and finding the characteristic through $(x_j, 0)$ by a numerical solution of (4.2a) 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.

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

$$u(x,t) = u^0(x - at). \tag{1.5}$$

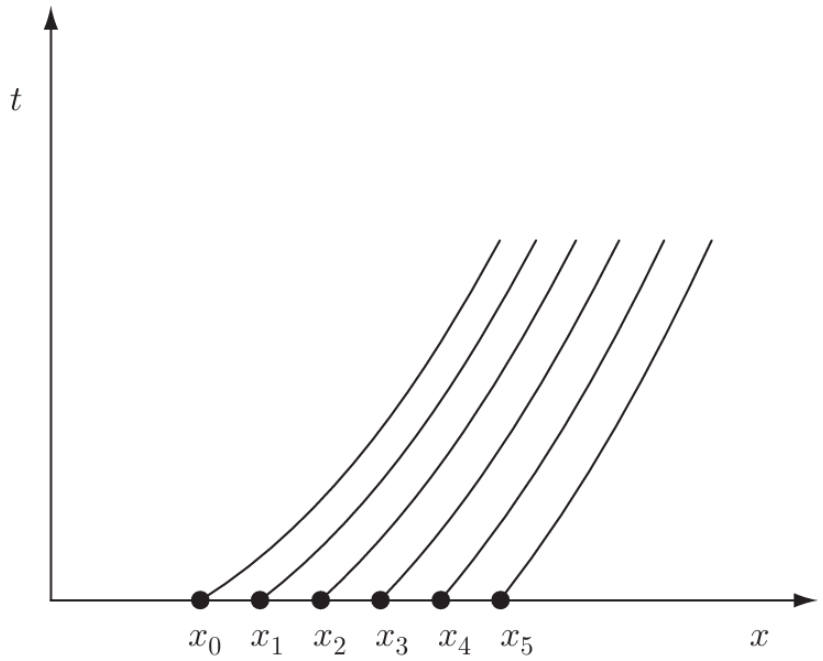


Fig. 4.1. Typical characteristics for $u_t + a(x,t)u_x = 0$.

Figure 1.1: Characteristic lines.

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

$$u(x,t) = u^0(x - a(u(x,t))t), \tag{1.6}$$

until the time when this breaks down because the characteristics can now envelope or cross each other in some other manner - see the following Section x.6.

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

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

where $\mathbf{u} = \mathbf{u}(x,t)$ 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 (1.6) 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

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

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

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

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 Λ 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

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

If it is possible to define a vector of **Riemann invariants** $\mathbf{r} = \mathbf{r}(\mathbf{u})$ such that

$$\mathbf{r}_t = S\mathbf{u}_t \tag{1.15}$$

and

$$\mathbf{r}_x = S\mathbf{u}_x, \tag{1.16}$$

then we can write

$$\mathbf{r}_t + \Lambda\mathbf{r}_x = 0 \tag{1.17}$$

which is a direct generalisation of the scalar case whose solution we have given in (1.5). However, now each component of Λ will usually depend on all the components of $\mathbf{r}$ so that the characteristics will be curved. Moreover, although these Riemann invariants can always be defined for a system of two equations, for a larger system this is not always possible.

To apply the method of characteristics to problems like (1.6), 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.**

However, to use such a technique in a direct way in two space dimensions does become excessively complicated: for there we have characteristic surfaces and many more complicated solution phenomena to describe. Thus, even though this chapter is only on one-dimensional problems, in line with our general philosophy set out in the first chapter, **we shall not consider the method of characteristics in any more detail.** Instead we shall confine our considerations to methods based on a fixed mesh in space: and although the length of the time step may vary from step to step it must be the same over all the space points. We shall start with explicit methods on a uniform mesh.

1.2 The CFL condition

1.2.1 The CFL condition for the upwind scheme

Courant, Friedrichs and Lewy, in their fundamental 1928 paper [1] 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 Fig. 1.2.

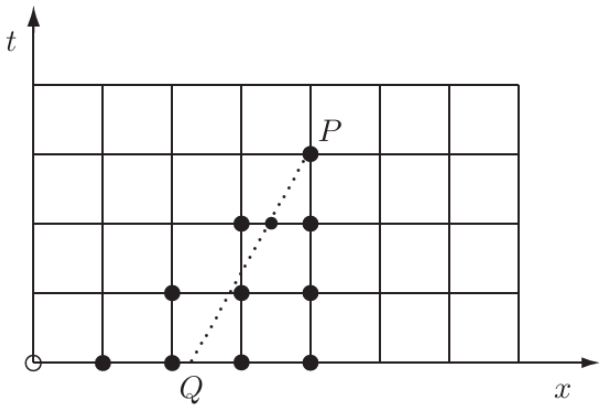


Fig. 4.2. Typical domain of dependence.

Figure 1.2: Domain of dependence.

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

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0. \tag{1.18}$$

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

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{a\Delta t}{\Delta x}(U_j^n - U_{j-1}^n) \\ &= (1-\nu)U_j^n + \nu U_{j-1}^n, \end{aligned} \tag{1.19}$$

where

$$\nu = \frac{a\Delta t}{\Delta x}. \tag{1.20}$$

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. This is illustrated in Fig. 1.2.

Proposition 1.2.1 *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.*

[1] Courant, R., Friedrichs, K.O. and Lewy, H. (1928), Über die partiellen Differenzengleichungen der mathematischen Physik, Math. Ann. **100**, 32–74.

Proposition 1.2.2 *CFL is necessary but not sufficient condition for stability. (CFL $\Leftarrow \not\Rightarrow$ stable)*

What we have thus obtained can also be regarded as a necessary condition for the stability of this difference scheme, somewhat similar to the condition for the stability of the explicit scheme for the parabolic equation in Chapter 2, but more obviously **applicable to problems with variable coefficients, or even nonlinear problems**. 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.

1.2.2 The stability for the upwind scheme by Fourier analysis

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:

$$U_j^{n+1} = \begin{cases} U_j^n - a\frac{\Delta t}{\Delta x}\Delta_{+x}U_j^n & \text{if } a < 0, \\ U_j^n - a\frac{\Delta t}{\Delta x}\Delta_{-x}U_j^n & \text{if } a > 0. \end{cases} \tag{1.21}$$

If a is not a constant, but a function of x and t, we must specify which value is used in (1.21). We shall for the moment assume that we use $a(x_j, t_n)$, but still write a without superscript or subscript and $\nu = a\Delta t/\Delta x$ as in (1.20) when this is unambiguous.

This scheme clearly satisfies the CFL condition when (1.24) is satisfied, and a Fourier analysis gives for the constant $a > 0$ case the amplification factor

$$\lambda \equiv \lambda(k) = 1 - (a\Delta t/\Delta x)(1 - e^{-ik\Delta x}) \equiv 1 - \nu(1 - e^{-ik\Delta x}). \tag{1.22}$$

This leads to

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

which gives

$$|\lambda|^2 = 1 - 4\nu(1-\nu)\sin^2\frac{1}{2}k\Delta x. \tag{1.23}$$

It follows that $|\lambda(k)| \leq 1$ for all k provided that $0 \leq \nu \leq 1$. 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, in agreement with the von Neumann condition.

1.2.3 The CFL condition for a scheme using a central difference in space

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

$$|a|\Delta t \leq \Delta x, \tag{1.24}$$

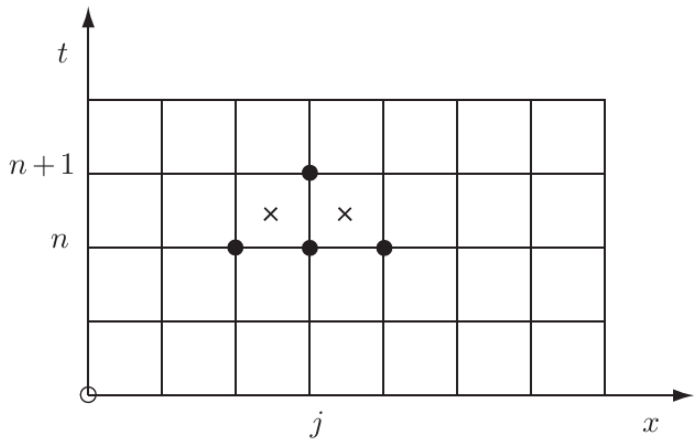


Fig. 4.4. General three-point scheme; the points marked × are used for the two-step Lax–Wendroff method.

Figure 1.3: General three-point scheme; used for the two-step Lax-Wendroff method.

as we see from Fig. 1.3; $\nu := |a|\Delta t/\Delta x$ is often called the **CFL number**.

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

$$\frac{U^{n+1}_j - U^n_j}{\Delta t} + a\frac{U^n_{j+1} - U^n_{j-1}}{2\Delta x} = 0. \tag{1.25}$$

If we satisfy (1.24) the CFL condition holds for either sign of a.

1.2.4 The stability for a scheme using a central difference in space

But now in the case where a is constant, and **ignoring the effect of the boundary conditions (Cauchy problem)**, we can investigate the stability of the scheme by Fourier analysis, as we did for parabolic equations in Chapter 2. The Fourier mode

$$U^n_j = (\lambda)^n e^{ik(j\Delta x)} \tag{1.26}$$

satisfies the difference scheme (1.25) provided that the amplification factor λ satisfies

$$\lambda \equiv \lambda(k) = 1 - (a\Delta t/\Delta x) i \sin k\Delta x. \tag{1.27}$$

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 $U_j \propto (-1)^j$, 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.*

1.3 Error analysis of the upwind scheme

We notice that the scheme (1.21) can be written

$$U_j^{n+1} = \begin{cases} (1+\nu)U_j^n - \nu U_{j+1}^n & \text{if } a < 0, \\ (1-\nu)U_j^n + \nu U_{j-1}^n & \text{if } a > 0. \end{cases} \tag{1.28}$$

Notice also that all the coefficients in (1.28) are nonnegative so that a maximum principle applies, provided that $|\nu| < 1$ 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.

Remark 1.3.1 *Suppose that the region of interest is $0 \le x \le X$, 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 condition; this is a fundamental difference from the parabolic equations of Chapter 2, where we were always given a boundary condition at each end of the domain.*

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

$$\begin{aligned} T_j^n &:= \frac{u_j^{n+1} - u_j^n}{\Delta t} + a_j^n \frac{u_j^n - u_{j-1}^n}{\Delta x} \\ &\sim [u_t + \frac{1}{2}\Delta t u_{tt} + \cdots]_j^n + [a(u_x - \frac{1}{2}\Delta x u_{xx} + \cdots)]_j^n \\ &= \frac{1}{2}(\Delta t u_{tt} - a\Delta x u_{xx}) + \cdots . \end{aligned} \tag{1.29}$$

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

$$T_j^n = -\frac{1}{2}(1-\nu)a\Delta x u_{xx} + \cdots ; \tag{1.30}$$

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

$$e_j^{n+1} = (1-\nu)e_j^n + \nu e_{j-1}^n - \Delta t T_j^n \tag{1.31}$$

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

$$E^{n+1} := \max_j |e_j^{n+1}| \le E^n + \Delta t \max_j |T_j^n|. \tag{1.32}$$

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

$$|T_j^n| \le T \tag{1.33}$$

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

$$E^n \le n\Delta t T \le t_F T \tag{1.34}$$

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.

Remark 1.3.2 *Now let us consider a completely general set of values*

$$a_j^n := a(x_j, t_n); \quad j = 0, 1, \ldots, J. \tag{1.35}$$

It is clear that an equation similar to (1.31) holds at each point: if $a_j^n \geq 0$ and $j > 0$, then (1.31) 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.

Remark 1.3.3 ***We mentioned in Chapter 2 the difficulties which arise quite commonly in the analysis of parabolic equations when the given data function is discontinuous, or has discontinuous derivatives.*** *In that case the solution itself had continuous derivatives in the interior of the region, and our difficulty lay in finding bounds on the derivatives. For hyperbolic equations the situation is quite different. We have seen that the solution of our model problem is constant along the characteristics. Suppose that the initial function $u(x, 0) = u^0(x)$ has a jump discontinuity in the first derivative at $x = \xi$. Then clearly the solution $u(x, t)$ also has a similar discontinuity at all points on the characteristic passing through the point $(\xi, 0)$; the discontinuity is not confined to the boundary of the domain. Such a solution satisfies the differential equation everywhere except along the line of the discontinuity, while satisfying (1.5) everywhere; thus the latter can be regarded as defining a* ***generalised*** *solution of the differential equation, as distinct from a* ***classical solution*** *where the differential equation is satisfied at every point. Indeed, in this way we can define a solution for initial data $u^0(x)$ which itself has a jump discontinuity.*

Most practical problems for hyperbolic systems involve discontinuities of some form in the given data, or arising in the solution; for such problems the analysis given above in terms of the global truncation error and the maximum norm is of little use, since the derivatives involved do not exist everywhere in the domain. *For this reason, the truncation error is mainly of use for a local analysis, where in any case it would be better done by drawing the characteristic back to replace u_j^{n+1} by the value at its foot at time level t_n, i.e., $u(Q)$. The overall behaviour of a method, and its comparison with other methods, is often* ***more satisfactorily carried out by means of Fourier analysis, an analysis of its conservation properties or a modified equation analysis*** *- see later sections of this chapter and the next chapter.*

1.4 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

$$u(x, t) = e^{i(kx+\omega t)} \tag{1.36}$$

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

$$\omega = -ak. \tag{1.37}$$

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 (1.26) satisfies the upwind scheme provided that (1.22) holds. This leads to (1.23), showing that except in the special case $\nu = 1$ the mode is damped. The phase of the numerical mode is given by

$$\arg\lambda = -\tan^{-1}\left[\frac{\nu\sin k\Delta x}{(1-\nu)+\nu\cos k\Delta x}\right] \tag{1.38}$$

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

$$q \sim c_1 p + c_2 p^2 + c_3 p^3 + c_4 p^4 + \cdots \tag{1.39}$$

as $p \to 0$, then

$$\tan^{-1} q \sim c_1 p + c_2 p^2 + (c_3 - \frac{1}{3}c_1^3)p^3 + (c_4 - c_1^2 c_2)p^4 + \cdots . \tag{1.40}$$

We can now expand (1.38) and apply the lemma, giving

$$\begin{aligned} \arg\lambda &\sim -\tan^{-1}\left[\nu(\xi - \frac{1}{6}\xi^3 + \cdots)(1 - \frac{1}{2}\nu\xi^2 + \cdots)^{-1}\right] \\ &= -\tan^{-1}\left[\nu\xi - \frac{1}{6}\nu(1-3\nu)\xi^3 + \cdots\right] \\ &= -\nu\xi[1 - \frac{1}{6}(1-\nu)(1-2\nu)\xi^2 + \cdots], \end{aligned} \tag{1.41}$$

where we have written

$$\xi = k\Delta x. \tag{1.42}$$

Remark 1.4.2 *The case $\nu = 1$ is obviously very special, as the scheme then gives the exact result.*

Remark 1.4.3 *Apart from this, we have found that the upwind scheme always has an **amplitude error** which, from (1.23), is of order ξ^2 in one time step, corresponding to a global error of order ξ; and from (1.41) it has a **relative phase error** of order ξ^2, with the sign depending on the value of ν, and vanishing when $\nu = \frac{1}{2}$. The amplitude and relative phase errors are defined in more detail and plotted in Section 4.11, where they are compared with those of alternative schemes.*

Left as Homework.

