CHAPTER 6

Finite difference methods for hyperbolic equations

1. Scalar equations

1.1. Constant velocity advection in one dimension. The simplest example of a hyperbolic equation is the constant velocity advection equation

$$q_t + u\, q_x = 0 \tag{1.1}$$

with some initial condition $q(x, t = 0) = q_0(x)$. The equation can be solved along the entire real axis in x or some portion thereof. In numerical work we always have a finite subdomain which we shall conveniently choose as $[0, 2\pi]$ with a view to applying Fourier analysis later on. When using a finite subdomain the question of boundary conditions arises which we shall postpone by considering periodic boundary conditions $q(x + 2\pi, t) = q(x, t)$.

1.1.1. *Exact solution by characteristics.* A first attack on finding the solution to (1.1) is to try to reduce it to a simpler problem. One can ask whether there is any subdomain over which the equation can be cast in a simpler form. For instance we can inquire whether there are any particular curves within the (x, t) plane over which the equation simplifies. A general curve Γ of curvilinear parameter is given by

$$\Gamma : \; x = x(s), \; t = t(s) \tag{1.2}$$

and the infinitesimal change in q when going along Γ is

$$\frac{dq}{ds} = \frac{\partial q}{\partial t}\frac{dt}{ds} + \frac{\partial q}{\partial x}\frac{dx}{ds} \tag{1.3}$$

Comparing (1.3) with (1.1) we see that if we impose

$$\frac{dt}{ds} = 1, \; \frac{dx}{ds} = u \tag{1.4}$$

then by (1.1) we must have that

$$\frac{dq}{ds} = 0 \; . \tag{1.5}$$

This means that q is constant along the curves Γ defined by (1.4) which are $x = ut + C$. The curves are shown in Fig. (1)

1.1.2. *Finite difference methods.* We can construct numerical methods for (1.1) by the same approaches used for the heat equation.

FIGURE 1. Characteristic curves for $q_t + q_x = 0$.

Semi-discretization. Define a computational grid $x_j = jh$, $h = 2\pi/(M+1)$, $t^n = nk$ with step size h, k in space and time. Define $Q_j(t)$ to be the restriction of $q(x,t)$ to $x = x_j$

$$Q_j(t) = q(x_j, t),\ j = 0, 1, \ldots, M+1\ . \tag{1.6}$$

We can choose some approximation of the x derivative. For instance approximating

$$\frac{dq(x_j,t)}{dx} \cong \frac{\delta}{h} Q_j = \frac{Q_{j+1}(t) - Q_{j-1}(t)}{2h} \tag{1.7}$$

leads to the ODE system

$$\frac{d}{dt}\mathbf{Q} = -\frac{u}{2h} B\mathbf{Q} \tag{1.8}$$

with

$$\mathbf{Q} = \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_M \end{bmatrix}^T \tag{1.9}$$

$$B = \begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{bmatrix} \tag{1.10}$$

We can now try various ODE schemes to solve (1.8). Using Euler's method would lead to a FTCS scheme

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n - \frac{uk}{2h} B\mathbf{Q}^n = \left(I - \frac{uk}{2h} B\right)\mathbf{Q}^n\ . \tag{1.11}$$

If instead of Euler's scheme we use the midpoint method we obtain the update formula

$$\mathbf{Q}^{n+1} = \mathbf{Q}^{n-1} - \frac{uk}{h} B\mathbf{Q}^n \tag{1.12}$$

known as the *leap-frog* or *Dufort-Frankel* method.

Full discretization. Instead of the semi-discretization or method of lines approach we can also directly discretize both the space and time derivatives appearing in (1.1). A first-order forward in time, second-order centered in space discretization would lead to the FTCS scheme

$$Q_j^{n+1} = Q_j^n - \frac{uk}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right) \tag{1.13}$$

A modification of (1.13) of historical relevance is the Lax-Friedrichs scheme

$$Q_j^{n+1} = \frac{1}{2}\left(Q_{j+1}^n + Q_{j-1}^n\right) - \frac{uk}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right)\ . \tag{1.14}$$

This scheme is obtained by replacing Q_j^n with its arithmetic average using values to the left and to the right. Since the formula has not been derived from discretizations of the derivative which we know to be consistent with the original equation it is useful to determine the truncation error. The exact advection operator is

$$D = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} \tag{1.15}$$

and we have $Dq = 0$ according to (1.1). Our approximation of this operator is

$$\tilde{D}q(x_j, t^n) = \frac{q_j^{n+1} - \frac{1}{2}\left(q_{j+1}^n + q_{j-1}^n\right)}{k} + \frac{u}{2h}\left(q_{j+1}^n - q_{j-1}^n\right) \tag{1.16}$$

with $q_j^n = q(x_j, t^n)$. The truncation error is therefore

$$\tau_j^n = \left(\tilde{D} - D\right) q(x_j, t^n) = \frac{1}{k} q_j^{n+1} - \frac{1}{2k}\left(q_{j+1}^n + q_{j-1}^n\right) + \frac{u}{2h}\left(q_{j+1}^n - q_{j-1}^n\right) \tag{1.17}$$

We can now carry out a Taylor's series expansion around (x_j, t^n). To simplify notation q and its derivatives will be understood to be evaluated at (x_j, t^n) if not explicitly shown otherwise

(1.18)

$$\tau_j^n = \frac{1}{k}\left(q + kq_t + \frac{k^2}{2}q_{tt} + \ldots\right) - \frac{1}{2k}\left(q + hq_x + \frac{h^2}{2}q_{xx} + \ldots + q - hq_x + \frac{h^2}{2}q_{xx} + \ldots\right)$$

(1.19)

$$+ \frac{u}{2h}\left(q + hq_x + \frac{h^2}{2}q_{xx} + \frac{h^3}{6}q_{xxx} + \ldots - q + hq_x - \frac{h^2}{2}q_{xx} + \frac{h^3}{6}q_{xxx} + \ldots\right)$$

which gives

$$\tau_j^n = q_t + \frac{k}{2}q_{tt} + \ldots - \frac{h^2}{2k}q_{xx} + \ldots + uq_x + \frac{uh^2}{12}q_{xxx} + \ldots \tag{1.20}$$

Using (1.1) leads to the leading order error

$$\tau_j^n \cong \frac{k}{2}q_{tt} + \frac{uh^3}{12}q_{xxx} = O(k, h^2) \tag{1.21}$$

The analysis shows that the scheme is first order in time and second order in space. The Lax-Friedrichs scheme is consistent, i.e.

$$\lim_{k,h\to 0} \tau_j^n = 0\ . \tag{1.22}$$

Instead of centered finite differences, other approximations may be introduced. A good choice would be to use one-sided finite differences. This would take into account what we know from the exact solution to the advection equation: information travels along the characteristic lines. It would make sense to use finite differences which have a stencil that mimics this behavior. If $u > 0$ we would use left-sided differences and for $u < 0$ we would use right-sided differences. A first order approximation would be

$$Q_j^{n+1} = \begin{cases} Q_j^n - \dfrac{uk}{h}\left(Q_j^n - Q_{j-1}^n\right) & u \geq 0 \\ Q_j^n - \dfrac{uk}{h}\left(Q_{j+1}^n - Q_j^n\right) & u \leq 0 \end{cases} \tag{1.23}$$

This is known, naturally enough, as the *upwind scheme*.

