

Mat1062: Computational Methods for PDE

Mary Pugh

February 14, 2008

1 Ownership

These notes are built upon those of Rob Almgren who taught an analogous course in 2003. Whatever you learn of value from them is due to him. All mistakes and sources of confusion are to be blamed on me.

2 Hyperbolic equations

Now we will talk about *hyperbolic* equations. As you may recall, in the first lecture we introduced the basic “trichotomy” around which we structured the course. We take the particular case

$$a u_{xx} + 2b u_{xy} + c u_{yy} = \text{lower order terms.} \quad (1)$$

We have taken two dimensions for simplicity; higher dimensions are straightforward. This classification really only applies to *quasi-linear, second-order, scalar* PDEs, but we apply the ideas much more generally.

The coefficients a, b, c define a matrix $Q = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and an associated quadratic form $G(p) = p^T Q p$ for $p \in \mathbb{R}^2$.

- Q is positive definite if $G(p) > 0$ for all $p \in \mathbb{R}^2$. Similarly, it’s negative definite if $G(p) < 0$. If Q is positive or negative definite then the equation (1) is elliptic. Example $u_{xx} + u_{yy} = 0$, with $Q = I$. Solutions are determined completely by the boundary data; in effect the propagation speed is infinite since all points are coupled to all other points.
- If $G(p) \geq 0$ for all $p \in \mathbb{R}^2$ and there is some p_0 such that $G(p_0) = 0$ then equation (1) is parabolic. (Similarly if $G(p) \leq 0$ for all $p \dots$) In

this case we need lower-order terms in equation (1) to get a reasonable problem. Example $u_t = u_{xx}$ (changing y to t), with $Q = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. As discussed, the speed of propagation is a complicated concept; in a strict sense it is infinite but in a more realistic sense information propagates with “diffusive scaling” $x \sim t^{1/2}$. In homework we explored a nonlinear version that gave anomalous scaling $x \sim t^{1/3}$.

- A hyperbolic equation has $G(p) > 0$ for some p and $G(p) < 0$ for others. If $G(p_0) = 0$ then p_0 determines the direction and speed of propagation. For example if $u_{yy} - c^2 u_{xx} = 0$ then

$$Q = \begin{pmatrix} -c^2 & 0 \\ 0 & 1 \end{pmatrix} \implies G(p) = 0 \text{ for } p = \begin{pmatrix} 1 \\ c \end{pmatrix}, \begin{pmatrix} 1 \\ -c \end{pmatrix}.$$

It is frequently possible to write a second-order differential equation as a system of first-order equations. For example, this is always possible for an ordinary differential equation. It is not always possible for a PDE but usually is, since most physical problems are really based on first derivatives. For a second-order PDE, we look for a first-order system of the form

$$u_t + A u_x = 0, \tag{2}$$

where u is now an n -component function of (x, t) , and A is an $n \times n$ matrix. A system derived from a scalar second-order equation will have $n = 2$. I will use t and y interchangeably depending on what kind of equation we are talking about. Then the properties of the system are determined by the eigenvalues of the matrix A .

Elliptic example Consider the Laplace equation $u_{xx} + u_{yy} = 0$. Writing $v = u_y$ and $w = u_x$, we get the system

$$\begin{pmatrix} v \\ w \end{pmatrix}_y + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_x = 0,$$

The matrix A is *skew-symmetric* ($A^T = -A$) and thus has purely imaginary eigenvalues $\pm i$. (This system is in fact the Cauchy-Riemann equations of complex analysis, for $f(z) = v + iw$.) The complex eigenvalues are typical of elliptic problems, and indicate that this reduction is not very useful. Just as we discussed in stability analysis, complex speeds of propagation correspond to exponential growth and decay of Fourier modes, and indicate that it is not a good idea to specify only “initial” data.

Hyperbolic example Let's do the same for the wave equation

$$u_{tt} = c^2 u_{xx}.$$

Setting $v = u_t$ and $w = -cu_x$, we get the first-order system

$$\begin{pmatrix} v \\ w \end{pmatrix}_t + \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_x = 0.$$

Here A is symmetric; it has real eigenvalues $\pm c$ and eigenvectors $(1, \mp 1)^T$. These eigenvalues are exactly the finite speeds of propagation.