Some results obtained with the upwind scheme are displayed in Fig. 1.4. The problem consists of solving the equation

$$u_t + a(x,t)u_x = 0, \quad x \geq 0, \quad t \geq 0, \tag{1.43}$$

where

$$a(x,t) = \frac{1+x^2}{1+2xt+2x^2+x^4}, \tag{1.44}$$

with the initial condition

$$u(x,0) = \begin{cases} 1 & \text{if } 0.2 \le x \le 0.4, \\ 0 & \text{otherwise,} \end{cases} \tag{1.45}$$

and the boundary condition

$$u(0,t) = 0. \tag{1.46}$$

The exact solution of the problem is

$$u(x,t) = u(x^*,0) \tag{1.47}$$

where

$$x^* = x - \frac{t}{1+x^2}. \tag{1.48}$$

Since $a(x,t) \le 1$ 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.

Remark 1.4.4 ***Important observation: we see from Fig. 1.4 that the numerical solution is getting smoothing for the discontinuous initial condition. One of the reason is that the consistency error is dissipative as seen in (1.29).***

1.5 The Lax–Wendroff scheme

1.5.1 one-step 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 $u(Q)$. 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 ν. 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[2] of hyperbolic conservation laws; it takes the form

$$U_j^{n+1} = \frac{1}{2}\nu(1+\nu)U_{j-1}^n + (1-\nu^2)U_j^n - \frac{1}{2}\nu(1-\nu)U_{j+1}^n \tag{1.49}$$

which may be written as (**formulation with artificial viscosity**)

$$U_j^{n+1} = U_j^n - \nu\Delta_0 x U_j^n + \frac{1}{2}\nu^2\delta_x^2 U_j^n, \tag{1.50}$$

[2] Lax, P.D. and Wendroff, B. (1960), Systems of conservation laws, *Comm. Pure and Appl. Math.* **13**, 217–37.

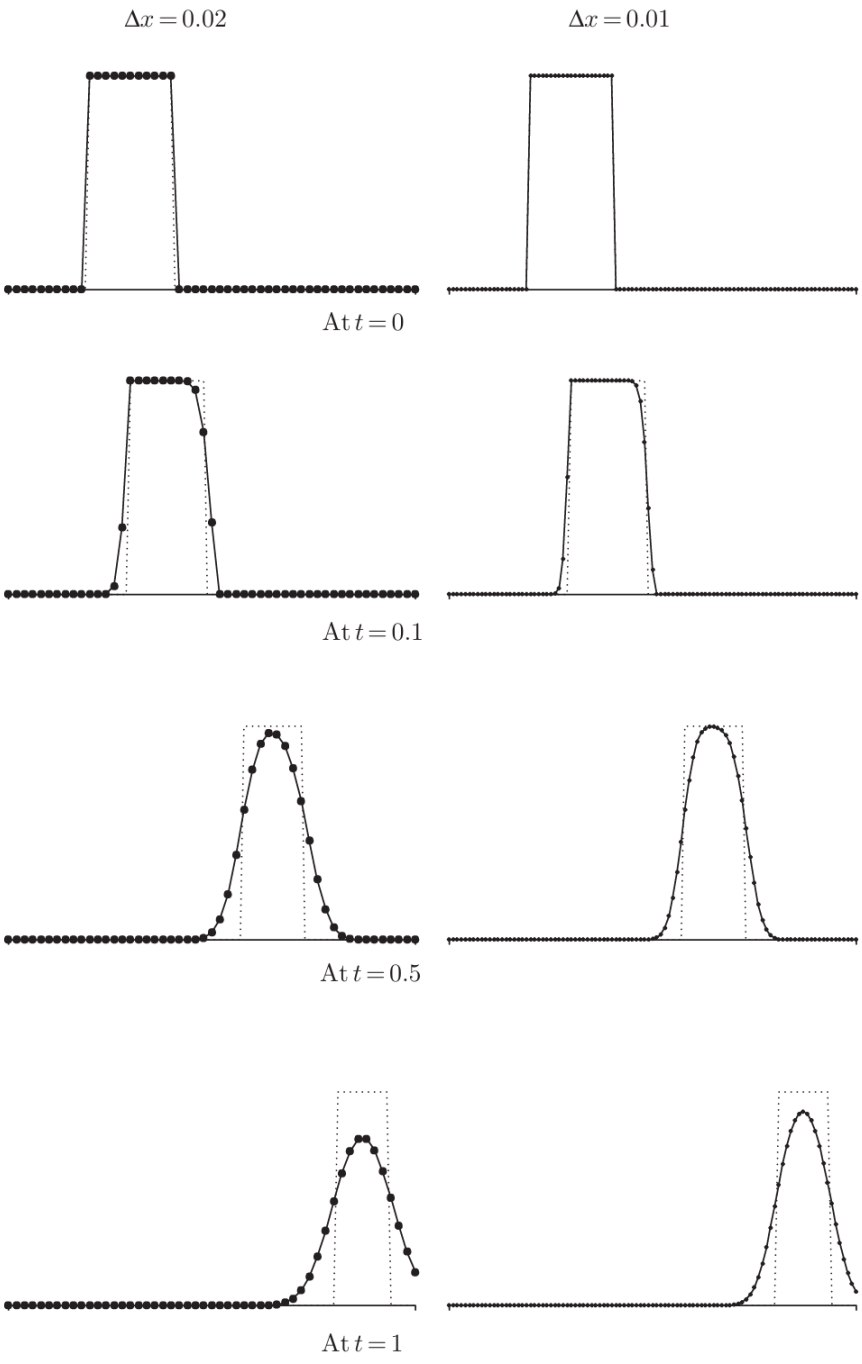


Fig. 4.6. Linear advection by the upwind method: problem (4.33), (4.34).

Figure 1.4: Linear advection by the upwind method.

where

$$\Delta_{0x}U(x,t) := \frac{1}{2}(\Delta_{+x} + \Delta_{-x})U(x,t) = \frac{1}{2}(U(x+\Delta x,t) - U(x-\Delta x,t)). \tag{1.51}$$

The usual Fourier analysis gives the amplification factor

$$\lambda(k) = 1 - i\nu \sin k\Delta x - 2\nu^2 \sin^2 \frac{1}{2}k\Delta x. \tag{1.52}$$

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

$$|\lambda|^2 = 1 - 4\nu^2(1-\nu^2)\sin^4 \frac{1}{2}k\Delta x. \tag{1.53}$$

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

$$\begin{aligned} \arg\lambda &= -\tan^{-1}\left[\tfrac{\nu \sin k\Delta x}{1-2\nu^2\sin^2\frac{1}{2}k\Delta x}\right] \\ &\sim -\nu\xi\left[1 - \frac{1}{6}(1-\nu^2)\xi^2 + \cdots\right]. \end{aligned} \tag{1.54}$$

Remark 1.5.1 *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 $\nu \sim 0$; but the error is always of one sign (corresponding to a phase lag) for Lax-Wendroff while it goes through a zero at $\nu = \frac{1}{2}$ for the upwind scheme. However,* ***the much smaller damping of the Lax-Wendroff scheme often outweighs the disadvantage of the larger phase error.***

Remark 1.5.2 ***Important observation: we see from Fig. 1.5 that the numerical solution is getting oscillatory for the discontinuous initial condition. One of the reason is that the consistency error is dispersive.***

Remark 1.5.3 *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*

$$u(x,t+\Delta t) = u(x,t) + \Delta t u_t(x,t) + \frac{1}{2}(\Delta t)^2 u_{tt}(x,t) + O((\Delta t)^3). \tag{1.55}$$

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

$$u_t = -au_x, \tag{1.56}$$

$$u_{tt} = -a_t u_x - au_{xt}, \tag{1.57}$$

$$u_{xt} = u_{tx} = -(au_x)_x, \tag{1.58}$$

which give

$$u_{tt} = -a_t u_x + a(au_x)_x. \tag{1.59}$$

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

$$U_j^{n+1} = U_j^n - a_j^n \Delta t \frac{\Delta_{0x}U_j^n}{\Delta x} + \frac{1}{2}(\Delta t)^2\left[-(a_t)_j^n \frac{\Delta_{0x}U_j^n}{\Delta x} + a_j^n \frac{\delta_x(a_j^n \delta_x U_j^n)}{(\Delta x)^2}\right]. \tag{1.60}$$

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 a_t at x_j. Note, however, that the scheme can be simplified by replacing $a_j^n + \frac{1}{2}\Delta t(a_t)_j^n$ 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 Fig. 1.5 are obtained by applying this scheme to the same problem (1.43)-(1.44), (1.45)-(1.46) used to test the upwind scheme, with the same mesh sizes. Comparing the results of Fig. 1.4 and Fig. 1.5 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 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 $O((\Delta x)^{1/2})$ for the upwind scheme and $O((\Delta x)^{2/3})$ for the Lax-Wendroff scheme. The error therefore tends to zero rather slowly as the mesh size is reduced.*

The oscillations in Fig. 1.5 arise because the Lax-Wendroff scheme does not satisfy a maximum principle. We see from (1.49) that with $\nu > 0$ the coefficient of U_{j+1}^n is negative, since we require that $\nu \leq 1$ 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, (1.43,1.44), but replace the initial condition (1.45)-(1.46) by

$$u(x,0) = \exp[-10(4x-1)^2]. \tag{1.61}$$

The results are illustrated in Fig. 1.6. 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.

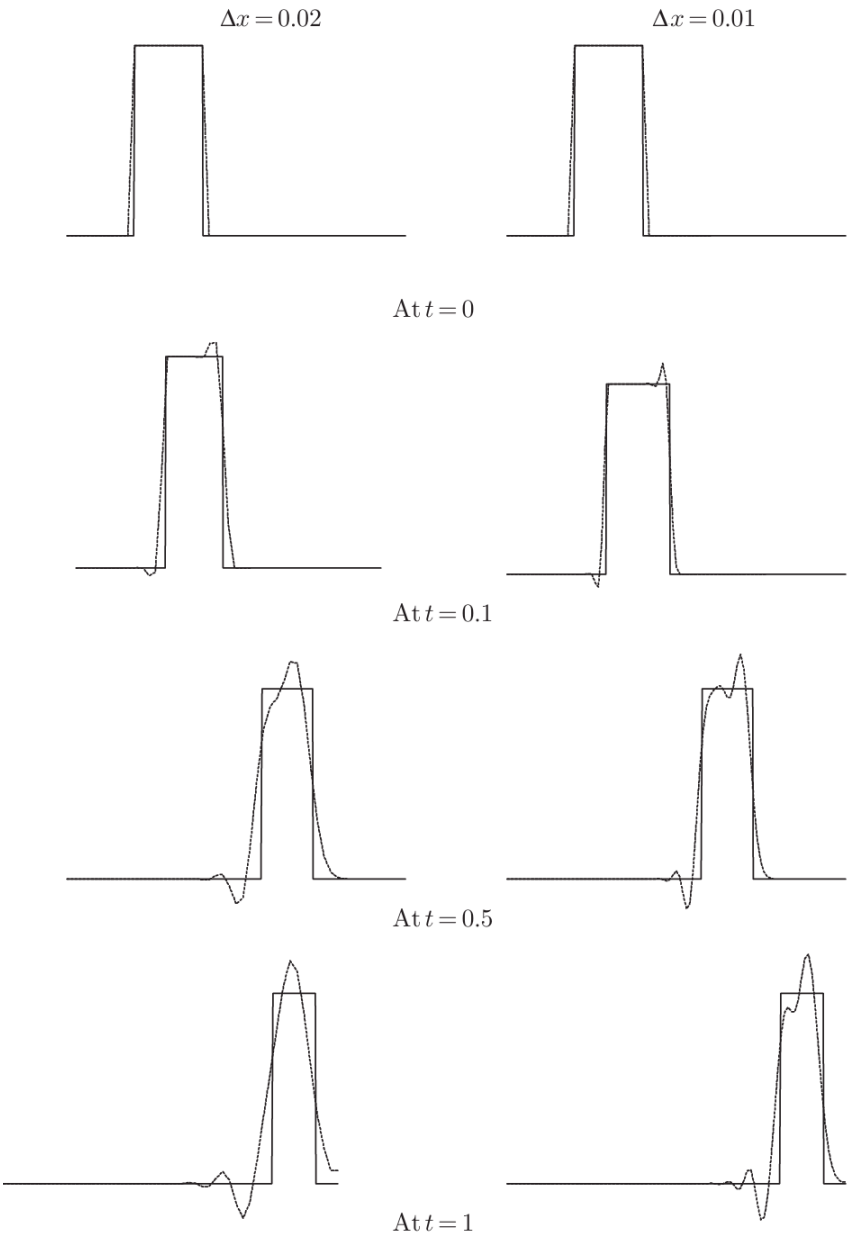


Fig. 4.7. Linear advection by the Lax–Wendroff method: problem (4.33), (4.34).

Figure 1.5: Linear advection by the Lax-Wendroff method.

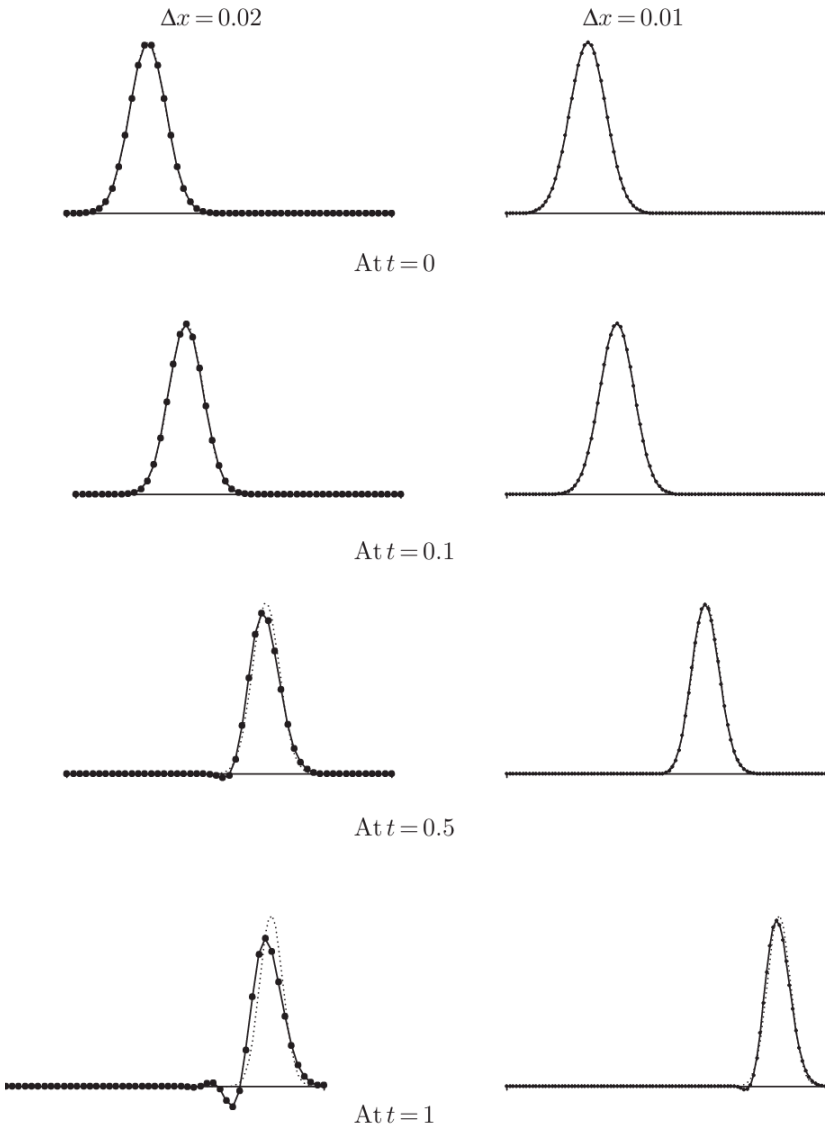


Fig. 4.8. Linear advection by the Lax–Wendroff method: (4.33) with the data (4.45).