Taylor series approach. A procedure useful in deriving higher order finite difference approximations for (1.1) is the Taylor series approach. In practical work it is economical to only store two time levels at any given stage in the computation. The general prescription for attaining higher order for the time derivative would involve keeping more terms from the operator series

$$\frac{\partial}{\partial t} = \frac{1}{k}\left(\Delta_+ - \frac{\Delta_+^2}{2} + \frac{\Delta_+^3}{3} - \ldots\right)\ . \tag{1.24}$$

This would be incovenient since the time stencil of the scheme would become wider and we would need to store more than two time levels. We can however use (1.1) to convert time derivatives into spatial derivatives

$$q_t = -uq_x \tag{1.25}$$

$$q_{tt} = \frac{\partial}{\partial t}(q_t) = -\frac{\partial}{\partial t}(uq_x) = -u\frac{\partial}{\partial x}(q_t) = u\frac{\partial}{\partial x}(uq_x) = u^2 q_{xx} \tag{1.26}$$

The Taylor series approach can now be applied to obtain as high an order of approximation in time as needed

$$q(t+k) = q + kq_t + \frac{k^2}{2}q_{tt} + \frac{k^3}{6}q_{ttt} + \dots \ . \tag{1.27}$$

As an example, let us construct a second order scheme by truncating

$$q(t+k) \cong q + kq_t + \frac{k^2}{2}q_{tt} \tag{1.28}$$

We now replace the time derivatives with spatial derivative

$$q(t+k) \cong q - ukq_x + \frac{u^2k^2}{2}q_{xx} \tag{1.29}$$

and use second order accurate, centered finite differences to approximate the spatial derivatives. The resulting scheme is

$$Q_j^{n+1} = Q_j^n - \frac{uk}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{u^2k^2}{2h^2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \tag{1.30}$$

This is known as the *Lax-Wendroff scheme*.

Instead of centered finite differences we might want to use one-sided formulas to take into account the direction of the characteristics of (1.1). Let us construct a second order accurate one sided approximation using (**??**). One-sided differences can be obtained to arbitrary order of accuracy using the series

$$\frac{\partial}{\partial x} = \frac{1}{h}\left(\Delta_{x+} - \frac{\Delta_{x+}^2}{2} + \frac{\Delta_{x+}^3}{3} - \dots\right) \tag{1.31}$$

$$= \frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2} + \frac{\Delta_{x-}^3}{3} - \dots\right) \tag{1.32}$$

with the finite difference operators defined by

$$\Delta_{x+}q(x,t) = q(x+h,t) - q(x,t) \tag{1.33}$$

$$\Delta_{x-}q(x,t) = q(x,t) - q(x-h,t) \tag{1.34}$$

Let us assume that $u \geq 0$ and therefore that we will be using backward differences so that the computational stencil mimics the true domain of dependence. A second

order accurate approximation of q_x is given by

(1.35)
$$q_x(x_j,t^n) = \frac{\partial q}{\partial x}(x_j,t^n) \cong \frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2}\right) q(x_j,t^n)$$

$$= \frac{1}{h}\left[q(x_j,t^n) - q(x_{j-1},t^n) + \frac{1}{2}\left(q(x_j,t^n) - 2q(x_{j-1},t^n) + q(x_{j-2},t^n)\right)\right] \tag{1.36}$$

$$\cong \frac{1}{h}\left[Q_j^n - Q_{j-1}^n + \frac{1}{2}\left(Q_j^n - 2Q_{j-1}^n + Q_{j-2}^n\right)\right] \tag{1.37}$$

$$= \frac{1}{2h}\left(3Q_j^n - 4Q_{j-1}^n + Q_{j-2}^n\right) \tag{1.38}$$

The second derivative is obtained from

(1.39)
$$\frac{\partial^2}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial}{\partial x} = \frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2} + \frac{\Delta_{x-}^3}{3} - \dots\right)\frac{1}{h}\left(\Delta_{x-} + \frac{\Delta_{x-}^2}{2} + \frac{\Delta_{x-}^3}{3} - \dots\right)$$

(1.40)
$$= \frac{1}{h^2}\left(\Delta_{x-}^2 + \Delta_{x-}^3 + \frac{11}{12}\Delta_{x-}^3\right)$$

Note that in (**??**) we have neglected terms of $O(k^3)$. The exact solution of the advection equation is $q(x,t) = q_0(x-ut)$. The $x-ut$ argument sugests that similar step sizes should be used for x and t. This will be confirmed by our stability analysis below. So let us assume that $k = O(h)$. Since q_{xx} already has a k^2 factor in (**??**) we only need an $O(h)$ approximation of q_{xx}. The leading order error term from $k^2 q_{xx}$ will then be of $O(k^2 h) = O(k^3) = O(h^3)$. We can therefore truncate the series (1.40) to just the first term and approximate

$$\frac{\partial^2 q}{\partial x^2}(x_j,t^n) \cong \frac{1}{h^2}\Delta_{x-}^2 q(x_j,t^n) = \frac{1}{h^2}\left[q(x_j,t^n) - 2q(x_{j-1},t^n) + q(x_{j-2},t^n)\right] \tag{1.41}$$

$$\cong \frac{1}{h^2}\left(Q_j^n - 2Q_{j-1}^n + Q_{j-2}^n\right) \tag{1.42}$$

Note that this is different from the procedure used in deriving the Lax-Wendroff scheme where a second order accurate expression of q_{xx} was used. The reason is that in the Lax-Wendroff scheme the computational stencil already included Q_{j-1}^n, Q_j^n, Q_{j+1}^n from the approximation of the first derivative q_x. Since the second order accurate approximation of q_{xx} does not widen the stencil there is no penalty in using the more accurate, second order approximation of q_{xx}. In the one-sided scheme we are deriving here however, using a second order accurate approximation of q_{xx} would involve widening the computational stencil to include Q_{j-3}^n. This increases the arithmetic cost of applying the formula without noticeable gain so we choose to use an $O(h)$ approximation of q_{xx}. Combining the above results we obtain

$$Q_j^{n+1} = Q_j^n - \frac{uk}{2h}\left(3Q_j^n - 4Q_{j-1}^n + Q_{j-2}^n\right) + \frac{1}{2}\left(\frac{uk}{h}\right)^2\left(Q_j^n - 2Q_{j-1}^n + Q_{j-2}^n\right) \tag{1.43}$$

for $u > 0$, which is known as the *Beam-Warming scheme*.

1.1.3. *Stability analysis.* We now turn to the analysis of the stability of the various schemes introduced above. The analysis can be done using the techniques for systems of ODE's or using Von Neumann analysis. We shall carry out both procedures.

Semi-discretized system. The matrix B arising in the semi-discretized approach is skew-symmetric and will have purely imaginary eigenvalues. We can check this by explicitly calculating the eigenvalues. As usual, we guess that

$$\mathbf{W}_p = \begin{bmatrix} e^{iph} & e^{ip2h} & \cdots & e^{ipjh} & \cdots & e^{ipMh} \end{bmatrix} \tag{1.44}$$

will be an eigenvector since B discretizes a derivation operator. Computing the j^{th} component of $B\mathbf{W}_p$ we get

$$\left(B\mathbf{W}_p\right)_j = e^{ip(j+1)h} - e^{ip(j-1)h} = 2i\sin ph\ e^{ipjh} = 2i\sin ph\ \left(\mathbf{W}_p\right)_j \tag{1.45}$$

so the eigenvalue associated with $\mathbf{W}_p$ is

$$\lambda_p = 2i\sin ph \tag{1.46}$$

and is indeed purely imaginary.