Motivated by this example, we widen our definition of hyperbolic problems. Instead of considering second-order scalar equations, we start with systems of the form (2), of any size $n \geq 1$. A *linear hyperbolic system* is one whose matrix A has a full set of eigenvectors and only real eigenvalues. This is often attained by A being *symmetric* but not always.

Soon we will generalize these problems to replace Au by $F(u)$ where F is a nonlinear function $\mathbb{R}^n \rightarrow \mathbb{R}^n$. We denote $A(u) = \nabla_u F$, and define a *nonlinear hyperbolic conservation law* to be a system of the form $u_t + F(u)_x = 0$, whose gradient matrix A has real eigenvalues for all relevant values of u . (Various kinds of degeneracy are also possible and lead to interesting problems.) Linear and nonlinear hyperbolic systems are easily defined in more than one space dimension ($u(x, y, t)$, etc).

The theory of nonlinear hyperbolic systems is one of the centerpieces of applied mathematics. These systems are of great importance in applications, but also it is one of the few types of nonlinear problems that we really understand. In particular, the concept of *weak solution* was first fully developed in this context. Conservation is the key idea that makes it work.

Linear scalar problems

The simplest hyperbolic equation is the linear scalar problem

$$u_t + a u_x = 0, \tag{3}$$

with initial data $u(x, 0) = u_0(x)$, and a constant. The solution is clearly

$$u(x, t) = u_0(x - at) \tag{4}$$

as you may check by differentiation. That is, to know what the solution value u is at a point (x, t) , look back to $t = 0$, but at a location $x = at$. Information travels along the straight *characteristic lines* $x = at + \text{const}$.

It may be that $u_0(x)$ is not a differentiable function—has a discontinuity, say—so that the function $u(x, t)$ defined by (4) does not have derivatives u_t or u_x , and hence it is not clear what the PDE (3) means. But in the same way that we defined solutions of a second-order elliptic problem as the minimizer of an energy involving only first derivatives, we decide that (4) will be the *weak solution* of (3) whether or not u_0 has the required derivatives. This is based on our interpretation of the PDE (3) as representing a physical effect (convection) rather than just a relationship among derivatives.

Boundary conditions Take the domain to be $x > 0$, with Dirichlet boundary data $u(0, t) = g(t)$ on $x = 0$. Does there exist a solution having initial data $u_0(x)$ and this boundary data? Answer: yes, if and only if $a > 0$; this may be understood by tracing the characteristics in the (x, t) -plane.

If $a > 0$, then characteristics leaving $x > 0, t = 0$ head off to positive x . Further, the characteristics leaving the boundary $x = 0, t > 0$ also head into the positive half-line, and they fill in the information missing from the initial data. For any point (x, t) with $x > 0, t > 0$, we can follow back the trajectory $x - at = \text{const}$ to find a unique data point that determines the solution along that characteristic.

But if $a < 0$, then characteristics leaving the initial line $t = 0$ hit the boundary $x = 0, t > 0$. Unless $u_0(-at) = g(t)$ for all $t > 0$, then the values at the two ends of the characteristics are *inconsistent*. There is *no* function $u(x, t)$ that can take both endpoint values that are imposed at the two ends of the characteristic segment.

Nonconstant coefficients Suppose that $a = a(x, t)$ is a function of (x, t) . The equation may be written in one of two ways: either

$$u_t + a(x, t) u_x = 0 \tag{5}$$

or

$$u_t + (a(x, t) u)_x = 0. \tag{6}$$

The second form is conservative, so that $\int_a^b u(x, t) dx$ changes only due to fluxes across the endpoints. If everything is smooth, we may expand (6) as

$$u_t + a(x, t) u_x = -a_x u,$$

so (6) is like (5), with an extra source term that changes u .