Figure 1.6: Linear advection by the Lax-Wendroff method.

1.5.2 two-step Lax-Wendroff scheme

The Lax-Wendroff scheme can be derived in several ways. We shall derive it from a multi-step perspective. The idea is to compute (U_j^{n+1}) using not the time derivative at $(t = n\Delta t)$, but that at the half-step $(t = n\Delta t + \Delta t/2 = (n + 1/2)\Delta t)$

$$U_j^{n+1} = U_j^n + \Delta t \left(-a \frac{\partial U}{\partial x}\Big|_{j,n+1/2} \right)$$

To obtain the spatial derivative at the half-time step, we must have the function values at $t^{n+1/2}$, or

$$U_j^{n+1/2} = U_j^n + \frac{\Delta t}{2} \left(-a \frac{\partial U}{\partial x}\Big|_{j,n} \right).$$

The Lax-Wendroff method thus involves two steps:
1: First, compute $\frac{\partial U}{\partial x} |_{j,n+1/2}$ using central differences, that involve the mid-points $j + 1/2$ and $j - 1/2$:

$$\begin{aligned} U_{j-1/2}^{n+1/2} &= \frac{1}{2}(U_j^n + u_{j-1}^n) - a\frac{\Delta t}{2\Delta x}(U_j^n - U_{j-1}^n) \\ U_{j+1/2}^{n+1/2} &= \frac{1}{2}(U_{j+1}^n + u_j^n) - a\frac{\Delta t}{2\Delta x}(U_{j+1}^n - U_j^n) \end{aligned}$$

2: Compute U_j^{n+1} using the spatial derivative at $n + 1/2$:

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

Consistency, stability and convergence

The scheme is 2nd order both in time and space. To determine its stability we can express the scheme as:

$$U_j^{n+1} = \alpha U_{j-1}^n + \beta U_j^n + \gamma U_{j+1}^n$$

with

$$\begin{aligned} \alpha &= \frac{\nu}{2}(\nu + 1), \\ \beta &= 1 - \nu^2, \\ \gamma &= \frac{\nu}{2}(\nu - 1), \end{aligned}$$

where $\nu = \frac{a\Delta t}{\Delta x}$ is the CFL number. Assuming a solution of the type $\lambda^n e^{ik(m\Delta x)}$, the amplification factor is

$$\lambda = (1 + \nu^2(\cos k\Delta x - 1)) - i\nu \sin k\Delta x$$

which has a norm

$$|\lambda|^2 = 1 - \nu^2(1 - \nu^2)(1 - \cos k\Delta x)^2.$$

For the method to be stable, the condition is $|\lambda|^2 \leq 1$ which provides the following stability condition

$$1 - \nu^2 \geq 0 \Leftrightarrow \nu = \frac{a\Delta t}{\Delta x} \leq 1.$$

which is the well-known CFL condition.

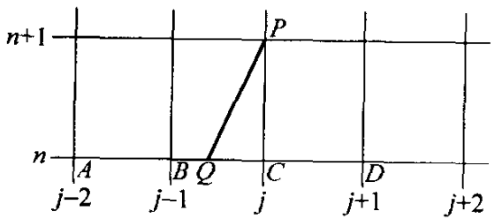


图 3.5

Figure 1.7: Construction of finite difference schemes using the characteristic line.

1.6 Construction of finite difference schemes using the characteristic line

利用偏微分方程的特征线来构造有限差分格式

特征线概念在双曲型方程中有很重要作用。借助于双曲型方程的解在特征线上为常数这一事实，可以构造出(1.1)式、(1.2)式的各种差分格式。为确定起见，假定$a > 0$。

设在$t = t_n$ 时间层上网格点A, B, C 和D 上的值已给定(已计算出的近似值或初值)。要计算出在$t = t_{n+1}$ 时间层上的网格点P 上的u 值，见图3.5。假定C, F, L 条件成立。那么过P 点特征线与BC 交于点Q。由微分方程解的性质知$u(P) = u(Q)$。但当Q 不是网格点时，$u(Q)$ 是未知的。由于$u(A), u(B), u(C)$ 和$u(D)$ 为$t = t_n$ 时间层上网格点上值已给定，因此可用插值方法给出$u(Q)$ 的近似值。利用B, C 两点上的值进行线性插值就可以得到

$$u(P) = u(Q) = (1 - a\nu)u(C) + a\nu u(B).$$

由此可推导出差分格式

$$U_j^{n+1} = U_j^n - a\nu(U_j^n - U_{j-1}^n),$$

其中$\nu = \frac{\tau}{h}$。这就是迎风格式。

如果改用B, D 两点进行线性插值。则有

$$u(P) = u(Q) = \frac{1}{2}(1 - a\nu)u(D) + \frac{1}{2}(1 + a\nu)u(B).$$

由此得到

$$U_j^{n+1} = \frac{1}{2}(1 - a\nu)U_{j+1}^n + \frac{1}{2}(1 + a\nu)U_{j-1}^n.$$

我们可以把此式改写为

$$U_j^{n+1} = \frac{1}{2}(U_{j-1}^n + U_{j+1}^n) - \frac{a\nu}{2}(U_{j+1}^n - U_{j-1}^n).$$

立即可以看出，这是Lax-Friedrichs 格式。

上面都是采用线性插值，当然我们也可以采用二次插值。如果使用B, C 和D 三个点进行抛物插值，则就得到

$$u(P) = u(Q) = u(C) - a\nu[u(C) - u(B)] - \frac{1}{2}a\nu(1 - a\nu)[u(B) - 2u(C) + u(D)].$$

由此得出差分格式

$$U_j^{n+1} = U_j^n - \frac{1}{2}a\nu(U_{j+1}^n - U_{j-1}^n) + \frac{1}{2}a^2\nu^2(U_{j+1}^n - 2U_j^n + U_{j-1}^n),$$

这就是Lax-Wendroff 格式。

如果我们不用B,C,D 三点进行插值，而是采用A,B,C 三点来进行抛物插值，可以得到

$$U_j^{n+1}=U_j^n-a\nu(U_j^n-U_{j-1}^n)-\frac{a\nu}{2}(1-a\nu)(U_j^n-2U_{j-1}^n+U_{j-2}^n),$$

此格式是二阶精度的。此格式由R. M. Beam 和R. F. Warming 于1976 年引入。因此一般称其为Beam-Warming 格式。这是二阶迎风格式。

为讨论格式(1.16)的稳定性，先求增长因子。

$$G(\tau,k)=1-2a\nu\sin^2(kh/2)-\frac{a\nu}{2}(1-a\nu)[4\sin^4\frac{kh}{2}-\sin^2 kh]-ia\nu\sin kh\left[1+2(1-a\nu)\sin^2\frac{kh}{2}\right],$$

$$|G(\tau,k)|^2=1-4a\nu(1-a\nu)^2(2-a\nu)\sin^4\frac{kh}{2}.$$

于是，当$a\nu\le 2$ 时有$|G(\tau,k)|\le 1$，由此推出，当$a\nu\le 2$ 时Beam-Warming 格式（1.16）是稳定的。

对于Beam-Warming 格式，当$a<0$ 时，格式变成

$$U_j^{n+1}=U_j^n+a\nu(U_{j+1}^n-U_j^n)-\frac{a\nu}{2}(1-a\nu)(U_{j+2}^n-2U_{j+1}^n+U_j^n). \tag{1.62}$$

仿上推导，（1.62）式的稳定性条件为$|a|\nu\le 2$。

由稳定性条件可以看出，对于固定空间步长，时间步长限制较宽。这有利于实际计算。

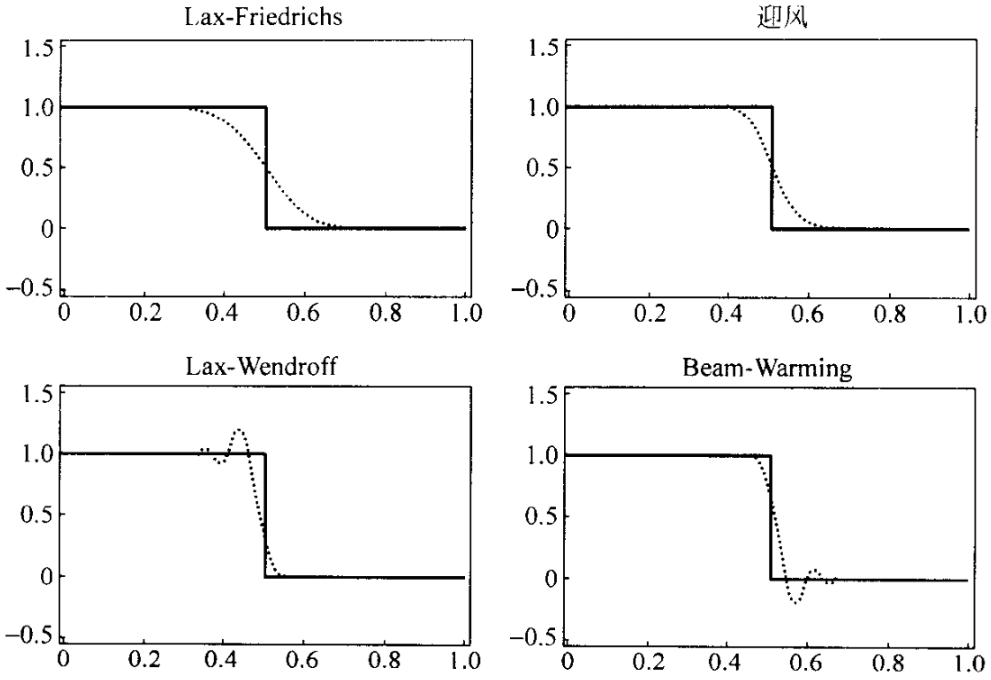


图 3.7

Figure 1.8: Comparison.

考虑初值问题

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, & x \in \mathbb{R},\, t \in [0,T], \\ u(x,0) = u_0(x), \end{cases}$$

其中

$$u_0(x) = \begin{cases} 1, & x \le 0, \\ 0, & x > 0. \end{cases}$$

取$h = 0.01, \lambda = \frac{1}{2}$，用Lax-Friedrichs 格式、迎风格式、Lax-Wendroff 格式以及Beam-Warming 格式，计算至$t_n = 0.5$ 时，计算结果与初值问题的解析解见图3.7[17]。对于前两个格式(Lax-Friedrichs 格式和迎风格式)把解抹平了。而后两个格式(Lax-Wendroff 格式和Beam-Warming 格式)出现了振荡。这些现象的出现是这些格式的正常现象。在拟线性双曲型方程组的间断解计算中为消去此类现象已研究出了很多良好的方法。

1.7 variable-coefficient equation

考虑简单的变系数方程的初值问题

$$\begin{cases} \frac{\partial u}{\partial t} + a(x,t)\frac{\partial u}{\partial x} = 0, & x \in \mathbb{R}, \quad 0 < t \le T. \\ u(x,0) = u_0(x), & x \in \mathbb{R}. \end{cases} \tag{1.63}$$

如果$a(x,t)$ 对x 和t 都是一次连续可微的，那么a 就光滑变化。情形与常系数相差不多。(1.63)式的特征线满足的方程为

$$\frac{dx}{dt} = a(x,t), \quad x(0) = x_0. \tag{1.64}$$

令$x = x(t, x_0)$ 和$u(x,t)$ 分别是方程(1.64) 和方程(1.63) 的解，那么

$$\frac{d}{dt}u(x(t,x_0),t) = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\cdot\frac{dx}{dt} = 0.$$

于是，方程(1.63) 的解沿特征线为常数。但我们要注意到此时的特征线是曲线(见图3.8)

$$u(x,t) = u_0(x_0), \quad x = x(t,x_0).$$

可以把常系数方程中推导的差分格式推广到变系数方程(1.63)。相应的Lax-Friedrichs 格式为

$$\frac{U_j^{n+1} - \frac{1}{2}(U_{j+1}^n + U_{j-1}^n)}{\tau} + a_j^n \frac{U_{j+1}^n - U_{j-1}^n}{2h} = 0, \tag{1.65}$$

其中$a_j^n = a(x_j, t_n)$.

(1.65)式是变系数差分格式,因此不能用Fourier 方法来讨论其稳定性。先采用能量不等式方法来讨论其稳定性。把(1.65)式改写为

$$U_j^{n+1} = \frac{1}{2}(U_j^n, + U_{j+1}^n) - \frac{1}{2}a_j^n\nu(U_{j+1}^n - U_{j-1}^n),$$

其中$\nu = \frac{\tau}{h}$。用U_j^{n+1} 乘上式两边

$$(U_j^{n+1})^2 = \frac{1}{2}(1 + a_j^n\nu)U_{j-1}^n U_j^{n+1} + \frac{1}{2}(1 - a_j^n\nu)U_{j+1}^n U_j^{n+1}.$$

假定网格比ν 满足条件

$$\max_j |a_j^n|\nu \le 1, \tag{1.66}$$

那么有

$$(U_j^{n+1})^2 \le \frac{1}{4}(1+a_j^n\nu)[(U_{j-1}^n)^2+(U_j^{n+1})^2]+\frac{1}{4}(1-a_j^n\nu)[(U_j^{n+1})^2+(U_{j+1}^n)^2]$$
$$=\frac{1}{4}(1+a_j^n\nu)(U_{j-1}^n)^2+\frac{1}{2}(U_j^{n+1})^2+\frac{1}{4}(1-a_j^n\nu)(U_{j+1}^n)^2.$$

从而得到

$$(U_j^{n+1})^2 \le \frac{1}{2}[(U_{j-1}^n)^2+(U_{j+1}^n)^2]+\frac{1}{2}a_j^n\nu[(U_{j-1}^n)^2-(U_{j+1}^n)^2].$$

用h 乘上式两边并对j 求和,记离散范数

$$\|U^n\|_h^2=\sum_{j=-\infty}^{\infty}(U_j^n)^2h,$$

那么有

$$\|U^{n+1}\|_h^2 \le \|U^n\|_h^2+\frac{1}{2}\sum_{j=-\infty}^{\infty}a_j^n\nu[(U_{j-1}^n)^2-(U_{j+1}^n)^2]h$$
$$=\|U^n\|_h^2+\frac{1}{2}\nu\sum_{j=-\infty}^{\infty}(a_{j+1}^n-a_{j-1}^n)(U_j^n)^2h.$$

如果

$$\left|\frac{\partial a}{\partial x}\right| \le M, \quad x\in\mathbb{R}, \quad t\in[0,T] \tag{1.67}$$

那么利用微分中值定理有

$$\left|a_{j+1}^n-a_{j-1}^n\right| \le 2Mh.$$

从而有

$$\|U^{n+1}\|_h^2 \le (1+m\tau)\|U^n\|_h^2.$$

重复使用上式有

$$\|U^n\|_h^2 \le e^{M\tau}\|U^0\|_h^2, \quad n\tau\le T.$$

这样我们证明了,当$a(x,t)$ 满足(1.67)式时,网格比满足条件(1.66),那么Lax-Friedrichs 格式稳定。

下面考虑初值问题(1.63)式、(1.63)式的迎风差分格式。与常系数情况的主要区别是$a(x,t)$ 是要变号的，因此不能要求利用一种形式写出，而是要随时考虑到$a(x,t)$ 的符号，这样可以写出差分格式

$$\frac{U_j^{n+1}-U_j^n}{\tau}+a_j^n\frac{U_j^n-U_{j-1}^n}{h}=0, \quad a_j^n>0. \tag{1.68}$$

$$\frac{U_j^{n+1}-U_j^n}{\tau}+a_j^n\frac{U_{j+1}^n-U_j^n}{h}=0, \quad a_j^n<0.$$

利用"冻结系数"方法得其稳定性条件为

$$\max_j|a_j^n|\nu\le 1.$$

在实际计算中也可把(1.68)式写成

$$\frac{U_j^{n+1}-U_j^n}{\tau}+a_j^n\frac{U_{j+1}^n-U_{j-1}^n}{2h}-\frac{1}{2h}|a_j^n|(U_{j+1}^n-2U_j^n+U_{j-1}^n)=0.$$

逼近(1.63)式、(1.63)式的Lax-Wendroff 格式需直接进行推导，由于系数依赖于x 和t，特别是依赖于t，推导得到的形式比较麻烦。

1.8 The Lax-Wendroff method for conservation laws

1.8.1 The Lax-Wendroff scheme

In practical situations a hyperbolic equation often appears in the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{1.69}$$

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

$$u_t + a u_x = 0, \tag{1.70}$$

where $a = a(u) = \partial f/\partial u$. It is then convenient to derive the Lax-Wendroff scheme directly for the conservation form (1.69). 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 (1.55) can now be written

$$u_t = -(f(u))_x \tag{1.71}$$

and

$$u_{tt} = -f_{xt} = -f_{tx} = -(a u_t)_x = (a f_x)_x. \tag{1.72}$$

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

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{\Delta x}\Delta_{0x} f(U_j^n) \\ &\quad + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 \delta_x \left[a(U_j^n)\delta_x f(U_j^n)\right], \end{aligned} \tag{1.73}$$

where the central difference $\delta_x f(U_j^n) = f(U_{j+1/2}^n) - f(U_{j-1/2}^n)$. It is clear that this reduces to (1.50) when $f(u) = au$ where a is constant.