To establish the stability region for the FTCS method (**??**) we use the eigenvalues λ of B in the criterion

$$|1+z| \le 1 \tag{1.47}$$

with $z = k\lambda$. It is immediately apparent that the scheme will be unconditionally unstable because λ is purely imaginary $\lambda = ai$ so

$$|1+z| = \sqrt{1+(ka)^2} > 1 \tag{1.48}$$

for all $a > 0$.

The interval of stability for the midpoint scheme is $\mathrm{Re}\, z = 0$, $|\mathrm{Im}(z)| \le 1$. Here we would have

$$z = -\frac{uk}{h} i\sin ph \tag{1.49}$$

and the method is stable for

$$\left|\frac{uk}{h}\right| \le 1\ . \tag{1.50}$$

Von Neumann analysis. Obtaining analytical expression for the matrices arising from the semi-discretized approach becomes increasingly difficult as we use more accurate approximations of the derivatives in the PDE or study PDE's more complex than the advection equation. Von Neumann analysis is typically simpler to apply. We start by determing the stability region for the FTCS scheme (1.13). Substituting a typical wavemode $Q^n = \hat{Q}^n e^{i\xi jh}$ we obtain

$$\hat{Q}^{n+1}e^{i\xi jh} = \hat{Q}^n e^{i\xi jh} - \frac{uk}{2h}\left(\hat{Q}^n e^{i\xi(j+1)h} - \hat{Q}^n e^{i\xi(j-1)h}\right)\ . \tag{1.51}$$

The amplification ratio is

$$G = \frac{\hat{Q}^{n+1}}{\hat{Q}^n} = 1 - \frac{uk}{h} i\sin \xi h \tag{1.52}$$

which is always greater than 1

$$|G| \ge 1\ . \tag{1.53}$$

Thus the scheme is unconditionally unstable as we expected from the semi-discretized stability analysis done above.

For the Lax-Friedrichs scheme we obtain

(1.54)
$$\hat{Q}^{n+1}e^{i\xi jh} = \frac{1}{2}\left(\hat{Q}^n e^{i\xi(j+1)h} + \hat{Q}^n e^{i\xi(j-1)h}\right) - \frac{uk}{2h}\left(\hat{Q}^n e^{i\xi(j+1)h} - \hat{Q}^n e^{i\xi(j-1)h}\right)$$

and the amplification factor is

$$G = \cos\xi h - \frac{uk}{h} i \sin\xi h \tag{1.55}$$

Let us introduce the notation

$$\nu = \frac{uk}{h},\ \theta = \xi h\ . \tag{1.56}$$

The stability condition is that

$$|G| = \cos^2\theta + \nu^2\sin^2\theta \leq 1 = \cos^2\theta + \sin^2\theta \tag{1.57}$$

from where we obtain

$$\left(\nu^2 - 1\right)\sin^2\theta \leq 0\ . \tag{1.58}$$

The inequality is satisfied for

$$|\nu| \leq 1\ . \tag{1.59}$$

The quantity ν that appears repeatedly in analysis of numerical schemes for the advection equation is known as the *Courant-Friedrichs-Lewy* number or more concisely as the *CFL number*. We say that the Lax-Friedrichs scheme is stable for CFL numbers up to 1, it being implicitly understood that we're considering the absolute value of the velocity $|u|$. From the stability criterion we obtain a bound on the time step that we can use in the Lax-Friedrichs scheme

$$k \leq \frac{h}{|u|}\ . \tag{1.60}$$

For the Lax-Wendroff scheme the amplification ratio is

$$G = 1 - \nu i \sin\theta + \nu^2\left(\cos\theta - 1\right) \tag{1.61}$$

We have

$$|G| = 1 + 2\nu^2\left(\cos\theta - 1\right) + \nu^4\left(\cos\theta - 1\right)^2 + \nu^2\sin^2\theta \tag{1.62}$$

$$= 1 - 4\nu^2\sin^2\frac{\theta}{2}\left(1 - \cos^2\frac{\theta}{2}\right) + 4\nu^4\sin^4\frac{\theta}{2} \tag{1.63}$$

The stability condition is $|G| \leq 1$ leads to

$$\left(\nu^2 - 1\right)\sin^2\frac{\theta}{2} \leq 0 \tag{1.64}$$

so again the domain of stability is

$$|\nu| \leq 1\ . \tag{1.65}$$

Lax-Wendroff is a more efficient scheme than Lax-Friedrichs since we obtain $O(h^2, k^2)$ precision as opposed to $O(h, k^2)$ under the same time step restriction $k \leq h/|u|$.

Turning now to the one-sided schemes, for upwind when $u > 0$ we have

$$G = 1 - \nu\left(1 - e^{-i\theta}\right) \tag{1.66}$$

$$|G| = 1 - 2\nu\left(1 - \cos\theta\right) + \nu^2\left(1 - \cos\theta\right)^2 + \nu^2\sin^2\theta \tag{1.67}$$

FIGURE 2. Amplification factor $|G(\nu,\theta)|$ for the Beam-Warming scheme evaluated at $\theta = m\pi/8$, $m = 0, 1, \ldots 16$.

The stability condition $|G| \leq 1$ leads to

$$-2\nu\left(1-\cos\theta\right)+\nu^2\left(1-\cos\theta\right)^2+\nu^2\sin^2\theta\leq 0 \tag{1.68}$$

which can be rewritten in terms of the half-angle $\theta/2$ to give

$$-4\nu\sin^2\frac{\theta}{2}+4\nu^2\sin^4\frac{\theta}{2}+4\nu^2\sin^2\frac{\theta}{2}\cos^2\frac{\theta}{2}\leq 0 \tag{1.69}$$

and finally

$$\nu(\nu-1)\leq 0 \tag{1.70}$$

so the stability region is again $\nu \leq 1$.

For the Beam-Warming scheme we have

$$G=1-\frac{\nu}{2}\left(3-4e^{-i\theta}+e^{-2i\theta}\right)+\frac{1}{2}\nu^2\left(1-2e^{-i\theta}+e^{-2i\theta}\right)\ . \tag{1.71}$$

Notice that as we look at more complicated schemes the amplification factors become increasingly difficult to evaluate analytically. We can however use a numerical evaluation of $G(\nu,\theta)$ to generate plots such as Fig. 2. From the plot we deduce that the stability region is $\nu \leq 2$.

1.1.4. *Lax equivalence theorem.* The importance of establishing consistency and stability for a finite difference scheme for the advection equation is that these two properties guarantee convergence by the Lax equivalence theorem.

THEOREM 4. *A finite difference scheme for a linear PDE is convergent if the scheme is consistent with the PDE and it is stable.*

Convergence means that

$$\lim_{k,h\to 0} Q_j^n = q(x_j, t^n) \tag{1.72}$$

where k, h go to zero in accordance with the stability criterion for the scheme. Convergence is obtained when the scheme is consistent, i.e. the truncation error goes to zero

$$\lim_{k,h\to 0} \tau_j^n = 0 \tag{1.73}$$

and the step sizes satisfy the stability criterion.