We may write (5) as

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

as long as

$$\frac{d\xi}{dt} = a(\xi(t), t). \quad (7)$$

Thus $u(x, t)$ is constant along *curved* characteristic paths, defined by the ODE (7). The value of the solution $u(x, t)$ for $t > 0$ is found by following back the curve $\xi(t)$ to its initial value $u_0(\xi(0))$. Note that (unless $a(\xi, t)$ is a really bad function) this evolution is uniquely defined forward and backward in time. Thus characteristics never cross, and this procedure uniquely defines the solution $u(x, t)$ everywhere. Boundary conditions may be handled as above.

The conservative form (6) may be written

$$\frac{d}{dt}u(\xi(t), t) = -a_x(\xi(t), t) u(\xi(t), t).$$

This equation also propagates data along the same characteristic curves, but with a nonzero right-hand side that modifies the value of u as the characteristics come closer together or move apart, so as to preserve conservation.

Systems of equations

Now consider $u_t + Au_x = 0$ where A is a given $n \times n$ matrix, constant for now. Let $V = (v_1, \dots, v_n)$ be the eigenvectors of A , and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ its eigenvalues, so that $AV = V\Lambda$. Then writing $u(x, t) = Vw(x, t)$, we see that

$$w_t + V^{-1}AV w_x = 0.$$

Since $V^{-1}AV = \Lambda$, this is equivalent to the n scalar equations

$$(w_j)_t + \lambda_j (w_j)_x = 0, \quad j = 1, \dots, n.$$

That is, the $n \times n$ linear system (2) has n characteristic speeds, given by the eigenvalues of A . The eigencomponents w_j propagate without interaction (unless there are boundaries in which case the boundary conditions could do things like transfer energy from one eigenvector to others).

If $A = A(x, t)$, then as above, the characteristic speeds $\lambda(x, t)$ and eigenvectors depend on space and time. We then get the system

$$w_t + \Lambda w_x = B(x, t) w, \quad B = -V^{-1} (V_t + AV_x).$$

The linear term on the right side couples the different eigenmodes together. For example, this is how linear waves reflect from gradients in material density or wave speed.

Numerical Methods

As usual, let us construct finite-difference methods on a grid with space step h and time step k , denoting by u_j^n our approximation to the solution value at $x = jh$, $t = nk$. We consider only the scalar first-order problem (3), with a constant. Extensions to the vector problem are sometimes obvious, sometimes not.

Explicit upwind We discussed this briefly when we talked about stability for parabolic problems. Approximate the derivatives as

$$\frac{u_j^{n+1} - u_j^n}{k} + a \frac{u_j^n - u_{j-1}^n}{h} = 0$$

giving the explicit scheme

$$u_j^{n+1} = u_j^n - \mu(u_j^n - u_{j-1}^n), \quad \mu = \frac{ak}{h}.$$

It is clear that this scheme is first-order accurate in h and k .

To study stability, we look for exact solutions in terms of Fourier modes $u_j^n = \omega^j \eta^n$. As before, we find the dispersion relation

$$\eta = 1 - \mu(1 - \bar{\omega}),$$

which is stable if $0 \leq \mu \leq 1$. In Figure 1 you can see what explicit upwinding does to a shock wave.

To find the truncation error, we assume $u(x, t)$ is a smooth solution of $u_t + au_x = 0$ and find

$$\begin{aligned} u(x_j, t_{n+1}) - u(x_j, t_n) + \mu(u(x_j, t_n) - u(x_{j-1}, t_n)) \\ = \frac{1}{2}k^2 u_{tt}(x_j, t_n) - \frac{a}{2}kh u_{xx}(x_j, t_n) + H.O.T. \end{aligned}$$

And so the truncation error is $\mathcal{O}(k^2, hk)$. This means that if one computes to the final time T using $M = T/k$ steps the final maximum error will be $\mathcal{O}(k, h)$. As a result, when testing convergence the natural “refinement path” is to keep μ constant as $h, k \rightarrow 0$ and the ratios of the errors will tend to 2 if one halves k and h with each refinement.

This scheme assumes $a \geq 0$; for $a < 0$ we need to take the difference on the other direction. For a system with a matrix A , it is not clear on which side we should take the difference, unless all the eigenvalues are of one sign.