Remark 1.8.1 *If we expand the last term in (1.73) we see that it involves the values of $a(U_{j-1/2}^n)$ and $a(U_{j+1/2}^n)$; in evaluating these we could set $a(U_{j\pm1/2}^n) := a(\frac{1}{2}(U_j^n + U_{j\pm1}^n))$, but a commonly used alternative is to replace them by $a(U_{j\pm1/2}^n) = \Delta_{\pm x} f(U_j^n)/\Delta_{\pm x} U_j^n$, reducing to $a(U_j^n)$ when $U_j^n = U_{j\pm1}^n$.*

Then writing

$$F_j^n = f(U_j^n)$$

and

$$A_{j\pm1/2}^n = a(U_{j\pm1/2}^n)$$

for either choice of the characteristic speeds, the scheme becomes

$$U_j^{n+1} = U_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left\{\left[1 - A_{j+1/2}^n \frac{\Delta t}{\Delta x}\right]\Delta_{+x} F_j^n + \left[1 + A_{j-1/2}^n \frac{\Delta t}{\Delta x}\right]\Delta_{-x} F_j^n\right\}. \tag{1.74}$$

1.8.2 Burgers' equation

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

$$u_t + uu_x = 0, \tag{1.75}$$

or in conservation form

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0. \tag{1.76}$$

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

$$u \equiv u(x,t) = u^0(x - tu(x,t)). \tag{1.77}$$

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 $dt/dx = 1/u^0(x_0)$ through the point $(x_0, 0)$, for each value of x_0.

This behaviour is a simple model for the formation of **shocks** in the flow of a gas. **Not only does the classical solution of (4.51) break down when the characteristics cross, but so too does the mathematical model of the physical situation. Viscosity becomes important in the steep gradients that occur and the full, viscous, Burgers' equation $u_t + u_{u_x} = \nu_v u_{xx}$ should be used. A thorough description of the situation is beyond the scope of this book but some key points are very pertinent to our emphasis on the use of the conservation law forms of equations.**

What one can hope to approximate beyond the point of breakdown is the limit of the solution to the viscous equation as the viscosity ν_v tends to zero. This will have a discontinuity, representing a shock, which for the conservation law $u_t + f_x = 0$ will move at the **shock speed**

$$S := \frac{[f(u)]}{[u]}, \tag{1.78}$$

where $[u]$ denotes the jump in the variable u; thus if the limiting value of u on the left is u_L and on the right is u_R, the shock moves at a speed

$$\frac{f(u_R) - f(u_L)}{u_R - u_L}, \tag{1.79}$$

which clearly tends to $a(u_L)$ as $u_R \to u_L$, i.e., in the limit of a 'weak' shock. Such a relation can be deduced by integrating the equation over a small box in the (x,t)-plane, aligned with and covering a portion of the shock, and then applying the Gauss divergence theorem. Solutions of the differential equation which are only satisfied in this averaged way are called **weak solutions**.

add shock PDE note

1.8.3 shock waves and rarefaction waves (激波与稀疏波)

方程假设

The equation is

$$u_t + (f(u))_x = 0, \qquad \text{or } u_t + a(u)u_x = 0,$$

where $a(u) = f'(u)$. f is a uniformly convex function and C^2, where uniformly convex function refers to that $f'' \geq f_0 > 0$ for some constant f_0.

特征线法

Example Consider

$$u_t + uu_x = 0. \tag{1.80}$$

The characteristic line is given by

$$\frac{dx(t)}{dt} = u(x,t).$$

Since equation (1.80) is nonlinear, the characteristic equation depends on the unknown function $u(x,t)$ itself! 每个$u(x,t)$ 会给出一组不同的特征线.

注意到

$$\frac{du(x(t),t)}{dt} = u_t + \frac{dx}{dt}u_x = u_t + uu_x = 0(!)$$

u沿特征线是常数，则

$$\frac{dx}{dt} = u(x(t),t) = constant!$$

(a) Each characteristic curve is a straight line. So each solution $u(x,t)$ has a family of straight lines (of various slopes) as its characteristics. 每条特征线都是直线，所以每个解$u(x,t)$ 有一族直线作为其特征线.

(b) The solution is constant on each such line. 沿每条直线解是常数.

(c) The slope of each such line is equal to the value of $u(x,t)$ on it. 每条直线的斜率是直线上的解$u(x,t)$ 的值.

Example The initial condition

$$u(x,0) = \phi(x).$$

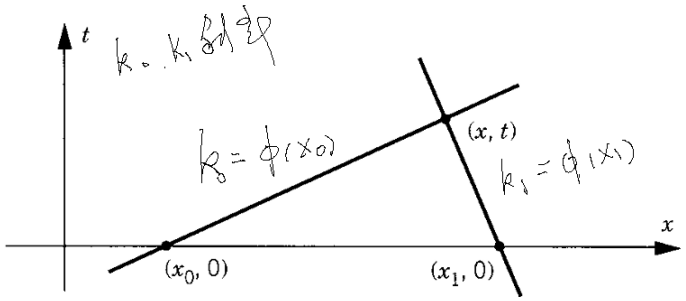


Figure 1

Figure 1.9: Lines with different slopes may intersect with each other.

三种解决办法：

(1) 规避所有相交的可能，当$\phi(x)$是x的增函数；

(2) 允许解的间断；

(3) 解存在非常局部的短时间内，而长时间行为未知。

一般方程

$$u_t + (f(u))_x = u_t + a(u)u_x = 0. \tag{1.81}$$

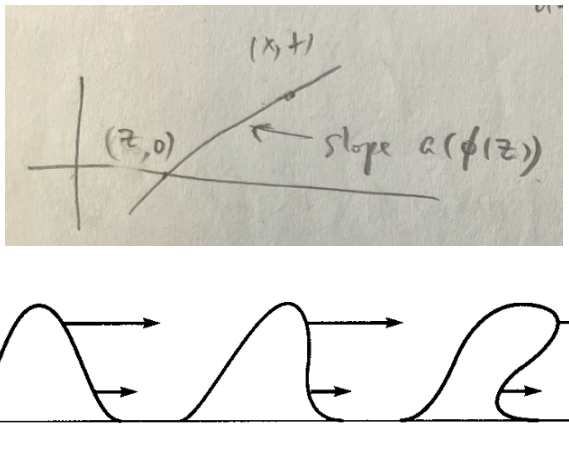


Figure 4

Figure 1.10: Wave may break.

$$\frac{dx}{dt} = a(u(x,t)).$$

$$\frac{d}{dt}u(x(t),t) = 0.$$

$$\frac{x-z}{t-0} = \frac{dx}{dt} = a(u(x,t)) = a(u(z,0)) = a(\phi(z)). \tag{1.82}$$

See upper panel of Fig. 1.10.

如果$a(\phi(z)) \le a(\phi(\omega))$，for all $z \le \omega$ ，则没有特征线相交，称为expansion wave or rarefaction wave（稀疏波）。

若有相关，则见bottom panel of 图1.10。

不连续解，激波，Rankine-Hugoniot condition

(1) $v \in D(Q) = C_0^\infty(Q)$, $Q = \mathbb{R} \times [0,\infty)$. v是试验函数(test function)，初始$u = \phi(x)$ at $t = 0$.

$$\begin{aligned} 0 &= \int_0^\infty dt \int_{-\infty}^\infty (u_t + (f(u)))_x v dx \\ &= -\int_0^\infty dt \int_{-\infty}^\infty u v_t dx - \int_{-\infty}^\infty uv dx\Big|_{t=0} - \int_0^\infty \int_{-\infty}^\infty f(u) v_x dx dt. \end{aligned}$$

则弱解满足

$$\int_0^\infty \int_{-\infty}^\infty (uv_t + f(u)v_x) dx dt + \int_{-\infty}^\infty \phi v dx\Big|_{t=0} = 0 \tag{1.83}$$

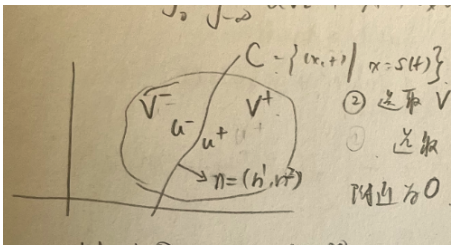


Figure 1.11: Domain for the discontinuous solution.

(2) 如图1.11，选取$V = V^- \cup V^+ \subset \mathbb{R} \times (0, \infty)$, $V^- \cap V^+ = C$. 选取v 在V^- 上有正交集. 且v 在∂V^- 附近为0. 则由(1.83),

$$0 = \int_0^\infty \int_{-\infty}^\infty (uv_t + f(u)v_x)dxdt = -\int_0^\infty \int_{-\infty}^\infty (u_t + (f(u))_x)vdxdt.$$

对任意在V^- 上有紧支集的试验函数v 成立，故

$$u_t + (f(u))_x = 0, \quad \text{in } V^-$$

同理

$$u_t + (f(u))_x = 0, \quad \text{in } V^+ \tag{1.84}$$

(3) Jump condition along shock.

选取在V 上有紧支集的试验v. 但v 在曲线C 上不一定为0. 由(1.83),

$$0 = \int_0^\infty \int_{-\infty}^\infty (uv_t + f(u)v_x)dxdt = \iint_{V^-} uv_t + f(u)v_x dxdt + \iint_{V^+} uv_t + f(u)v_x dxdt \tag{1.85}$$

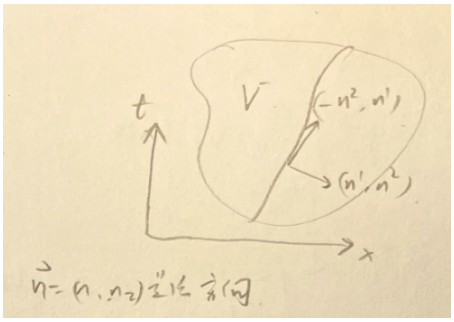


Figure 1.12: Domain for the discontinuous solution.

由Green公式，

$$\begin{aligned} \iint_{V^-} (uv_t + f(u))v_x dxdt &= -\iint_{V^-} \underbrace{[u_t + f(u)_x]}_{=0} vdxdt + \int_C (u^- n^2 + f(u^-)n^1)vdl \\ &= \int_C (u^- n^2 + f(u^-)n^1)vdl. \end{aligned}$$

这里, $n = (n^1, n^2)$是法方向。具体推导过程为，在标准Green公式中

$$\int Pdx + Qdt = \iint \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial t}\right) dxdt. \tag{1.86}$$

令

$$Q = fv, \quad P = -uv, \tag{1.87}$$

则得到

$$\iint \left(\frac{\partial(fv)}{\partial x} - \frac{\partial(-uv)}{\partial t} \right) dxdt = \int -uvdx + fvdt.$$

又$dl(-n^2, n^1) = (dx, dt)$, 则上式得到

$$\sim= \int (uvn^2 + fvn^1)dl. \tag{1.88}$$

同理，$\iint_{V^+}(uv_t + f(u)v_x)dxdt = -\int_C((u^+n^2 + f(u^+)n^1))vdl$. 综上可得，

$$\int_C \left[(u^-n^2 + f(u^-)n^1) - u^+n^2 - f(u^+)n^1 \right] vdl = 0. \tag{1.89}$$

$$\int_C \left[(f(u^-) - f(u^+))n^1 + (u^- - u^+)n^2 \right] vdl = 0. \quad \forall v \in D(V). \tag{1.90}$$

Then we have the line for discontinuity,

$$\Rightarrow C = \{(x,t)|x = s(t)\}. \quad (f(u^-) - f(u^+))n^1 + (u^- - u^+)n^2 = 0. \tag{1.91}$$

$$(n^1, n^2) \qquad \| \ (1, -\dot{s}). \tag{1.92}$$

$$(\dot{s}, 1) \qquad \| \ (-n^2, n^1). \tag{1.93}$$

$$f(u^-) - f(u^+) = (u^- - u^+)\dot{s}. \tag{1.94}$$

Thus, the velocity of the curve C is

$$\sigma = \dot{s} = \frac{f(u^-) - f(u^+)}{u^- - u^+}. \tag{1.95}$$

This is called **Rankine-Hugoniot formula**.

Remark 1.8.2 **虽然存在直接针对非守恒形式的数值格式，但对于包含激波（间断）的解，建议使用基于守恒形式的格式（如*Lax-Friedrichs ,Lax-Wendroff, Godunov, WENO*等）。这是因为只有守恒格式才能保证在网格细化时，其弱解收敛到满足*RH*跳跃条件（即正确的激波速度）的物理解。非守恒格式可能会计算出物理上不正确的激波。因此，除非我们只关心激波形成前的光滑解阶段，或者有特殊原因使用非守恒形式，否则在实践中，守恒格式是模拟*Burgers*方程（特别是激波）的标准和推荐。非守恒格式通常作为不能捕捉激波的反面教材。**

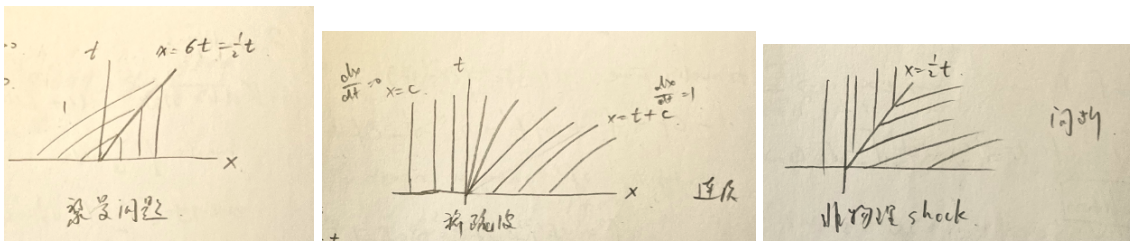


Figure 1.13: Domain for the discontinuous solution.