1.1.5. *Modified equations.* We have established a number of methods for solving the advection equation (1.1). Up to now we have characterized the error of any one scheme by its truncation error. Though indicative of the overall quality of an approximation, the precise nature of the error in the scheme is not apparent. It has proved very fruitful in the development of better methods to more accurately describe how a numerical approximation differs from the exact solution. A question one can ask is whether a given numerical scheme is perhaps a more accurate discretization of another PDE than the one it was originally designed for. Let us exemplify using the upwind scheme for the advection equation with $u > 0$

$$Q_j^{n+1} = Q_j^n - \nu\left(Q_j^n - Q_{j-1}^n\right)\ . \tag{1.74}$$

We know that this scheme is $O(k,h)$ accurate for the equation $q_t+uq_x=0$. Suppose that the scheme is an exact discretization of some unknown PDE $Ls=0$ with L an unknown differential operator and $s=s(x,t)$. Then we would have

$$s(x,t+k)=s(x,t)-\nu\left[s(x,t)-s(x-h,t)\right]\ , \tag{1.75}$$

exactly. Let us carry out Taylor series expansion of s around (x,t)

$$s+ks_t+\frac{k^2}{2}s_{tt}+\frac{k^3}{6}s_{ttt}+\ldots=s-\frac{uk}{h}\left[hs_x-\frac{h^2}{2}s_{xx}+\frac{h^3}{6}s_{xxx}-\ldots\right]\ . \tag{1.76}$$

To obtain a more concise notation the function arguments have been dropped. We obtain

$$s_t+us_x=-\frac{k}{2}s_{tt}+\frac{uh}{2}s_{xx}-\frac{k^2}{6}s_{ttt}-\frac{uh^2}{6}s_{xxx}+\ldots \tag{1.77}$$

This is of the form $As=E_{(h,k)}s$ with A the advection operator $A=\partial_t+u\partial_x$ and $E_{(h,k)}$ an operator giving the deviation of the modified equation from the advection equation. Note that if $k=h=0$ we obtain the advection equation for which the scheme (1.74) is $O(h,k)$ accurate. We can interpret (1.77) as stating that the scheme (1.74) is:

(1) first order accurate for

$$s_t+us_x=0 \tag{1.78}$$

(2) second order accurate for

$$s_t+us_x=-\frac{k}{2}s_{tt}+\frac{uh}{2}s_{xx} \tag{1.79}$$

(3) third order accurate for

$$s_t+us_x=-\frac{k}{2}s_{tt}+\frac{uh}{2}s_{xx}-\frac{k^2}{6}s_{ttt}-\frac{uh^2}{6}s_{xxx} \tag{1.80}$$

The equations obtained above are called *modified equations*. These statements can be verfied by explicit computation of the truncation error. For example let us compute the truncation error in applying (1.74) to (**??**)

$$\tau_j^n=\left(\tilde{D}-D\right)s(x_j,t^n) \tag{1.81}$$

The finite difference approximation operator is

$$\tilde{D}s(x_j,t^n)=\frac{s_j^{n+1}-s_j^n}{k}+\frac{u}{h}\left(s_j^n-s_{j-1}^n\right) \tag{1.82}$$

The exact operator for the modified equation (**??**) is

$$D=\frac{\partial}{\partial t}+u\frac{\partial}{\partial x}+\frac{k}{2}\frac{\partial^2}{\partial t^2}-\frac{uh}{2}\frac{\partial^2}{\partial x^2} \tag{1.83}$$

and we have $Ds=0$. We now expand s_j^{n+1}, s_{j-1}^n, s_{j+1}^n from (1.82) around (x_j,t^n) and obtain

$$\tau_j^n=\frac{1}{k}\left(s+ks_t+\frac{k^2}{2}s_{tt}+\ldots\right)-\frac{1}{k}s+\frac{u}{h}\left(s-s+hs_x-\frac{h^2}{2}s_{xx}+\frac{h^3}{6}s_{xxx}+\ldots\right) \tag{1.84}$$

$$\tau_j^n=s_t+\frac{k}{2}s_{tt}+us_x+\frac{uh}{2}s_{xx}+O(k^2,h^2)=Ds+O(k^2,h^2)=O(k^2,h^2) \tag{1.85}$$

so the truncation error is indeed of second order.

Now let us show the benefits of looking at the modified equation by using (**??**) for which the upwind scheme (1.77) is second order accurate. First we recast (**??**) so as to eliminate higher order derivatives in time. We can rewrite (**??**) as

$$s_t = -us_x - \frac{k}{2}s_{tt} + \frac{uh}{2}s_{xx} \tag{1.86}$$

and differentiate with respect to t to obtain

$$s_{tt} = -us_{xt} - \frac{k}{2}s_{ttt} + \frac{uh}{2}s_{xxt} \tag{1.87}$$

Replacing (1.87) in (**??**) gives

$$s_t + us_x = -\frac{k}{2}\left(-us_{xt} - \frac{k}{2}s_{ttt} + \frac{uh}{2}s_{xxt}\right) + \frac{uh}{2}s_{xx} \tag{1.88}$$

$$= \frac{uk}{2}s_{xt} + \frac{uh}{2}s_{xx} + O(k^2, h^2, kh) \tag{1.89}$$

We can neglect the higher order terms since this is consistent with the order of accuracy used in obtaining (**??**). Differentiating (**??**) with respect to x yields

$$s_{tx} = -us_{xx} - \frac{k}{2}s_{ttx} + \frac{uh}{2}s_{xxx} \tag{1.90}$$

and replacing this in (**??**) gives

$$s_t + us_x = -\frac{u^2k}{2}s_{xx} + \frac{uh}{2}s_{xx} = \frac{uh}{2}(1-\nu)s_{xx} \tag{1.91}$$

Equation (1.91) is the usual way to express the modified equation for the upwind scheme applied to the advection equation to second order. It shows that the upwind scheme does indeed model the advection equation in the limit $h \to 0$. For finite step sizes however the upwind scheme more accurately models the equation (1.91). The difference between (1.91) and the advection equation is the term

$$\frac{uh}{2}(1-\nu)s_{xx} \tag{1.92}$$

Note that this is a diffusive term whose effect is to smooth out any variations in $s(x,t)$ as long as $|1-\nu| \geq 0$ as has been seen in the study of the heat equation. The condition $|1-\nu| \geq 0$ is exactly the stability criterion for the upwind scheme. Indeed if $\nu > 1$ then we would obtain a negative diffusion coefficient for which the initial value problem is ill-defined. We can see that at exactly $\nu = 1$ there is no diffusion indicating that for $\nu = 1$ the upwind scheme achieves higher order accuracy for the advection equation. When $\nu < 1$ the error in the upwind scheme with respect to the true solution $q(x,t)$ of the advection equation will be diffusive: gradients will be smoothed out instead of being simply advected.

Now that we have seen the nature of the error introduced by the upwind scheme applied to the advection equation, we can also use this information to derive better schemes. Since the error is known to be given by (1.92) we can change the upwind scheme

$$Q_j^{n+1} = Q_j^n - \nu\left(Q_j^n - Q_{j-1}^n\right) \tag{1.93}$$

to counteract the known error by including a discretization of (1.92)

$$Q_j^{n+1} = Q_j^n - \nu\left(Q_j^n - Q_{j-1}^n\right) - k\frac{uh}{2}(1-\nu)\frac{Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n}{h^2} \tag{1.94}$$

Working this through leads to the scheme

$$Q_j^{n+1} = Q_j^n - \frac{\nu}{2}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{\nu^2}{2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right). \tag{1.95}$$

Thus we have obtained the Lax-Wendroff scheme (1.30) via another route.