The CFL condition The restriction $0 \leq \mu \leq 1$ can easily be understood in terms of the *CFL condition*: the domain of dependence of the PDE must be included in the domain of dependence of the discrete scheme. If the scheme has no way for information to get where it needs, then the solution necessarily goes unstable.

In this case, since information propagates one grid step per time step, grid point (jh, nk) receives information from the initial data in the range $((j-n)h, 0)$ to $(jh, 0)$: the domain of dependence of this discrete scheme is $[(j-n)h, jh]$. Analytically, the solution at (jh, nk) is determined by the initial data at the point $(jh - ank, 0) = (jh - n\mu h, 0)$. The CFL condition is then: $(j-n)h \leq jh - n\mu h \leq jh$. Since we've assumed that $a > 0$ the CFL condition is $0 \leq \mu \leq 1$. Another way to have seen this is to simply look at the n th and $n+1$ st time levels. The discrete solution at $(jh, (n+1)k)$ depends on the discrete solution at $((j-1)h, nk)$ and (jh, nk) . The analytic solution is determined by $u(jh - ak, nk) = u(jh - \mu h, nk)$. The CFL condition is then: $jh - h \leq jh - \mu h \leq jh$ again yielding $0 \leq \mu \leq 1$.

Note that the time-step constraint is $\mu \leq 1 \implies k \leq h/a$. The timestep is linearly bounded by h . This condition is typical of hyperbolic problems, as opposed to the quadratic $k \leq Ch^2$ for explicit methods for parabolic problems. Since it is fairly mil, implicit methods are much less popular.

Implicit upwind Although it is not widely used in practice (I have used this in a problem where the propagation speeds were very fast), let us consider the natural generalization

$$u_j^{n+1} = u_j^n - \mu(u_j^{n+1} - u_{j-1}^{n+1}),$$

which requires solving a “bidiagonal” linear system at each step. Since the matrix is lower-triangular, it is easily solved by forward substitution. Thus each computed value u_j^{n+1} influences values $u_{j+1}^{n+1}, u_{j+2}^{n+1}, \dots$ to the *right*, but none to the left. In each time step, information from a particular point u_j^n can propagate arbitrarily rapidly to the right, but not at all to the left.

More formally, the dispersion relation is

$$\eta = \frac{1}{1 + \mu(1 - \bar{\omega})},$$

which is stable for any $\mu \geq 0$ and unstable for $\mu < 0$, so the qualitative explanation above is correct.

Centered difference If you don't want to worry about the sign of a , a natural idea is to take

$$u_j^{n+1} = u_j^n - \frac{\mu}{2} (u_{j+1}^n - u_{j-1}^n),$$

which satisfies the CFL condition if $|\mu| \leq 1$. The dispersion relation is

$$\eta = 1 - i\mu \operatorname{Im} \omega$$

which is *unstable for any* μ . And so satisfying the CFL condition is not sufficient for stability.

Lax-Friedrichs How can we use a symmetric difference formula so as to avoid worrying about the sign of a ? We can stabilize the centered difference formula by making a small modification: replacing u_j^n by the average of the two values on either side. This gives

$$\begin{aligned} u_j^{n+1} &= \frac{1}{2}(u_{j-1}^n + u_{j+1}^n) - \frac{\mu}{2}(u_{j+1}^n - u_{j-1}^n) \\ &= u_j^n - \frac{\mu}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{1}{2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n). \end{aligned}$$

Thus in effect, we add a diffusive term to the right side, to damp the oscillations introduced by the centered difference. In Figure 1 you can see what Lax-Friedrichs does to a shock wave.