Example Consider the Riemann problem for $a(u) = u$, $f(u) = \frac{1}{2}u^2$. The initial condition is

$$\begin{aligned} \phi &= 0, \quad \text{for} \quad x > 0. \\ \phi &= 1, \quad \text{for} \quad x < 0. \end{aligned}$$

The speed of the discontinuity is

$$\sigma = \dot{s} = \frac{f(u^-) - f(u^+)}{u^- - u^+} = \frac{\frac{1}{2} \cdot 1^2 - \frac{1}{2} \cdot 0^2}{1 - 0} = \frac{1}{2} \tag{1.96}$$

so that the line is

$$x = \sigma t = \frac{1}{2}t \tag{1.97}$$

Example Problem The Burgers' equation is $a(u) = u$ with initial condition

$$\begin{aligned} \phi &= 1, \quad \text{for} \quad x > 0. \\ \phi &= 0, \quad \text{for} \quad x < 0. \end{aligned}$$

The first possible solution is

$$u_1 = \begin{cases} 1 & x > t \\ \frac{x}{t} & 0 < x < t \\ 0 & x < 0. \end{cases} \tag{1.98}$$

The second possible solution is

$$u_2 = \begin{cases} 0 & x < \frac{t}{2} \\ 1 & x > \frac{t}{2}. \end{cases} \tag{1.99}$$

Which one is physically correct? We could argue that the continuous one is preferred. But there may be other situations where neither one is continuous. Here both mathematicians and physicists are guided by the concept of entropy in gas dynamics. It requires that the wave speed just behind the shock is greater than the wave speed just ahead of it; that is, the wave behind the shock is "catching up" to the wave ahead of it. Mathematically, this means that on a shock curve we have the **Entropy Criterion**

$$a(u^-) > \sigma > a(u^+).$$

Notice that above is satisfied for Solution 1 but not for Solution 2. Therefore, Solution 2 is rejected. Finally, the definition of a shock wave is complete. Along its curves of discontinuity it must satisfy both RH formula and entropy criterion. For further discussion of shocks, see [Wh] or [Sm].

对于间断解，沿着间断曲线，要满足RH formula和entropy criterion.

Leave Example 3 on page 142 as Homework.

1.8.4 The upwind scheme

In Fig. 1.14 are shown early results obtained from the conservation law (1.76) using the initial data (1.61). A shock develops at $t_c = 0.25$ approximately and grows in strength, though the whole solution will

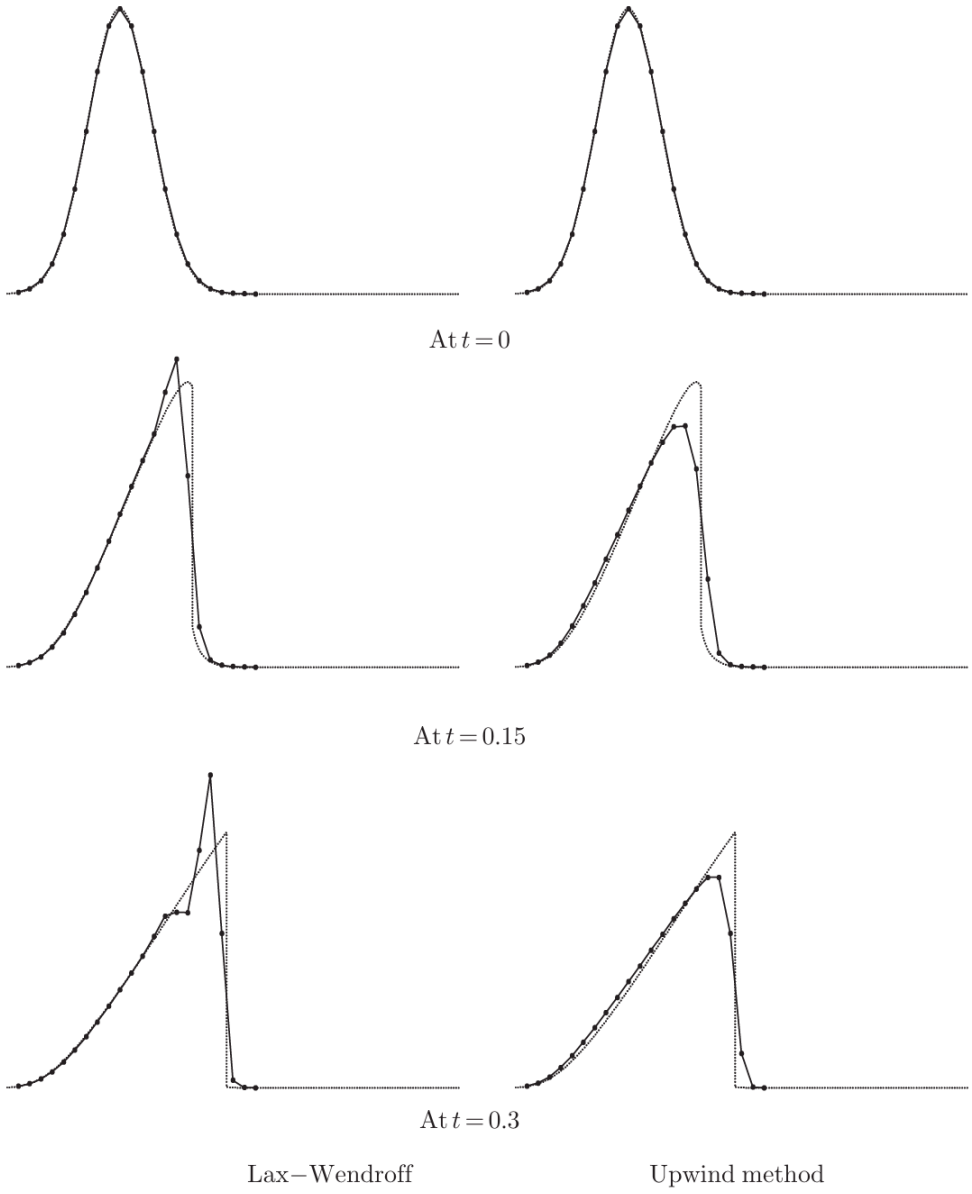


Fig. 4.10. Burgers' equation with initial data (4.45), approximated by the Lax–Wendroff method (on the left) and the upwind method (on the right).

Figure 1.14: Linear advection by the Lax-Wendroff method.

eventually decay to zero. This weak solution of the equation is shown as a dotted line, with its approximation by the Lax-Wendroff scheme shown on the left. For comparison we have also shown on the right of Fig. 1.14 the approximation obtained with the upwind scheme, which we write in the form

$$U_j^{n+1} = U_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left\{\left[1 - \text{sgn } A^n_{j+\frac{1}{2}}\right]\Delta_{+x}F_j^n + \left[1 + \text{sgn } A^n_{j-\frac{1}{2}}\right]\Delta_{-x}F_j^n\right\} \tag{1.100}$$

where the preferred choice is $A^n_{j+\frac{1}{2}} := \Delta_{\pm x}F_j^n/\Delta_{\pm x}U_j^n$, reducing to $a(U_j^n)$ when $U_j^n = U^n_{j\pm1}$; this form clearly generalises (1.21) and is directly comparable with (1.74). The greater accuracy of the Lax-Wendroff method away from the shock is apparent from these figures; but the oscillations that develop behind the shock, in contrast to their absence with the upwind method, prompt the idea of adaptively switching between the two schemes, as pioneered in the work of van Leer. [3]

1.8.5 systems of equations

One of the great strengths of the Lax-Wendroff method is that it can be extended quite easily to systems of equations. Instead of (1.69) and (1.71,1.72) we have

$$\mathbf{u}_t = -\mathbf{f}_x, \quad \mathbf{u}_{tt} = -\mathbf{f}_{tx} = -(\mathbf{A}\mathbf{u}_t)_x = (\mathbf{A}\mathbf{f}_x)_x \tag{1.101}$$

where A is the Jacobian matrix as in (1.11); then (1.73) simply becomes

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \left(\frac{\Delta t}{\Delta x}\right)\Delta_{0x}\mathbf{f}(\mathbf{U}_j^n) \tag{1.102}$$

$$+\frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 \delta_x\left[A(\mathbf{U}_j^n)\delta_x\mathbf{f}(\mathbf{U}_j^n)\right], \tag{1.103}$$

and (1.74) is exactly the same except for the use of vectors $\mathbf{U}_j^n$ and $\mathbf{F}_j^n$.

In the special case where $\mathbf{f}(\mathbf{u}) = A\mathbf{u}$ and A is a constant matrix, this reduces to

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \left(\frac{\Delta t}{\Delta x}\right)A\Delta_{0x}\mathbf{U}_j^n + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 A^2\delta_x^2\mathbf{U}_j^n. \tag{1.104}$$

For this problem a Fourier analysis is possible. Each of the components of the vector $\mathbf{U}$ will be a multiple of the same Fourier mode, and we look for a solution of (1.104) which is of the form

$$\mathbf{U}_j^n = \lambda^n e^{ikj\Delta x}\hat{\mathbf{U}} \tag{1.105}$$

where $\hat{\mathbf{U}}$ is a constant vector. This is a solution provided that

$$\left\{\lambda I - \left[I - i\left(\frac{\Delta t}{\Delta x}\right)\sin(k\Delta x)A - 2\left(\frac{\Delta t}{\Delta x}\right)^2\sin^2(\frac{1}{2}k\Delta x)A^2\right]\right\}\hat{\mathbf{U}} = 0. \tag{1.106}$$

For this to hold $\hat{\mathbf{U}}$ has to be an eigenvector of A; if μ is the corresponding eigenvalue we write $\nu = \mu\Delta t/\Delta x$ and obtain

$$\lambda = 1 - i\nu\sin k\Delta x - 2\nu^2\sin^2\frac{1}{2}k\Delta x \tag{1.107}$$

[3] van Leer, B. (1974), Towards the ultimate conservative difference scheme. II monotonicity and conservation combined in a second order scheme, *J. of Comput. Phys.* **14**, 361–70.

which is precisely the same as (1.52). Thus we can deduce a necessary condition for the scheme to be stable as

$$\frac{\rho \Delta t}{\Delta x} \leq 1 \tag{1.108}$$

where ρ is the **spectral radius** of A, the largest of the magnitudes of the eigenvalues of A, which is a generalisation of our earlier concept of stability to systems of equations. We leave until the next chapter a consideration of whether this necessary von Neumann condition is a sufficient condition for stability.

It is an important advantage of the Lax-Wendroff scheme that the stability condition involves only the magnitudes of the eigenvalues, not their signs, so that its form does not have to be changed with a switch in sign. **We have seen in (1.21) and (1.100) how the upwind scheme for a single equation uses either a forward or backward difference, according to the sign of a. This is much more difficult for a system of equations.** It requires that, at each point, we find the eigenvalues and eigenvectors of an approximate Jacobian matrix $\hat{A}$, express the current vector in terms of the eigenvectors, and use forward or backward differences for each eigenvector according to the sign of the corresponding eigenvalue; the eigenvectors are then combined together again to give the solution at the new time level. **The Lax-Wendroff method avoids this considerable complication,but at the cost of the oscillations shown in Figs. 1.5 and 1.14.** There has thus been considerable development of methods which combine the advantages of the two approaches, as has already been mentioned and as we will discuss further in the next section.

1.8.6 two-step Lax Wendroff method

Lastly, a convenient and often-used variant of the Lax-Wendroff scheme is the two-step method. Values are predicted at points $(x_{j+1/2}, t_{n+1/2})$ and then a centred scheme used to obtain the final values at t_{n+1} (see Fig. 1.3):

$$\mathbf{U}_{j+1/2}^{n+1/2} = \frac{1}{2}\left(\mathbf{U}_j^n + \mathbf{U}_{j+1}^n\right) - \frac{1}{2}(\Delta t/\Delta x)\left[\mathbf{f}\left(\mathbf{U}_{j+1}^n\right) - \mathbf{f}\left(\mathbf{U}_j^n\right)\right], \tag{1.109}$$

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - (\Delta t/\Delta x)\left[\mathbf{f}\left(\mathbf{U}_{j+1/2}^{n+1/2}\right) - \mathbf{f}\left(\mathbf{U}_{j-1/2}^{n+1/2}\right)\right]. \tag{1.110}$$

For the linear case where $\mathbf{f} = A\mathbf{u}$ with constant A, we leave it as an exercise to show that on elimination of the intermediate value $\mathbf{U}_{j+1/2}^{n+1/2}$ we obtain exactly the same result as the standard one-step Lax-Wendroff method (1.104). **For nonlinear problems and variable coefficients the two variants do not yield the same results. One of the most important advantages of (1.109,1.110) is that it avoids the need to calculate the Jacobian matrix A.**

1.9 Finite volume schemes

1.9.1 basic idea of finite volume method

Many of the methods that are used for practical computation with conservation laws are classed as **finite volume methods**, and that in (1.109,1.110) is a typical example. Suppose we take the system of equations

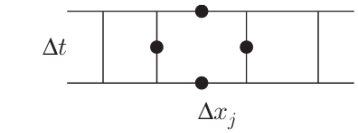
$\mathbf{u}_t + \mathbf{f}_x = 0$ in conservation law form and integrate over a region Ω in (x,t)-space; using the Gauss divergence theorem this becomes a line integral,

$$\iint_\Omega (\mathbf{u}_t + \mathbf{f}_x) dxdt \equiv \iint_\Omega \mathrm{div}(\mathbf{f}, \mathbf{u}) dxdt \tag{1.111}$$

$$= \oint_{\partial\Omega} [\mathbf{f} dt - \mathbf{u} dx]. \tag{1.112}$$

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

$$(\mathbf{u}_{\mathrm{top}} - \mathbf{u}_{\mathrm{bottom}})\Delta x + (\mathbf{f}_{\mathrm{right}} - \mathbf{f}_{\mathrm{left}})\Delta t = 0. \tag{1.113}$$



(a)

Figure 1.15: Finite volume scheme with mid-point 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 Fig. 1.15: if we denote by $\mathbf{U}_j^n$ 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

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - (\Delta t/\Delta x_j)\left(\mathbf{F}_{j+1/2}^{n+1/2} - \mathbf{F}_{j-1/2}^{n+1/2}\right). \tag{1.114}$$

It remains to calculate the fluxes from the set of $\mathbf{U}_j^n$ values, for example by the Taylor expansion used in the two-step Lax-Wendroff method: that is, solution values on the cell sides are calculated by the formula (1.109) and these are substituted into equation (1.110), which is exactly of the form (1.114).