The procedure can be continued to higher orders. We can now ask what is the modified equation more accurately described by the Lax-Wendroff scheme. Repeating the same procedures as above we first write

$$s(x,t+k) = s(x,t) - \frac{\nu}{2}\left[s(x+h,t) - s(x-h,t)\right] + \frac{\nu^2}{2}\left[s(x+h,t) - 2s(x,t) + s(x-h,t)\right] \tag{1.96}$$

and then carry out Taylor series expansions around (x,t) to obtain

$$s + ks_t + \frac{k^2}{2}s_{tt} + \frac{k^3}{6}s_{ttt} + \ldots = s - \frac{uk}{2h}\left[2hs_x + \frac{h^3}{3}s_{xxx} + \ldots\right] + \tag{1.97}$$

$$\frac{u^2k^2}{2h^2}\left[h^2 s_{xx} + \frac{h^4}{12}s_{xxx}\right] \tag{1.98}$$

from where

$$s_t + us_x = -\frac{k}{2}\left(s_{tt} - u^2 s_{xx}\right) - \frac{k^2}{6}s_{ttt} - \frac{uh^2}{6}s_{xxx} + O(k^3, h^3) \tag{1.99}$$

Note the appearance of the $O(k)$ term. Had we carried out the Taylor expansion for the advection equation this term would have been proportional to $q_{tt} - u^2 q_{xx}$ which is zero according to (1.26). Here we cannot assume that $s_{tt} - u^2 s_{xx}$ is zero a priori. We must carry the term in the ensuing computations, expecting that it will give a higher order correction. Let us neglect the $O(k^3, h^3)$ contributions and proceed with our technique of replacing higher order time derivatives with spatial derivatives using

$$s_{tt} = -us_{xt} - \frac{k}{2}\left(s_{ttt} - u^2 s_{xxt}\right) \tag{1.100}$$

$$s_{ttt} = -us_{xtt}. \tag{1.101}$$

Higher order terms have been dropped since they would lead to $O(k^3, h^3)$ contributions in (1.99). Our intermediate result is

$$s_t + us_x = -\frac{k}{2}\left[-us_{xt} - \frac{k}{2}\left(s_{ttt} - u^2 s_{xxt}\right) - u^2 s_{xx}\right] + \frac{k^2}{6}us_{xtt} - \frac{uh^2}{6}s_{xxx} \tag{1.102}$$

and we continue by eliminating mixed derivatives. In the above formula we wish to express s_{xtt} in terms of x derivatives to $O(1)$

$$s_{xtt} = s_{ttx} = (s_t)_{tx} = (-us_x)_{tx} = -u\,(s_t)_{xx} = u^2 s_{xxx}\ . \tag{1.103}$$

We also need to express s_{xt} in terms of x derivatives to $O(k,h)$

$$s_{xt} = s_{tx} = -us_{xx} - \frac{k}{2}\left(s_{ttx} - u^2 s_{xxx}\right) \tag{1.104}$$

and s_{ttt}, s_{xxt} to $O(1)$

$$s_{ttt} = -u^3 s_{xxx},\ s_{xxt} = -us_{xxx}\ . \tag{1.105}$$

Replacing in (1.102) leads to

(1.106)
$$s_t+us_x = -\frac{k}{2}\left\{-u\left[-us_{xx}-\frac{k}{2}\left(s_{ttx}-u^2s_{xxx}\right)\right]-u^2s_{xx}-\frac{k}{2}\left(-u^3s_{xxx}+u^3s_{xxx}\right)\right\}+\frac{u}{6}\left(k^2u^2-h^2\right)s_{xxx}$$

which simplifies to

$$s_t + us_x = \frac{k^2}{4}\left(s_{ttx}-u^2s_{xxx}\right)+\frac{u}{6}\left(k^2u^2-h^2\right)s_{xxx} \tag{1.107}$$

Since $s_{ttx} = u^2 s_{xxx}$ to $O(1)$ we obtain in final

$$s_t + us_x = -\frac{uh^2}{6}\left(1-\nu^2\right)s_{xxx}\ . \tag{1.108}$$

The third order derivative now obtained shows that the Lax-Wendroff scheme introduces a *dispersive error* with different wave numbers traveling at different speeds. As expected, the dispersive error is proportional to h^2 since the Lax-Wendroff scheme is second order. A scheme more accurate than Lax-Wendroff could be obtained by adding a correction term modeling the dispersive error. Since this involves a third-order derivative the stencil of the scheme would become wider by at least one unit thereby entailing more computational work.

1.2. Non-linear scalar equations. We have introduced a number of finite difference methods for the simple constant-velocity advection equation (1.1). Of course, there is hardly much need for a numerical method in order to solve (1.1). Rather we have used (1.1) as a model problem to study the properties of numerical schemes on a simple case. We now proceed to consider more complicated problems and investigate how the methods already derived apply to these problems.

A general first order, hyperbolic scalar equation is given by

$$q_t + u(x,t,q)q_x = \sigma(x,t,q) \tag{1.109}$$

where u may be interpretred as local advection velocity that depends in general upon x, t and q. In a wide range of problems equations of the form

$$q_t + f(q)_x = \sigma(x,t,q) \tag{1.110}$$

arise where f is known as the *flux function*. If f is differentiable we can write

$$q_t + f_q q_x = \sigma(x,t,q) \tag{1.111}$$

so f_q plays the role of the local advection velocity. Generally u, f depend on q so that the equations become non-linear in q. Equation (1.110) is said to be in *conservative form* as opposed to (1.109) which is said to be in *non-conservative form*. Generally we say that a PDE is in conservative form when it can be expressed as the space-time divergence of a vector field. For equation (1.110) the vector field would be $(q, f(q))$ and the space-time divergence is $\nabla_{(t,x)}\cdot = (\partial_t, \partial_x)\cdot$ so another way of writing (1.110) is

$$\nabla_{(t,x)} \cdot (q, f(q)) = 0\ . \tag{1.112}$$

An initial value problem is defined by specifying a solution domain along x and an initial condition $q_0(x)$.

FIGURE 3. Characteristic curves for $q_t + e^{x+t} q_x = -\beta q$.

1.2.1. *Solution by characteristics.* We can solve (1.109) using the method of characteristics. We again ask whether there are any special curves Γ within the (x,t) plane on which (1.109) reduces to a simpler form. Along the curves specified by the differential system

$$\frac{dt}{ds} = 1, \ \frac{dx}{ds} = u(x,t,q) \tag{1.113}$$

we do indeed obtain the simpler form

$$\frac{dq}{ds} = \sigma. \tag{1.114}$$

The essential difference with respect to the constant-velocity case is that the curves are no longer simple straight lines but depend on x, t and q. Let us consider some examples in order to see the complications involved.

Variable-velocity advection. Consider the equation

$$q_t + u(x,t) q_x = \sigma \tag{1.115}$$

which describes the advection of the unknown field variable q by an imposed velocity field $u(x,t)$. The velocity field is not influenced by q itself; q is said to be a *passive tracer*. The characteristic curves are given by the ODE

$$\frac{dx}{dt} = u(x,t) \ . \tag{1.116}$$

Note that we are no longer guaranteed that the characteristics exist for all times as they did for the constant-velocity advection equation. This is the case only if u is uniformly Lipschitz.