The dispersion relation is

$$\eta = \operatorname{Re} \omega - i\mu \operatorname{Im} \omega$$

which is stable for $|\mu| \leq 1$ (it is an ellipse centered at 0, of major radius 1 and minor radius μ). The truncation error is

$$\frac{k^2}{2}u_{tt}(x_j, t_n) - \frac{h^2}{2}u_{xx}(x_j, t_n) + H.O.T.$$

and so the truncation error is $\mathcal{O}(k^2, h^2)$. This means that if one takes the “refinement path” of keeping μ constant as $h, k \rightarrow 0$ then if one computes to the final time T using $M = T/k$ steps the final maximum error will be $\mathcal{O}(k, h) = \mathcal{O}(h)$. As a result, the ratios of the errors will tend to 2 if one halves k and h with each refinement.

Again, when testing convergence the natural “refinement path” is to keep μ constant as $h, k \rightarrow 0$ and the ratios of the errors will tend to 2 if one halves k and h with each refinement.

In this way, we see that Lax-Friedrichs is no more accurate than explicit upwinding but at least we don't need to know the sign of a a priori.

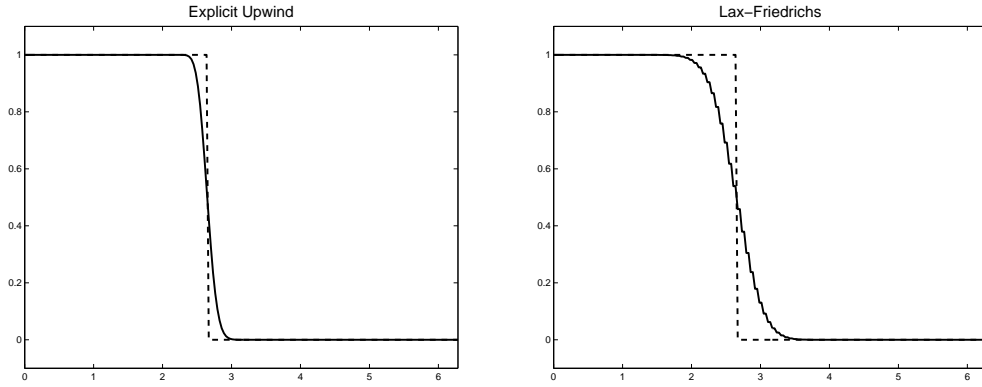


Figure 1: The advection equation $u_t + au_x = 0$ on $[0, 2\pi]$. The advection speed is $a = 1/2$. There are 200 subintervals in space ($h = \pi/100$). The initial data is a step function that jumps at $x = \pi - 1$. The time-step is $k = 1/100$. The approximate solutions (solid lines) and exact solutions (dashed lines) are shown at $t = 1$. Left: the explicit upwind method. Right: the Lax-Friedrichs method. Note that the Lax-Friedrichs method has smeared the shock out more.

Lax-Wendroff We know that if u is the analytic solution then $u(x_j, t_{n+1}) = u(x_j - ak, t_n) = u(x_j - \mu h, t_n)$. And so a natural idea is to approximate $u(x_j - \mu h, t_n)$ using *quadratic* interpolation among the three values $u_{j\pm 1}^n$ and u_j^n . The quadratic function $q(x)$ that takes values $q(\pm h) = u_{j\pm 1}$ and $q(0) = u_j$ is $q(x) = \alpha x^2 + \beta x + \gamma$, with

$$\alpha = \frac{1}{2h^2}(u_{j+1} - 2u_j + u_{j-1}), \quad \beta = \frac{1}{2h}(u_{j+1} - u_{j-1}), \quad \gamma = u_j.$$

Then $u_j^{n+1} \sim u(jh - \mu h, t_n) \sim q(-\mu h)$ and so we set

$$\begin{aligned} u_j^{n+1} &= \frac{\mu(1+\mu)}{2} u_{j-1}^n + (1-\mu^2) u_j^n - \frac{\mu(1-\mu)}{2} u_{j+1}^n \\ &= u_j^n - \frac{\mu}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{\mu^2}{2}(u_{j-1}^n - 2u_j^n + u_{j+1}^n). \end{aligned}$$

This is like the Lax-Friedrichs scheme but with a different coefficient on the diffusive term. In Figure 2 you can see what Lax-Wendroff does to a shock wave.