Note, however, that in (1.114) 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 (1.114)

$$\sum_{j=k}^{l} \Delta x_j \left(\mathbf{U}_j^{n+1} - \mathbf{U}_j^n\right) + \Delta t \left(\mathbf{F}_{l+1/2}^{n+1/2} - \mathbf{F}_{k-1/2}^{n+1/2}\right) = 0, \tag{1.115}$$

which exactly mirrors the conservation property of the differential equation. In the case of the Lax-Wendroff scheme, though, if $\mathbf{U}_j^n$ 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,

$$\mathbf{u}(x_{j+1/2}, t_n + \Delta t/2) = \mathbf{u}(x_{j+1/2}, t_n) - \frac{1}{2}\Delta t f_x(x_{j+1/2}, t_n) + O((\Delta t)^2); \tag{1.116}$$

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

$$\mathbf{U}_{j+1/2}^{n+1/2} = \frac{\Delta x_{j+1}\mathbf{U}_j^n + \Delta x_j \mathbf{U}_{j+1}^n - \Delta t\left[f\left(\mathbf{U}_{j+1}^n\right) - f\left(\mathbf{U}_j^n\right)\right]}{\Delta x_j + \Delta x_{j+1}} \tag{1.117}$$

which generalises (1.109) to a general mesh. We will normally avoid this extra complexity in what follows and revert to our usual assumption of uniform mesh spacing.

1.9.2 other finite volume schemes

p 149 of Lu Jinfu

1. 守恒律和Riemann 问题

一般地考虑

$$\frac{\partial u}{\partial t}+\frac{\partial f(u)}{\partial x}=0,\quad t>0,\quad x\in\mathbb{R}, \tag{1.118}$$

此方程称为(单个)守恒律。初始条件为

$$u(x,0)=u_0(x),\quad x\in\mathbb{R}, \tag{1.119}$$

那么称(1.118)、(1.119)式为守恒律的初值问题。设u 为初值问题(1.118)、(1.119)式的解，并设f 为u 的连续可微函数。令

$$a(u)=f'(u), \tag{1.120}$$

那么(1.118)式可以写成

$$\frac{\partial u}{\partial t}+a(u)\frac{\partial u}{\partial x}=0. \tag{1.121}$$

考虑守恒律(1.118)式及初值

$$u(x,0)=u_0(x)=\begin{cases}u_l, & x<0,\\ u_r, & x>0.\end{cases} \tag{1.122}$$

这样的初值问题称为**Riemann 问题**。

仅对$f''(u)>0$ 的情况进行讨论。

如果u 是初值问题(1.118)和(1.122)式的解，那么对任意常数$\lambda>0$，由

$$u_\lambda(x,t)=u(\lambda x,\lambda t) \tag{1.123}$$

定义的函数u_λ 也是初值问题(1.118)和(1.122)式的解。由于我们是求惟一解，所以仅考虑依赖于$\xi=\frac{x}{t}$ 的解就可以了。这种类型的解称为**自相似解或自型解**。

Riemann 问题(1.118)和(1.122)式的解可分3 种情况进行讨论。

(1) $u_l=u_r$，此时$u=u_r=u_l$ 是古典解（连续可微解）。

(2) $u_l>u_r$，由于$f''>0$，所以$a(u)$ 单调增加，因此有$a(u_l)>a(u_r)$。此时解是联结u_l 和u_r 的速度为s的**激波**，其表达式可以写为

$$u(x,t)=\begin{cases}u_l, & \text{当}x<st,\\ u_r, & \text{当}x>st.\end{cases} \tag{1.124}$$

激波图示见图1.16(a)。

(3) $u_r>u_l$，此时不产生激波，出现连续解，称其为**稀疏波**。由于$f''>0$，我们有$a(u_r)>a(u_l)$，方程$\eta=a(u(\eta))$ 有解$u(\eta)$，其中$a(u_r)>\eta>a(u_l)$。稀疏波的具体表达式是

$$u(x,t)=\begin{cases}u_l, & \text{当}\frac{x}{t}<a(u_l),\\ a^{-1}\left(\frac{x}{t}\right), & \text{当}a(u_l)<\frac{x}{t}<a(u_r),\\ u_r, & \text{当}\frac{x}{t}>a(u_r),\end{cases} \tag{1.125}$$

其中$a^{-1}\left(\frac{x}{t}\right)$ 由$a(a^{-1}(\eta))=\eta$ 所确定。稀疏波图示见图1.16(b)。

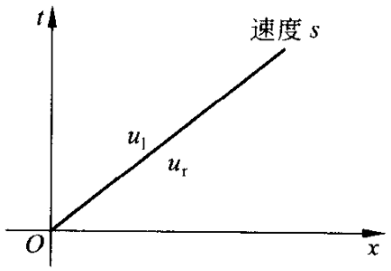


图 6.4

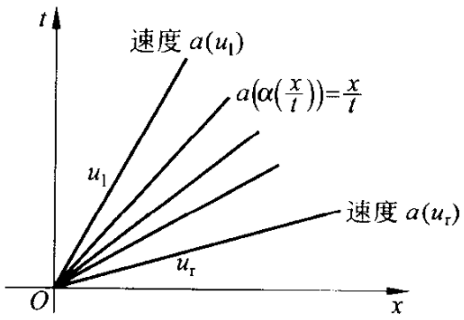


图 6.5

Figure 1.16: Shock wave and rarefaction wave.

2. Lax-Friedrichs 差分格式

设$\mathcal{D}$ 为x-t 平面上一有界区域，对守恒律(1.118)式在$\mathcal{D}$ 上积分并用Green 公式有

$$\iint_{\mathcal{D}} \left(\frac{\partial u}{\partial t}+\frac{\partial f(u)}{\partial x}\right) dxdt = \int_{\Gamma}(fdt-udx)=0, \tag{1.126}$$

其中$\Gamma=\partial\mathcal{D}$，即$\mathcal{D}$ 的边界。为推导差分格式，取$\mathcal{D}=\{(x,t)|x_{j-1}\le x\le x_{j+1}, t_n\le t\le t_{n+1}\}$，并令节点$A(x_{j+1},t_n), B(x_{j+1},t_{n+1}), C(x_{j-1},t_{n+1}), D(x_{j-1},t_n)$ 见图1.17。

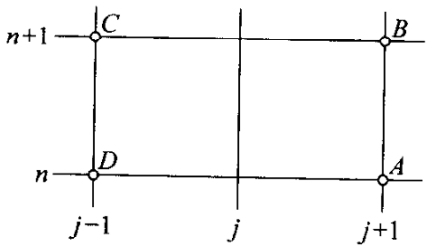


图 6.6

Figure 1.17: Lax-Friedrichs scheme.

Γ 为矩形$ABCD$ 的边界，那么

$$\int_{\Gamma}(fdt-udx)=\int_{DA}(-u)dx+\int_{BC}(-u)dx+\int_{AB}f(u)dt+\int_{CD}f(u)dt, \tag{1.127}$$

上式右端采用数值积分来近似。第1 个积分用梯形公式，第2 个积分用中矩形公式，第3、第4 个积分用下矩形公式，这样可以得到逼近守恒律的Lax-Friedrichs 格式

$$\frac{u_j^{n+1}-\frac{1}{2}(u_{j-1}^n+u_{j+1}^n)}{\tau}+\frac{f(u_{j+1}^n)-f(u_{j-1}^n)}{2h}=0. \tag{1.128}$$

在双曲型方程的求解中网格比$\lambda=\frac{\tau}{h}$，Lax-Friedrichs 格式(1.128)的截断误差为$O(\tau+h)$。(1.128)式为非线性差分格式。对稳定性不做严格讨论，而采用线性化稳定性分析方法来粗略讨论(1.128)式的稳定性。先把守恒律(1.118)式写成非守恒形式$\frac{\partial u}{\partial t}+a(u)\frac{\partial u}{\partial x}=0$，再设$a(u)$ 与u 无关，那么差分格式(1.128)化为

$$\frac{u_j^{n+1}-\frac{1}{2}(u_{j+1}^n+u_{j-1}^n)}{\tau}+a\frac{u_{j+1}^n-u_{j-1}^n}{2h}=0. \tag{1.129}$$

此差分格式在第3章中讨论过，其稳定性条件为$|a|\lambda \le 1$，类似于冻结系数方法可以得到

$$\max_j |a(u_j^n)|\lambda \le 1. \tag{1.130}$$

此条件为格式(1.128)的稳定性条件。在实际计算中，上述条件还需增加一些安全系数，如可取

$$\max_j |a(u_j^n)|\lambda \le 1-\epsilon, \tag{1.131}$$

其中ϵ 为一个正的小数。从稳定性条件(1.130)可以看出，这个条件不但依赖于网格点的位置，而且还依赖于初值问题的解u。由此可知，使用较为复杂，一般可采用试算的办法来确定稳定性。

3. 守恒型差分格式

回忆下，Lax-Friedrichs 格式(1.128)可以改写为

$$u_j^{n+1} = u_j^n - \lambda(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n), \tag{1.132}$$

其中

$$g_{j+\frac{1}{2}}^n = g(u_j^n, u_{j+1}^n) = -\frac{1}{2\lambda}(u_{j+1}^n - u_j^n) + \frac{1}{2}[f(u_j^n) + f(u_{j+1}^n)], \tag{1.133}$$

并注意到有

$$g(\omega,\omega) = f(\omega), \quad \forall \omega \in \mathbb{R}. \tag{1.134}$$

此情形以后将常用到。

下面定义逼近守恒律(1.118)式的守恒型差分格式。

定义 2.1 称差分格式

$$u_j^{n+1} = u_j^n - \lambda(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n) \tag{1.135}$$

是**守恒型差分格式**，其中

$$g_{j+\frac{1}{2}}^n = g(u_{j-l+1}^n, u_{j-l+2}^n, \cdots, u_{j+r}^n), \tag{1.136}$$

并称其为数值通量。

为了使守恒型差分格式(1.135)与守恒律(1.118)是相容的，g 必须满足

$$g(\omega,\omega,\cdots,\omega) = f(\omega), \quad \omega \in \mathbb{R}. \tag{1.137}$$

相容性条件(1.137)在一定意义下反映了数值通量g 和物理通量f 的相容性。差分格式(1.135)可以推广为

$$u_j^{n+1} = u_j^n - \lambda[\theta(g_{j+\frac{1}{2}}^{n+1} - g_{j-\frac{1}{2}}^{n+1}) + (1-\theta)(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n)], \quad 0 \le \theta \le 1. \tag{1.138}$$

当$\theta = 0$ 时，(1.138)式即为(1.135)式，是显式格式；当$\theta = 1$ 时，(1.138)式为隐式格式。

守恒型差分格式是由守恒律推导而得到，即为守恒律的离散化。它与守恒律一样，可以保持物理量的某种守恒性质。

由前面推导可知，Lax-Friedrichs 格式是相容的守恒型差分格式。守恒型格式的收敛性以及收敛到物理解等问题在此不做讨论了，有兴趣的读者可参考文献[8,16]。下面仅举常用守恒型差分格式的例子。

例 2.1 迎风格式

在守恒律(1.118)中，假定$a(u)=f'(u)\geq 0$，那么构造差分格式

$$u_j^{n+1}=u_j^n-\lambda[f(u_j^n)-f(u_{j-1}^n)] \tag{1.139}$$

此格式称为迎风格式，$g_{j+\frac{1}{2}}^n=g(u_j^n,u_{j+1}^n)=f(u_j^n)$，相容性是显然的。

例 2.2 Engquist-Osher 格式

在守恒律(1.118)中，令

$$\chi(u)=\begin{cases}1, & \text{如果}f'(u)>0,\\ 0, & \text{如果}f'(u)<0.\end{cases} \tag{1.140}$$

并定义

$$f_+(u)=\int_0^u\chi(s)f'(s)ds, \tag{1.141}$$

$$f_-(u)=\int_0^u(1-\chi(s))f'(s)ds. \tag{1.142}$$

那么，Engquist-Osher 格式定义为

$$u_j^{n+1}=u_j^n-\lambda[\Delta_+f_-(u_j^n)+\Delta_-f_+(u_j^n)], \tag{1.143}$$

其中$\Delta_+f_j=f_{j+1}-f_j$; $\Delta_-f_j=f_j-f_{j-1}$. 在(1.136)中取$l=1$,

$$g_{j+\frac{1}{2}}^n=g(u_j^n,u_{j+1}^n)=f_-(u_{j+1}^n)+f_+(u_j^n), \tag{1.144}$$

$$\begin{aligned}g(\omega,\omega) &= f_-(\omega)+f_+(\omega)\\ &= \int_0^\omega\chi(s)f'(s)ds+\int_0^\omega(1-\chi(s))f'(s)ds\\ &= \int_0^\omega f'(s)ds=f(\omega).\end{aligned}$$

因此Engquist-Osher 格式是相容的守恒律差分格式。其实此格式可看作迎风格式的改进。

例 2.3 Godunov 格式

假定在$t=t_n$ 这个时间层上$u_j^n(j=0,\pm1,\cdots)$ 已经求出，定义

$$v^n(x)=u_j^n,\quad x\in I_{j-\frac{1}{2}}=(x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}). \tag{1.145}$$

显然，$v^n(x)$ 是一个阶梯函数。令

$$u(x,t_n)=v^n(x), \tag{1.146}$$

那么(1.118)与(1.145)式构成了一个初值问题。为把这个初值问题的解从t_n 推进到t_{n+1}，首先必须求解这个初值问题。在每个区间$[x_j,x_{j+1}]$上，这个初值问题定义了一个局部Riemann问题，因此(1.118)、(1.145)式定义了一系列的（局部）Riemann问题。利用本章第1节的讨论，知其解仅依赖于$\frac{x}{t}$、u_l 和u_r，因此可把解表示成

$$u(x,t)=R\left(\frac{x}{t};u_l,u_r\right). \tag{1.147}$$

如果

$$\sup_j |a(u^n_{j+\frac{1}{2}})|\lambda \le \frac{1}{2}, \tag{1.148}$$

那么每个（局部）Riemann问题的解是不相交的。利用(1.147)式，可以得到初值问题(1.118)和(1.145)式的解为

$$u_n(x,t) = R\left(\frac{x - x_{j+\frac{1}{2}}}{t - t_n}; u^n_j, u^n_{j+1}\right), x \in I_j, t \in [t_n, t_{n+1}], \tag{1.149}$$

其中$I_j = [x_j, x_{j+1}]$。为了得到t_{n+1}时刻的$u^{n+1}_j, j = 0, \pm 1, \cdots$。取

$$u^{n+1}_j = \frac{1}{h}\int_{I_{j-\frac{1}{2}}} u_n(x, t_n + \tau)dx. \tag{1.150}$$

注意到(1.149)式定义的$u_n(x,t)$ 为初值问题(1.118)、(1.145)式的解，因此满足积分关系式。所以在

$$D = \{(x,t)|x_{j-\frac{1}{2}} \le x \le x_{j+\frac{1}{2}}, t_n \le t \le t_{n+1}\} \tag{1.151}$$

上积分有

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_n(x,t_n)dx - \int_{t_n}^{t_{n+1}} f[u_n(x_{j+\frac{1}{2}},t)]dt - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_n(x,t_{n+1})dx + \int_{t_n}^{t_{n+1}} f[u_n(x_{j-\frac{1}{2}},t)]dt = 0. \tag{1.152}$$

注意到

$$u_n(x_{j+\frac{1}{2}},t) = R(0; u^n_j, u^n_{j+1}), \quad u_n(x_{j-\frac{1}{2}},t) = R(0; u^n_{j-1}, u^n_j) \tag{1.153}$$

是不依赖于t 的，所以有

$$\int_{t_n}^{t_{n+1}} f[u_n(x_{j+\frac{1}{2}},t)]dt = f[R(0; u^n_j, u^n_{j+1})]\tau, \tag{1.154}$$

$$\int_{t_n}^{t_{n+1}} f[u_n(x_{j-\frac{1}{2}},t)]dt = f[R(0; u^n_{j-1}, u^n_j)]\tau. \tag{1.155}$$