EXAMPLE 7. *Consider the velocity field*

$$u(x,t) = x + t, \tag{1.117}$$

the initial condition

$$q_0(x) = \sin x, \tag{1.118}$$

and the source term

$$\sigma = -\beta q \ . \tag{1.119}$$

The characteristic curves are

$$x(t) = Ce^t - t - 1 \tag{1.120}$$

which are shown in Fig. (3). At $t = 0$ the characteristic labeled by C passes through the x coordinate $x_0 = C - 1$. Along each characteristic the variable-velocity advection equation reduces to the ODE

$$\frac{dq}{dt} = -\beta q \tag{1.121}$$

which has the solution $q(x,t) = Ae^{-\beta t}$. We have to determine the constant A from the initial conditions. Through any given point (x,t) there passes the characteristic curve labeled by $C = e^{-t}(x+t+1)$. This particular characteristic curve will intersect

FIGURE 4. Solution of $q_t + (x+t)q_x = -\beta q$, $q(x, t=0) = \sin x$ for $\beta = 0.1$ at $t = 0, 0.2, \ldots, 1$.

FIGURE 5. Crossing characteristics for inviscid Burgers equation with initial condition $q_0(x) = \sin x$ (shown in thick line).

the x-axis at $x_0 = C - 1$ and this is the position from which we must take the initial value for q

$$q(x,t) = q_0(e^{-t}(x+t+1)-1)e^{-\beta t} = \sin(e^{-t}(x+t+1)-1)e^{-\beta t} \ . \tag{1.122}$$

We have found the solution to the PDE using the simpler expression of the PDE along the characteristics. The solution can be verified by direct substitution in (1.115) and is depicted in Fig. (4). The initial condition is spread out due to the spreading out of the characteristic curves and attenuated due to the source term σ.

Burgers equation. A model equation used extensively in the study of non-linear equations is

$$q_t + qq_x = 0 \tag{1.123}$$

known as the *inviscid Burgers equation*. It is given in non-conservative form above. In conservative form it becomes

$$q_t + \left(\frac{q^2}{2}\right)_x = 0 \tag{1.124}$$

so the flux function is

$$f(q) = q^2/2 \ . \tag{1.125}$$

The characteristic curves are given by

$$\frac{dx}{dt} = q(x,t) \tag{1.126}$$

and along a characteristic curve Γ equation (1.123) reduces to

$$\left(\frac{dq}{ds}\right)_\Gamma = 0 \ , \tag{1.127}$$

i.e. there is no variation in q along the characteristic. This implies that the slope of each characteristic curve is constant and specified by the initial condition $q(x, t = 0) = q_0(x)$.

The type of difficulties that arise for non-linear equations is immediately apparent from the consideration of simple initial conditions. Consider $q_0(x) = \sin x$. The characteristics are sketched in Fig. 5. The problem is that the characteristic curves cross one another. At such a crossing point it is not apparent what the correct value of q should be since different values are being transported along each of the crossing characteristics.

To get a better idea of what is happening it is useful to simplify the initial condition as much as possible. This leads to the so-called *Riemann problem*

$$q_0(x) = \begin{cases} q_l & x < 0 \\ q_r & x > 0 \end{cases} \tag{1.128}$$

Let us try to solve Burgers equation for this initial condition.

If $q_l > q_r$ characteristics from $x < 0$ will overtake those from $x > 0$. This will occur on some ray from the origin of equation $x = st$. To the left of this separating ray we will observe the value q_l while to the right we will observe the value q_r. The solution is therefore

$$q(x,t) = \begin{cases} q_l & x < st \\ q_r & x > st \end{cases} \tag{1.129}$$

The initial discontinuity propagates at a velocity s. The discontinuity is called a *shock* using the language of compressible gas dynamics and s is the *shock velocity*. The shock velocity can be determined by using the integrating Burgers equation over a domain having the shock as its diagonal $[st_1, st_2] \times [t_1, t_2]$

$$\int_{st_1}^{st_2} \int_{t_1}^{t_2} [q_t + f(q)_x]\, dtdx = 0 \tag{1.130}$$

from where

$$s = \frac{f(q_r) - f(q_l)}{q_r - q_l}\ . \tag{1.131}$$

If $q_l < q_r$ two solutions are possible. We can again have the shock solution (1.129) but also the solution

$$q(x,t) = \begin{cases} q_l & x < q_l t \\ x/t & q_l t \le x \le q_r t \\ q_r & x > q_r t \end{cases} \tag{1.132}$$

called a *rarefaction solution*, again using terms from gas dynamics. This an even worse conundrum, not only can discontinuities arise which invalidate the differentiation operations but multiple solutions seem to be possible. Clearly something is wrong and a way to correct the model that led to equation (**??**) must be found. From the physical point of view certain effects have been neglected, namely the viscosity of the fluid and we might be led to studying the viscous Burgers equation

$$q_t + qq_x = \nu q_{xx} \tag{1.133}$$

as a remedy to the difficulties encountered. This can be done and leads to smooth solutions with very large gradients in the regions where shocks would have formed for the inviscid Burgers equation. These large gradients are difficult to resolve properly requiring very fine grids, much finer than needed elsewhere in the solution domain. So a way that enables us to still work with the inviscid equation is quite useful.

1.2.2. *Weak solutions.* The possibility of crossing characteristic curves is indicative with a breakdown of the modeling assumptions that led to a certain hyperbolic PDE. In this situation one must revisit the method by which a certain PDE is derived and consider the validity of all intermediate hypotheses used in the derivation. Burgers equation serves as a useful example. The PDE

$$q_t + f(q)_x = 0 \tag{1.134}$$

with $f = q^2/2$ was proposed as a model for fluid flow in which the quantity q is conserved but being advected by itself. The correct formulation of a conservation principle is through the integral statement

$$\int_{x_1}^{x_2} [q(x,t_2) - q(x,t_1)]\, dx = -\int_{t_1}^{t_2} [f(q(x_2,t)) - f(q(x_1,t))]\, dt \tag{1.135}$$

the one-dimensional expression of (1.10). In this form one can replace

$$q(x,t_2) - q(x,t_1) = \int_{t_1}^{t_2} \frac{\partial q}{\partial t}(x,t)dt \tag{1.136}$$

$$f(q(x_2,t)) - f(q(x_1,t)) = \int_{x_1}^{x_2} \frac{\partial f}{\partial x}dx \tag{1.137}$$

and obtain Burgers equation by going to the limits $t_2 \to t_1$, $x_2 \to x_1$ if the derivatives $\partial q/\partial t$, $\partial f/\partial x$ exist. However one cannot do this if q is discontinuous. In this case only the integral form (1.135) is valid.

Nonetheless it is typically much more convenient to work with differential equations instead of integral equations. Therefore it is useful to extend the meaning we associate to "q is a solution of a PDE" to cover the case where q might be discontinuous at a few points. This is done through the techniques of the theory of distributions by requiring that q satisfy a certain integral condition. Namely we consider the integral

$$I = \int_0^\infty \int_{-\infty}^{+\infty} \phi\, [q_t + f(q)_x]\, dxdt \tag{1.138}$$

with ϕ a smooth function of finite support and impose $I = 0$. Typically we require that ϕ be at least differentiable. We can integrate by parts to obtain

$$\int_0^\infty \int_{-\infty}^{+\infty} [\phi_t q + \phi_x f(q)]\, dxdt = -\int_{-\infty}^{+\infty} \phi(x,0)q(x,0)dx\ . \tag{1.139}$$

By this technique all differentiation operations on q have been removed. We say that q is a *weak solution* of (1.134) if (2.28) is satisfied for all ϕ from some space of test functions such as $\phi \in C^1(\mathbb{R}\times\mathbb{R})$.

1.2.3. *Difficulties of finite difference methods for non-linear hyperbolic equations.* The possibility of shocks for non-linear hyperbolic equations should alert us to possible difficulties with the finite difference methods we have introduced for the linear advection equation. Since these are based upon Taylor series expansions of $q(x,t)$ and q can be discontinuous, the expansions will break down and not be valid near the discontinuities. Nevertheless, we would expect the methods to be adequate in regions where q is smooth.