The dispersion relation is

$$\eta = 1 - i\mu \operatorname{Im} \omega - \mu^2(1 - \operatorname{Re} \omega) = 1 - \mu^2 + \mu^2 \operatorname{Re} \omega - i\mu \operatorname{Im} \omega$$

which is stable for $|\mu| \leq 1$ (it is an ellipse centered on $1 - \mu^2$, with major axis μ^2 and minor axis $|\mu|$).

The truncation error is

$$\frac{1}{6}k^3 u_{ttt}(x_j, t_n) + \frac{a}{6}kh^2 u_{xxx}(x_j, t_n) + H.O.T.$$

and so the truncation error is $\mathcal{O}(k^3, kh^2)$. This means that if one computes to the final time T using $M = T/k$ steps the final maximum error will be $\mathcal{O}(k^2, h^2)$. As a result, when testing convergence the natural “refinement path” is to keep μ constant as $h, k \rightarrow 0$ and the ratios of the errors will tend to 4 if one halves k and h with each refinement.

In this way, we see that the Lax-Wendroff scheme is more accurate and does not require knowledge of the sign of a .

Beam-Warming This method is based on the same idea as Lax-Wendroff, but it assumes that we know the sign of a . If $a > 0$ the quadratic interpolation is taken among the three *upwind* points: u_{j-2} , u_{j-1} , and u_j resulting in:

$$u_j^{n+1} = u_j^n + \frac{\mu}{2} (-u_{j-2}^n + 4u_{j-1}^n - 3f_j^n) + \frac{\mu^2}{2} (u_{j-2}^n - 2u_{j-1}^n + u_j^n)$$

The truncation error is

$$\frac{k^3}{6} u_{ttt}(x_j, t_n) + k \left(\frac{1}{2} a^2 kh - \frac{1}{3} h^2 \right) u_{xxx}(x_j, t_n)$$

leading to the same accuracy as the Lax-Wendroff scheme. Beam-Warming is stable for $0 \leq \mu \leq 2$, which is a wider range (allowing for a larger timestep) at the cost of needing to be certain of the sign of a . In Figure 2 you can see what Beam-Warming does to a shock wave.

A Caveat All of the truncation errors above were made assuming that u is a smooth solution of the advection equation. But as you know, the advection equation has the exact solution $u(x, t) = u_0(x - at)$. Whatever smoothness the initial data has (or doesn't have) will be inherited by the solution at later times. As a result, the convergence arguments are assuming that the initial data is “smooth enough”. If you start with nasty initial data you have no expectation of seeing the desired convergence. Please see the hand-out on the course webpage for more on this.

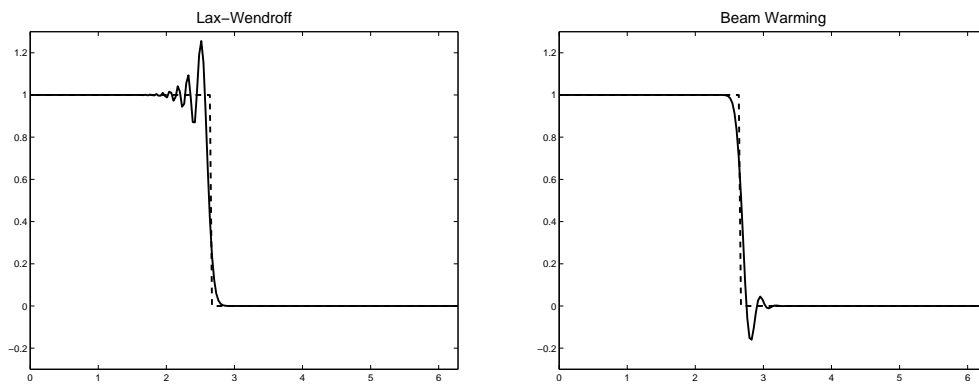


Figure 2: The same as in Figure 1 except that here the left plot shows the discrete solution as computed by Lax-Wendroff and the right plot shows the discrete solution as computed by the Beam-Warming method. Note that both methods have dispersive effects near the shock. Lax-Wendroff has dispersive ripples behind the shock. Beam-Warming has them ahead of the shock.