由此得

$$u^{n+1}_j = u^n_j - \lambda(f^n_{j+\frac{1}{2}} - f^n_{j-\frac{1}{2}}), \tag{1.156}$$

其中

$$f^n_{j+\frac{1}{2}} = f(u^n_{j+1/2}) = f[R(0; u^n_j, u^n_{j+1})]. \tag{1.157}$$

其中$f^n_{j+\frac{1}{2}} = f(u^n_{j+1/2})$, $u^n_{j+\frac{1}{2}} = R(0; u^n_j, u^n_{j+1})$. 差分格式(1.156) 称为**Godunov 差分格式**。

例 2.4 Roe 方法

在Godunov 方法中要求解真实的Riemann 问题,这是很费时间的,因此近似地求解Riemann 问题很重要。Roe 方法为此而建立。在Roe 方法中不是求Riemann 问题

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \\ u(x,0) = u_0(x), \end{cases} \tag{1.158}$$

其中

$$u_0(x)=\begin{cases}u_l, & x<0,\\ u_r, & x>0,\end{cases} \tag{1.159}$$

而是求解相应的局部线性化问题。即用局部线性化方程

$$\frac{\partial u}{\partial t}+a(u_l,u_r)\frac{\partial u}{\partial x}=0 \tag{1.160}$$

来替代(单个)守恒律,其中$a(u_l,u_r)$ 满足

$$f(u_r)-f(u_l)=a(u_l,u_r)(u_r-u_l). \tag{1.161}$$

线性化的Riemann 问题的解可以表示为

$$r\left(\frac{x}{t};u_l,u_r\right)=\begin{cases}u_l, & \frac{x}{t}<s\\ u_r, & \frac{x}{t}>s\end{cases} \tag{1.162}$$

其中s 为间断传播速度,事实上$s=a(u_l,u_r)$。(1.162)式给出了非线性Riemann 问题的近似解。

Roe 方法与Godunov 方法差别在于一个求解近似Riemann 问题,一个求解真实Riemann 问题。显然,Roe 方法很容易,但在计算中会出现非物理解。

4. 数值例子

采用初值问题

$$\begin{cases}\frac{\partial u}{\partial t}+\frac{\partial}{\partial x}\left(\frac{u^2}{2}\right)=0,\\ u(x,0)=u_0(x).\end{cases} \tag{1.163}$$

其中

$$u_0(x)=\begin{cases}1, & 若x<0.3,\\ 0, & 若x\geq 0.3.\end{cases} \tag{1.164}$$

用Lax-Friedrichs 格式，Engquist-Osher 格式，Godunov 格式以及Lax-Wendroff 格式的计算结果见图1.18。

由计算结果可以看出，对于Lax-Friedrichs，Engquist-Osher，Godunov 格式把间断的过渡区域拉得过宽，特别地，Lax-Friedrichs 格式更严重。这是一阶格式的普遍问题。对于Lax-Wendroff 格式，在间断附近存在不应有的振荡，这也是二阶精度格式普遍存在的问题。

1.9.3 total variation diminishing

1. total variation diminishing

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

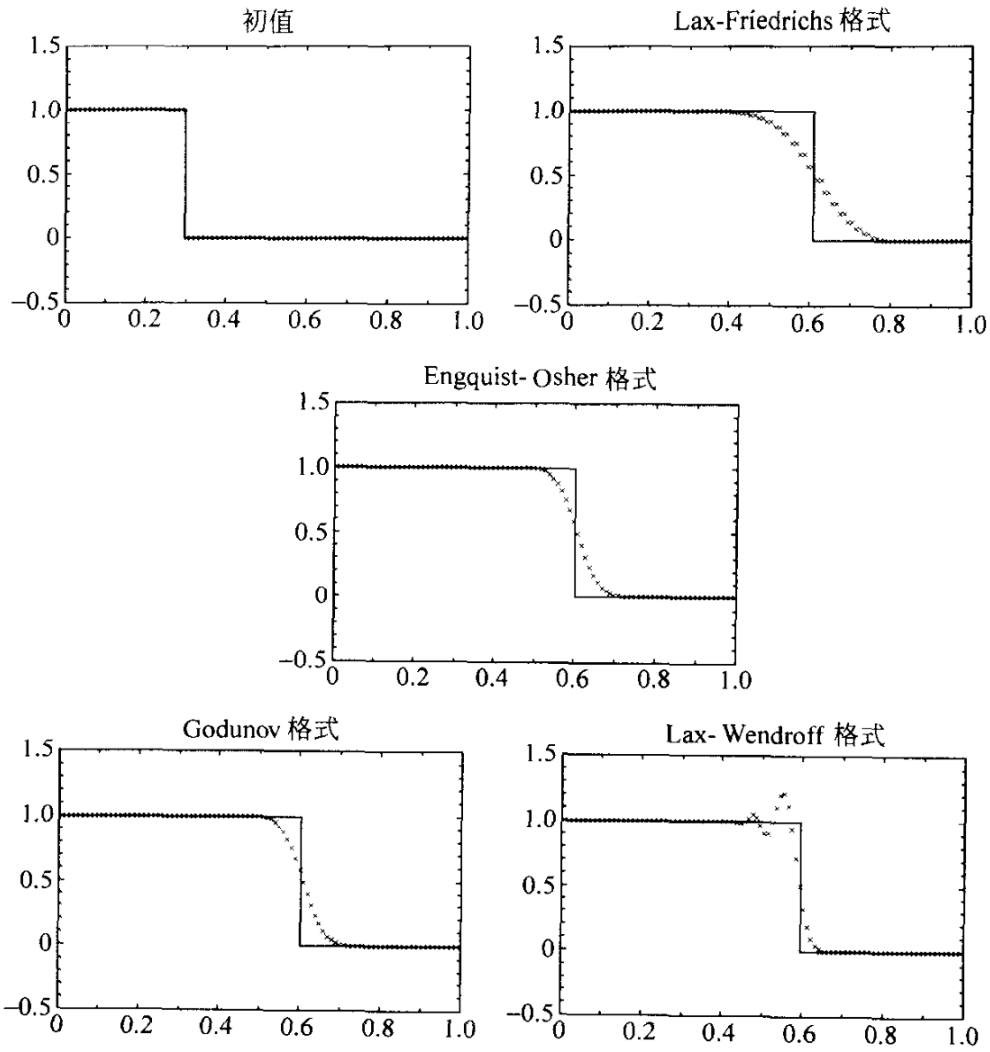


图 6.7

Figure 1.18: Numerical results for Burgers' equation using various schemes.

by controlling the **total variation** of the solution: on a finite domain $[0, X]$ divided into J cells, with U^n taking the value U_j^n in cell j at time level n, we can define the total variation as

$$\mathrm{TV}(U^n) := \sum_{j=1}^{J-1} |U_{j+1}^n - U_j^n| = \sum_{j=1}^{J-1} |\Delta_{+x} U_j^n|. \tag{1.165}$$

More generally, for the exact solution $u(x,t)$, $\mathrm{TV}(u(\cdot,t))$ 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 $|u(\xi_{j+1},t) - u(\xi_j,t)|$. 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 aside 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.

A key property of the solution of a conservation law such as (1.69) is that $\mathrm{TV}(u(\cdot,t))$ is a nonincreasing function of t – which can be deduced informally from the constancy of the solution along the characteristics described by (1.5). Thus we define **TVD (total variation diminishing)** schemes as those for which we have $\mathrm{TV}(U^{n+1}) \le \mathrm{TV}(U^n)$. This concept is due to Harten [4] who established the following useful result:

Theorem 1.9.1 ***(Harten)*** *A scheme is TVD if it can be written in the form*

$$U_j^{n+1} = U_j^n - C_{j-1}\Delta_{+x}U_{j-1}^n + D_j\Delta_{+x}U_j^n, \tag{1.166}$$

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

$$C_j \ge 0, \quad D_j \ge 0 \quad \textit{and} \quad C_j + D_j \le 1 \quad \forall j. \tag{1.167}$$

Proof. Taking the forward difference of (1.166), and freely using the identity $\Delta_{+x}U_{j-1} \equiv \Delta_{-x}U_j$, we get

$$U_{j+1}^{n+1} - U_j^{n+1} = \Delta_{+x}U_j^n - C_j\Delta_{+x}U_j^n + C_{j-1}\Delta_{+x}U_{j-1}^n + D_{j+1}\Delta_{+x}U_{j+1}^n - D_j\Delta_{+x}U_j^n \tag{1.168}$$

$$= (1 - C_j - D_j)\Delta_{+x}U_j^n + C_{j-1}\Delta_{+x}U_{j-1}^n + D_{j+1}\Delta_{+x}U_{j+1}^n. \tag{1.169}$$

By the hypotheses of (1.167), all the coefficients on the right of this last expression are nonnegative. So we can take absolute values to obtain

$$\left|\Delta_{+x}U_j^{n+1}\right| \le (1 - C_j - D_j)\left|\Delta_{+x}U_j^n\right| + C_{j-1}\left|\Delta_{+x}U_{j-1}^n\right| + D_{j+1}\left|\Delta_{+x}U_{j+1}^n\right|, \tag{1.170}$$

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

2. Roe upwind scheme

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 (1.100) with $A_{j\pm1/2}^n := \Delta_{\pm x}F_j^n / \Delta_{\pm x}U_j^n$. This corresponds to the scalar case of the scheme due to Roe, and is best considered as a finite volume scheme in which the fluxes of (1.114) are given by

$$F_{j+1/2}^{n+1/2} = \begin{cases} f(U_j^n) & \text{when } A_{j+1/2}^n \ge 0, \\ f(U_{j+1}^n) & \text{when } A_{j+1/2}^n < 0; \end{cases} \tag{1.171}$$

[4] Harten, A. (1983), High resolution schemes for hyperbolic conservation laws, *J. Comput. Phys.* **49**, 357–93.

or, equivalently,

$$F^{n+1/2}_{j+1/2} = \frac{1}{2}\left[\left(1+\operatorname{sgn} A^n_{j+1/2}\right)F^n_j + \left(1-\operatorname{sgn} A^n_{j+1/2}\right)F^n_{j+1}\right]. \tag{1.172}$$

Then, comparing (1.100) with (1.166) after replacing the flux difference $\Delta_{-x}F^n_j$ by $A^n_{j-1/2}\Delta_{-x}U^n_j$, we are led to setting

$$C_{j-1} = \frac{1}{2}\frac{\Delta t}{\Delta x}\left(1+\operatorname{sgn} A^n_{j-1/2}\right)A^n_{j-1/2}. \tag{1.173}$$

This is clearly always nonnegative, thus satisfying the first condition of (1.167). Similarly, we set

$$D_j = -\frac{1}{2}\frac{\Delta t}{\Delta x}\left(1-\operatorname{sgn} A^n_{j+1/2}\right)A^n_{j+1/2}, \tag{1.174}$$

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

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

which is just the CFL number. Hence the last condition of (1.167) corresponds to the CFL stability condition; we have shown that the Roe first order upwind scheme is TVD when Δt is chosen so that it is stable.

3. Lax-Wendroff scheme

On the other hand, if we attempt to follow similar arguments with the Lax-Wendroff scheme in the corresponding form of (1.74) and write $\nu^n_{j+1/2}$ for $A^n_{j+1/2}\Delta t/\Delta x$, we are led to setting

$$C_j = \frac{1}{2}\nu^n_{j+1/2}(1+\nu^n_{j+1/2}), \quad \text{and} \quad D_j = -\frac{1}{2}\nu^n_{j+1/2}(1-\nu^n_{j+1/2}), \tag{1.175}$$

both of which have to be nonnegative. Then the third condition of (1.167) requires that the CFL condition $(\nu^n_{j+1/2})^2 \le 1$ be satisfied, and the only values that $\nu^n_{j+1/2}$ can take to satisfy all three conditions are $-1, 0$ and $+1$; this is clearly impractical for anything other than very special cases.

4. Engquist and Osher

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, and it is particularly successful in modelling shocks. However, it needs modification for the handling of some rarefaction waves. For instance, suppose that the inviscid Burgers equation (1.76) is given the initial data $\{U^0_j = -1, \text{ for } j \le 0; U^0_j = +1, \text{ for } j > 0\}$, which should lead to a spreading rarefaction wave. Then it is clear from (1.171) that in Roe's scheme all the fluxes would be equal to $\frac{1}{2}$ so that the solution would not develop at all. The problem is associated with the **sonic points** which occurs for $u = 0$ where the characteristic speed a is zero – more precisely, with the **transonic rarefaction wave** that we have here in which the characteristic speeds are negative to the left of this point and positive to the right. For a general convex flux function $f(u)$, suppose that it has a (unique) sonic point at $u = u_s$. Then an alternative finite volume scheme has a form that replaces the flux function of (1.172) by

$$F^{n+1/2}_{j+1/2} = \frac{1}{2}\left[(1+\operatorname{sgn} A^n_j)F^n_j + (\operatorname{sgn} A^n_{j+1} - \operatorname{sgn} A^n_j)f(u_s) + (1-\operatorname{sgn} A^n_{j+1})F^n_{j+1}\right]. \tag{1.176}$$

This scheme, which uses the signs of the characteristic speeds $\{A_j^n = a(U_j^n)\}$ rather than those of the divided differences $\{A_{j+1/2}^n\}$, is due to Engquist and Osher [5] and has been very widely used and studied; it differs from the Roe scheme only when a sonic point occurs between U_j^n and U_{j+1}^n, it is also TVD (see Exercise 11) and it correctly resolves the transonic rarefaction wave.

Remark 1.9.2 *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 (1.166) and satisfying the conditions (1.167). 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 (1.49), that second order accuracy leads directly to these coefficients, as in (1.171), and hence the violation of the TVD conditions except in very special cases. From another viewpoint, in our 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 [6] 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,[7] 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. In all such schemes only the interface fluxes are changed in the finite volume update procedure of (1.114). How this is done using the recovered approximation $\tilde{U}^n(\cdot)$ is beyond the scope of this book; we merely note that it is based on solving local evolution problems in the manner initiated by the seminal work of Godunov.[8] However, we must observe that to obtain a TVD approximation in this way it is necessary to place restrictions on the recovery process. A typical constraint is that it is monotonicity preserving; that is, if the $\{U_j^n\}$ are monotone increasing then so must be $\tilde{U}^n(\cdot)$.

[5] Engquist, B. and Osher, O. (1981), One-sided difference approximations for nonlinear conservation laws, *Math. Comp.* **36**, 321–52.

[6] van Leer, B. (1979), Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method, *J. Comput. Phys.* **32**, 101–36.

[7] Colella, P. and Woodward, P.R. (1984), The piecewise parabolic method (PPM) for gas-dynamical simulations, *J. Comput. Phys.* **54**, 174–201.

[8] Godunov, S.K. (1959), A difference scheme for numerical computation of discontinuous solutions of equations of fluid dynamics, *Mat. Sb.* **47**, 271–306.

1.10 The leap-frog scheme

1.10.1 advection equation

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 Fig. 1.19. For (1.69) it has the form

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} + \frac{f(U_{j+1}^n) - f(U_{j-1}^n)}{2\Delta x} = 0, \tag{1.177}$$

or

$$U_j^{n+1} = U_j^{n-1} - (a\Delta t/\Delta x)\left[U_{j+1}^n - U_{j-1}^n\right]. \tag{1.178}$$

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.