Let us see how we would apply the methods to a non-linear equation, taking Burgers equation as an example. One possibility is to interpret q as the local advection velocity u. The upwind method for

$$q_t + qq_x = 0 \tag{1.140}$$

then becomes

$$Q_j^{n+1} = Q_j^n - \begin{cases} Q_j^n \left(Q_j^n - Q_{j-1}^n\right) & \text{if } Q_j^n \geq 0 \\ Q_j^n \left(Q_{j+1}^n - Q_j^n\right) & \text{if } Q_j^n < 0 \end{cases} \tag{1.141}$$

and the Lax-Wendroff methods reads

$$Q_j^{n+1} = Q_j^n - \frac{Q_j^n k}{2h}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{\left(Q_j^n\right)^2 k^2}{2h^2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \ . \tag{1.142}$$

Applying this for a Riemann problem leads to a numerical solution similar to the exact shock solution but with oscillations near the shock (Fig. 1.2.3). There is also a smearing of the shock, instead of sharp discontinuity we have a smoothing of q in the vicinity of the shock. Far from the shock the numerical solution is quite good however. This therefore leads to the search for so-called *high-resolution algorithms* that are able to preserve a high order of accuracy away from discontinuities and also sharply capture discontinuities.

2. Systems of hyperbolic equations

2.1. Linear systems.

2.1.1. *Classification of linear systems.* Consider now that we are interested in the simultaneous time evolution of a number of quantities

$$q = \left[\begin{array}{cccc} q_1 & q_2 & \dots & q_m \end{array}\right]^T \tag{2.1}$$

that satisfy

$$q_t + \mathbf{A}q_x = 0 \tag{2.2}$$

with $\mathbf{A}$ a constant $m \times m$ matrix of real numbers. Such a system is said to be hyperbolic if the eigenvectors of $\mathbf{A}$ form a basis for real m-vectors.

EXAMPLE 8. *The second order wave equation is given in canonical form as*

$$\phi_{tt} - c^2\phi_{xx} = 0 \ . \tag{2.3}$$

It can be reduced to a system of two first-order equations by introducing

$$u = \phi_t, \ v = \phi_x \tag{2.4}$$

We have

$$u_t - c^2 v_x = 0 \tag{2.5}$$

and since $\phi_{xt} = \phi_{tx}$

$$v_t - u_x = 0 \tag{2.6}$$

In vector form we obtain

$$\frac{\partial}{\partial t}\left[\begin{array}{c} u \\ v \end{array}\right] + \left[\begin{array}{cc} 0 & -c^2 \\ -1 & 0 \end{array}\right]\frac{\partial}{\partial x}\left[\begin{array}{c} u \\ v \end{array}\right] = 0 \tag{2.7}$$

or

$$q_t + \mathbf{A}q_x = 0 \tag{2.8}$$

with

$$q = \left[\begin{array}{c} u \\ v \end{array}\right], \ \mathbf{A} = \left[\begin{array}{cc} 0 & -c^2 \\ -1 & 0 \end{array}\right] \tag{2.9}$$

The eigenvalues of $\mathbf{A}$ *are* $\lambda_{1,2} = \pm c$ *and the eigenvectors are*

$$r_1 = \left[\begin{array}{c} c \\ 1 \end{array}\right], \ r_2 = \left[\begin{array}{c} -c \\ 1 \end{array}\right] \tag{2.10}$$

The eigenvectors are independent for $c \neq 0$ and therefore they form a basis for the space of real 2-vectors. The system (2.8) is hyperbolic.

EXAMPLE 9. *Applying the same procedure to the Laplace equation*

$$\phi_{tt} + \phi_{xx} = 0 \tag{2.11}$$

leads to the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{2.12}$$

whose eigenvalues are $\lambda_{1,2} = \pm i$ and eigenvectors are

$$\begin{bmatrix} i \\ 1 \end{bmatrix}, \begin{bmatrix} -i \\ 1 \end{bmatrix} \tag{2.13}$$

These have complex values and are not a basis for real 2-vectors. The system (2.12) is not hyperbolic, it is elliptic.

2.1.2. *Solution by method of characteristics and reduction to diagonal form.* For hyperbolic systems we can apply a procedure similar to that used for systems of ODE's. We can write

$$\mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1} \tag{2.14}$$

with

$$\mathbf{\Lambda} = diag\,\{\lambda_1, \lambda_2, \ldots, \lambda_m\} \tag{2.15}$$

$$\mathbf{T} = [r_1\ r_2\ \cdots\ r_m] \tag{2.16}$$

$$\mathbf{A}r_j = \lambda_j r_j,\ j = 1, 2, \ldots, m\ . \tag{2.17}$$

and write

$$q_t + \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1}q_x = 0\ . \tag{2.18}$$

Introducing the notation

$$w = \mathbf{T}^{-1}q \tag{2.19}$$

we obtain

$$w_t + \mathbf{\Lambda}w_x = 0\ . \tag{2.20}$$

Since $\mathbf{\Lambda}$ is a diagonal matrix, the equations of the original system have been decoupled and we can write the scalar j^{th}component equation

$$w_t^{(j)} + \lambda_j w_x^{(j)} = 0\ , \tag{2.21}$$

for $j = 1, 2, \ldots, m$. These are now simple constant-velocity advection equations for which we know the solution

$$w^{(j)}(x,t) = w_0^{(j)}(x - \lambda_j t) \tag{2.22}$$

with $w_0^{(j)}$ given by the initial conditions on q

$$w_0 = \mathbf{T}^{-1}q_0\ . \tag{2.23}$$

The value of each individual component of $w^{(j)}$ is constant along the family of characteristics $x - \lambda_j t = C_j$. Therefore w are known as the *conservative variables.*

From a knowledge of the conservative variable solution we can recover the solution for the original variables

$$q = \mathbf{T}w \ . \tag{2.24}$$

2.1.3. *Finite difference methods.* The finite difference methods derived for the constant-velocity advection equation can be applied formally to hyperbolic systems also. For example, the Lax-Wendroff scheme is

$$Q_j^{n+1} = Q_j^n - \frac{k}{2h}\mathbf{A}\left(Q_{j+1}^n - Q_{j-1}^n\right) + \frac{k^2}{2h^2}\mathbf{A}^2\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \tag{2.25}$$

There are some new features though due to the fact that there is no longer just a single "advection" or characteristic velocity. Let us try to apply the upwind method to the system (2.8). It is not apparent what the upwind direction should be for q. We can ascertain this for the conservative variables though. We have $\lambda_1 = -c$, $\lambda_2 = c$,

$$r_1 = \begin{bmatrix} c \\ 1 \end{bmatrix},\ r_1 = \begin{bmatrix} -c \\ 1 \end{bmatrix} \tag{2.26}$$

$$T = \begin{bmatrix} c & -c \\ 1 & 1 \end{bmatrix},\ T^{-1} = \begin{bmatrix} \frac{1}{2c} & \frac{1}{2} \\ -\frac{1}{2c} & \frac{1}{2} \end{bmatrix},\ \Lambda = \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix} \tag{2.27}$$

and the conservative variable system is

$$\frac{\partial}{\partial t}\begin{bmatrix} w^{(1)} \\ w^{(2)} \end{bmatrix} + \begin{bmatrix} -c & 0 \\ 0 & c \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} w^{(1)} \\ w^{(2)} \end{bmatrix} = 0 \ . \tag{2.28}$$