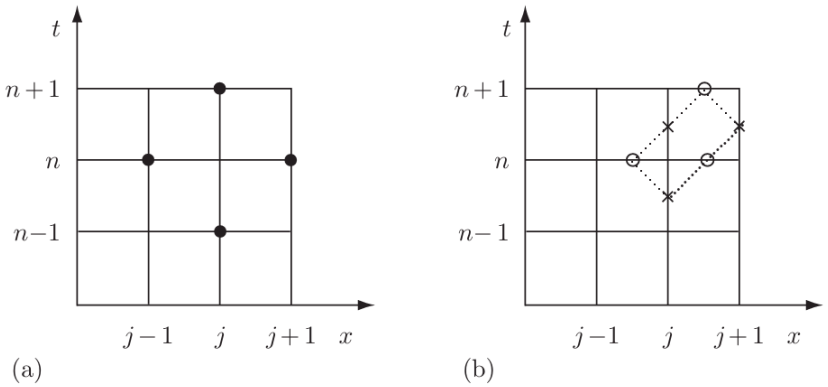


Fig. 4.14. The leap-frog scheme: (a) unstaggered; (b) staggered, × = V and o = W.

Figure 1.19: The leap-frog scheme (a) unstaggered; (b) staggered.

It is clear from Fig. 1.19(a) that the CFL condition requires that $|\nu| \leq 1$, as for the Lax-Wendroff scheme. When $f = au$ with constant a the usual Fourier analysis leads to a quadratic for $\lambda(k)$:

$$\lambda^2 - 1 + 2i\nu\lambda \sin k\Delta x = 0 \tag{1.179}$$

with solutions

$$\lambda(k) = -i\nu \sin k\Delta x \pm \left[1 - \nu^2 \sin^2 k\Delta x\right]^{1/2}. \tag{1.180}$$

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 $|\nu| \leq 1$: 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 (1.180) we obtain a mode that provides a good approximation to the differential equation, namely the 'true' mode λ_T given by

$$\begin{aligned} \arg \lambda_T &= -\sin^{-1}(\nu \sin k\Delta x) \\ &\sim -\nu\xi \left[1 - \frac{1}{6}(1-\nu^2)\xi^2 + \cdots\right]. \end{aligned} \tag{1.181}$$

Note that the phase error here has the same leading term as the Lax-Wendroff scheme – see (1.54). On the other hand, taking the negative root gives the spurious mode

$$\lambda_S \sim (-1)\left[1 + i\nu\xi - \frac{1}{2}\nu^2\xi^2 + \cdots\right], \tag{1.182}$$

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 Fig. 1.20 illustrate the application of the leap-frog 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**.

1.10.2 wave equation

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

$$u_{tt} = a^2 u_{xx}, \tag{1.183}$$

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

$$v_t + aw_x = 0, \tag{1.184}$$

$$w_t + av_x = 0. \tag{1.185}$$

Because of the pattern of differentials here, a staggered form of the leap-frog method can be used that is much more compact than (1.178): as indicated in Fig. 1.19(b) we have V and W at different points and a staggered scheme can be written

$$\frac{V_j^{n+1/2} - V_j^{n-1/2}}{\Delta t} + a\frac{W_{j+1/2}^n - W_{j-1/2}^n}{\Delta x} = 0, \tag{1.186}$$

$$\frac{W_{j+1/2}^{n+1} - W_{j+1/2}^n}{\Delta t} + a\frac{V_{j+1}^{n+1/2} - V_j^{n+1/2}}{\Delta x} = 0, \tag{1.187}$$

or

$$\delta_t V + \nu\delta_x W = 0, \quad \delta_t W + \nu\delta_x V = 0, \tag{1.188}$$

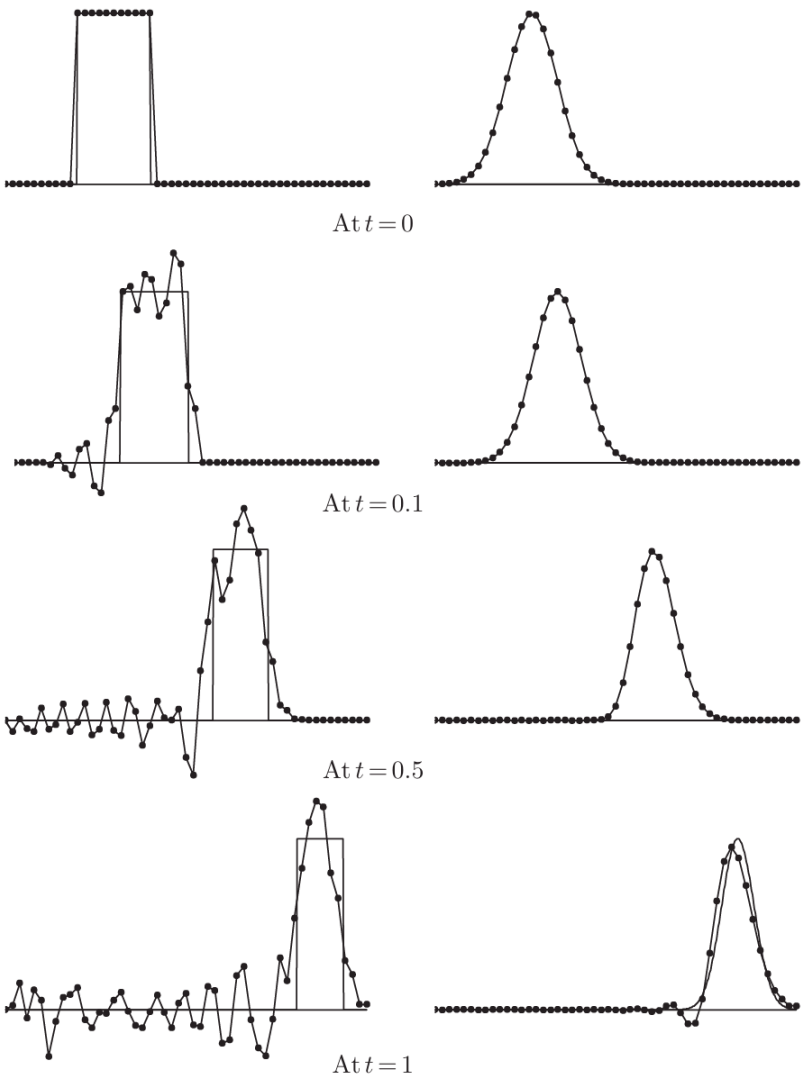


Fig. 4.15. Linear advection by the leap-frog scheme with $\Delta t = \Delta x = 0.02$ for (a) a square pulse and (b) Gaussian initial data.

Figure 1.20: Linear advection by the leap-frog scheme with $\Delta t = \Delta x = 0.02$ for (a) a square pulse and (b) Gaussian initial data.

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

$$(V^{n-1/2}, W^n) = \lambda^n e^{ikx}(\hat{V}, \hat{W}) \tag{1.189}$$

where $\hat{V}$ and $\hat{W}$ are constants. These will satisfy the equations (1.186,1.187) if

$$\begin{pmatrix} \lambda - 1 & 2i\nu \sin \frac{1}{2}k\Delta x \\ 2i\lambda\nu \sin \frac{1}{2}k\Delta x & \lambda - 1 \end{pmatrix} \begin{pmatrix} \hat{V} \\ \hat{W} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1.190}$$

This requires the matrix in (1.190) to be singular, so that

$$\lambda^2 - 2(1 - 2\nu^2 \sin^2 \frac{1}{2}k\Delta x)\lambda + 1 = 0 \tag{1.191}$$

with solutions given by

$$\lambda_\pm(k) = 1 - 2\nu^2 s^2 \pm 2i\nu s[1 - \nu^2 s^2]^{1/2}, \tag{1.192}$$

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 $|\nu| \leq 1$, in which case $|\lambda_\pm| = 1$. The phases are given by

$$\arg \lambda_\pm = \pm \sin^{-1}\left(2\nu s[1 - \nu^2 s^2]^{1/2}\right) \sim \pm\nu\xi \left[1 - \frac{1}{24}(1 - \nu^2)\xi^2 + \cdots\right]. \tag{1.193}$$

Note that the two roots of (1.191) are just the squares of the roots of (1.179) with Δx replaced by $\frac{1}{2}\Delta x$; hence the expansion in (1.193) corresponds to that in (1.181) 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

$$V_j^{n+1/2} = (U_j^{n+1} - U_j^n)/\Delta t, \quad W_{j+1/2}^n = -a(U_{j+1}^n - U_j^n)/\Delta x \tag{1.194}$$

into the equations (1.186,1.187), (1.188) gives

$$(\delta_t^2 - \nu^2\delta_x^2)U_j^n = 0, \tag{1.195}$$

the simplest central difference representation of the second order wave equation (1.183) 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.

1.11 Hamiltonian ODE systems and symplectic schemes

With the growing power of computers, it becomes more feasible to numerically simulate many-body systems with a large number of degrees of freedom. However, usual numerical algorithms such as Euler scheme or **Runge-Kutta methods** do not conserve the total energy of a Hamiltonian system for long time simulations. Hamiltonian systems possess symplectic structures, which we define below. Therefore, efficient and precise algorithms that capture this crucial characteristic of Hamiltonian system, are needed. Here, we

describe a class of such algorithms called **symplectic integrators** and provide all the necessary formulas and parameters for one particular method, i.e., the sixth order Yoshida method [9].

Suppose, we study a system of N particles. At any given time the position of the system is described by the set of $d \cdot N$ coordinates and $d \cdot N$ momenta, which we will denote as a pair (q, p) assuming that both components are d-dimensional vectors. In Hamiltonian mechanics, the evolution of the system is given by the Hamiltonian function $H(p, q)$ via the canonical equations of motion

$$\begin{cases} \dot{q}_j = \frac{\partial H}{\partial p_j}, \\ \dot{p}_j = -\frac{\partial H}{\partial q_j}. \end{cases} \tag{1.196}$$

These equations define a time flow of the phase space – to find the position (q_τ, p_τ) of the system at time τ one can integrate Eq. (1.196) up to $t = \tau$ with the initial conditions (q_0, p_0) given at $t = 0$. By definition, the time flow of the phase space is symplectic if it preserves the differential form

$$\omega = dp \wedge dq. \tag{1.197}$$

In order to construct the symplectic integrator, we provide a formal description of the Hamiltonian flow given by Eqs. (1.196). Define $z = (p, q)$ and the Poisson bracket $\{\cdot, \cdot\}$ as

$$\{f, g\} = \frac{\partial f}{\partial q}\frac{\partial g}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial g}{\partial q}. \tag{1.198}$$

Then Eqs. (1.196) can be written in the form

$$\dot{z} = \{z, H(z)\}. \tag{1.199}$$

By introducing the notation $D_H = \{\cdot, H(\cdot)\}$ for the differential operator, Eq. (1.199) can be rewritten as

$$\dot{z} = D_H z, \tag{1.200}$$

and its formal solution at $t = \tau$ is given by

$$z(\tau) = e^{\tau D_H} z(0). \tag{1.201}$$

Since the total energy is conserved, we can write

$$H(z(\tau)) = H(z(0)). \tag{1.202}$$

Suppose, the total energy of the system is a sum of the kinetic energy that is a function of p only and a potential energy that is a function of q only

$$H(q, p) = T(p) + V(q). \tag{1.203}$$

Then the corresponding differential operators are denoted as

$$\begin{cases} D_T = \frac{\partial T}{\partial p}\frac{\partial}{\partial q}, \\ D_V = -\frac{\partial V}{\partial q}\frac{\partial}{\partial p}, \end{cases} \tag{1.204}$$

[9] H. Yoshida, Phys. Lett. A 150, 262 (1990).

and thus the formal solution (1.201) becomes

$$z(\tau) = e^{\tau(D_T+D_V)}z(0). \tag{1.205}$$

Since the differential operators D_T and D_V are non-commutative, exponential of the sum of two operators in Eq. (1.205) is not equal to the product of exponentials of the individual components. Instead, for any non-commutative differential operators A and B, the approximate relationship holds

$$e^{\tau(A+B)} = e^{\tau A}e^{\tau B} + o(\tau^2). \tag{1.206}$$

This is closely related to the operator splitting method. Our goal is to build the n-th order symplectic scheme, and in order to achieve this we generalize Eq. (1.206) and construct the expansion of the form

$$e^{\tau(A+B)} = e^{c_1\tau A}e^{d_1\tau B}e^{c_2\tau A}e^{d_2\tau B} \times \cdots \times e^{c_k\tau A}e^{d_k\tau B} + o(\tau^{n+1}), \tag{1.207}$$

where $c_1 \ldots c_k$ and $d_1 \ldots d_k$ are real numbers. Now, the solution for $z(\tau)$ is approximated by

$$z'(\tau) = \left(\prod_{j=1}^{k} e^{c_j\tau A}e^{d_j\tau B}\right) z(0), \tag{1.208}$$

Note that Eq. (1.208) provides a symplectic mapping in the phase space, i.e., because it consists of a series of elementary symplectic mappings $e^{c_j\tau A}$ and $e^{d_j\tau B}$. Moreover, using Taylor expansions of the operators (1.204) up to the first order in τ, the corresponding elementary mappings are explicitly computable

$$q^{(j)} = q^{(j-1)} + \tau c_j \frac{\partial T}{\partial p}(p^{(j-1)}), \tag{1.209}$$

$$p^{(j)} = p^{(j-1)} - \tau d_j \frac{\partial V}{\partial q}(q^{(j)}), \tag{1.210}$$

for $j = 1$ to $j = k$. Naturally, the question arises whether the expansion of the form (1.207) exists for any order n. It turns out that for every even order n there exists at least one set of exact coefficients $c_1 \ldots c_k$ and $d_1 \ldots d_k$ so that Eq. (1.208) provides an order n approximate solution for the dynamical equation (1.199). The general approach of finding the coefficients c_j and d_j is as follows. We expand the LHS of Eq. (1.207) in powers of τ up to the order n. Then we equate the coefficients of the corresponding powers on both sides of Eq. (1.207) and obtain a system of equations for c_j and d_j. The resulting coefficients for $n = 6$ are given by the above reference

$$\begin{aligned}
c_1 &= 0.392256805238780 \quad d_1 = 0.784513610477560,\\
c_2 &= 0.510043411918458 \quad d_2 = 0.235573213359357,\\
c_3 &= -0.471053385409757 \quad d_3 = -1.177679984178870,\\
c_4 &= 0.068753168252518 \quad d_4 = 1.315186320683906,\\
c_5 &= 0.068753168252518 \quad d_5 = -1.177679984178870,\\
c_6 &= -0.471053385409757 \quad d_6 = 0.235573213359357,\\
c_7 &= 0.510043411918458 \quad d_7 = 0.784513610477560,\\
c_8 &= 0.392256805238780 \quad d_8 = 0.0,
\end{aligned}$$

To summarize, here we presented the numerical algorithm for solving the Hamiltonian equations of motion (1.196) using the symplectic algorithm given by Eqs. (1.209) and (1.210). Note that strictly speaking a

symplectic algorithm may not preserve energy as well as an explicit Runge-Kutta method for a fixed time step τ. However, it is superior to an explicit Runge-Kutta method since the system energy is bounded when computed by a symplectic algorithm in contrast to the unbounded energy when computed by an explicit Runge-Kutta method. This property of a symplectic algorithm becomes important when a long time (statistical) behavior of a Hamiltonian system is studied.