This system can be discretized in an upwind manner and we obtain the scheme

$$\left(W^{(1)}\right)_j^{n+1} = \left(W^{(1)}\right)_j^n + \frac{ck}{h}\left[\left(W^{(1)}\right)_{j+1}^n - \left(W^{(1)}\right)_j^n\right] \tag{2.29}$$

$$\left(W^{(2)}\right)_j^{n+1} = \left(W^{(2)}\right)_j^n - \frac{ck}{h}\left[\left(W^{(2)}\right)_j^n - \left(W^{(2)}\right)_{j-1}^n\right] \tag{2.30}$$

In matrix form this reads

$$W_j^{n+1} = (1-\nu)\,W_j^n + \mathbf{C}W_{j-1}^n + \mathbf{D}W_{j+1}^n \tag{2.31}$$

with

$$\nu = \frac{ck}{h},\ \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & \nu \end{bmatrix},\ \mathbf{D} = \begin{bmatrix} \nu & 0 \\ 0 & 0 \end{bmatrix} \tag{2.32}$$

Multiplying by T leads to

$$Q_j^{n+1} = (1-\nu)\,Q_j^n + \mathbf{T}\mathbf{C}\mathbf{T}^{-1}Q_{j-1}^n + \mathbf{T}\mathbf{D}\mathbf{T}^{-1}Q_{j+1}^n \tag{2.33}$$

$$\mathbf{T}\mathbf{C}\mathbf{T}^{-1} = \frac{\nu}{2}\begin{bmatrix} 1 & -c \\ -\frac{1}{c} & 1 \end{bmatrix},\ \mathbf{T}\mathbf{D}\mathbf{T}^{-1} = \frac{\nu}{2}\begin{bmatrix} 1 & c \\ \frac{1}{c} & 1 \end{bmatrix} \tag{2.34}$$

This is the upwind scheme for the system (2.8).

2.2. Non-linear systems.

2.2.1. *Classification.* Non-linear systems are written as

$$q_t + f(q)_x = 0 \tag{2.35}$$

with

$$q = \begin{bmatrix} q_1 & q_2 & \dots & q_m \end{bmatrix}^T,\ f = \begin{bmatrix} f_1 & f_2 & \dots & f_m \end{bmatrix}^T \tag{2.36}$$

The classification of non-linear systems is made in accordance with the properties of the Jacobian of f with respect to q

$$f_q = \begin{bmatrix} \frac{\partial f_1}{\partial q_1} & \frac{\partial f_1}{\partial q_2} & \dots & \frac{\partial f_1}{\partial q_m} \\ \frac{\partial f_2}{\partial q_1} & \frac{\partial f_2}{\partial q_2} & \dots & \frac{\partial f_2}{\partial q_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial q_1} & \frac{\partial f_m}{\partial q_2} & \dots & \frac{\partial f_m}{\partial q_m} \end{bmatrix} \tag{2.37}$$

If the eigenvectors of f_q form a basis for q-vectors the system is said to be hyperbolic, otherwise it is elliptic or parabolic. Note that in this case the eigenvectors typically depend on the variables q themselves so that the same system of equations may be hyperbolic in some regions and elliptic in others. The classification of PDE's as hyperbolic, parabolic and elliptic may be more familiar from the classification of second order equations. Let us show the equivalence of the two usages.

The canonical elliptic second order PDE is the Laplace equation

$$\phi_{tt} + \phi_{xx} = 0\ . \tag{2.38}$$

We reduce it to a system of first-order PDE's by introducing $u = \phi_t,\ v = \phi_x$. The Laplace equation states $u_t + v_x = 0$ and we also have $u_x = v_t$ by the equality of mixed derivatives. These two relations can be written in matrix form as

$$q_t + \mathbf{A} q_x = 0 \tag{2.39}$$

$$q = \begin{bmatrix} u \\ v \end{bmatrix},\ \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = 0 \tag{2.40}$$

The matrix $\mathbf{A}$ has the eigenvalues $\lambda_{1,2} = \pm i$ and eigenvectors $r_{1,2} = \begin{bmatrix} \pm i & 1 \end{bmatrix}$. The eigenvectors $r_{1,2}$ do not form a basis for two-component real vectors such as q so the system is classified as elliptic in accord with the second-order Laplace equation's classification.

The canonic hyperbolic second order PDE is the wave equation

$$\phi_{tt} - \phi_{xx} = 0\ . \tag{2.41}$$

Following the same procedure we arrive at the study of the eigensystem of

$$\mathbf{B} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \tag{2.42}$$

which is given by $\lambda_{1,2} = \pm 1$, $r_{1,2} = \begin{bmatrix} \pm 1 & 1 \end{bmatrix}$. The eigenvectors now do form a basis for two-component real vectors and the system is classified as hyperbolic as expected from the wave equation.

Finally, the typical parabolic equation is

$$\phi_x = \phi_{tt} \tag{2.43}$$

for which we denote $u = \phi_t$ to obtain

$$q_t + Cq_x = \sigma \tag{2.44}$$

with

$$q = \begin{bmatrix} \phi \\ u \end{bmatrix}, \ \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \ \sigma = \begin{bmatrix} u \\ 0 \end{bmatrix} . \tag{2.45}$$

The eigensystem of $\mathbf{C}$ is $\lambda_{1,2} = 0$, $r_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}$, $r_2 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ which does not form a basis for two-component real vectors. Note that in this case the rank of $\mathbf{C}$ is less than the dimension of the system; this is characteristic of parabolic equations.

2.2.2. *Solution by characteristics.* Let $\mathbf{A}$ be the Jacobian matrix for a non-linear hyperbolic system

$$q_t + \mathbf{A}(q)q_x = 0, \tag{2.46}$$

with q a vector with m components. By the definition of a hyperbolic system we know that $\mathbf{A}$ can be represented as

$$\mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^{-1} . \tag{2.47}$$

The difference with respect to the linear system case is that the matrices $\mathbf{T}$, $\mathbf{\Lambda}$ are no longer constant but depend on q and hence on (x, t). Nevertheless, we can follow the same procedure of reduction to characteristic form locally for some neighborhood of a point (x_0, t_0) where $q(x, t) = q_0$. We can write

$$q(x, t) = q_0 + \tilde{q}(x, t) \tag{2.48}$$

where $\tilde{q}$ is the perturbation from the value q_0. System (2.46) can now be written

$$\tilde{q}_t + \mathbf{A}_0\tilde{q}_x = 0 , \tag{2.49}$$

from where we obtain

$$\tilde{w}_t + \mathbf{\Lambda}_0\tilde{w}_x = 0 \tag{2.50}$$

with the perturbation characteristic variables given by

$$\tilde{w} = \mathbf{T}^{-1}\tilde{q}_0 . \tag{2.51}$$

The characteristic system (2.50) leads to the ODE's

$$\frac{d\tilde{w}^{(i)}}{ds_i} = 0, i = 1, \ldots, m . \tag{2.52}$$

where d/ds_i indicates the derivative along the i^{th} characteristic direction whose slope is given by the λ_{0i} eigenvalue of $\mathbf{A}_0$

$$\frac{d}{ds_i} = \frac{\partial}{\partial t} + \lambda_{0i}\frac{\partial}{\partial x} . \tag{2.53}$$

A solution to (2.46) can be found by locally solving the ODE's (2.52). This is the *method of characteristics* for non-linear hyperbolic systems.