Mat1062: Introductory Numerical Methods for PDE

Mary Pugh

February 5, 2009

1 Ownership

These notes are the joint property of Rob Almgren and Mary Pugh.

2 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

$$u_t + a u_x = 0 \tag{1}$$

with a constant. Extensions to the vector problem are sometimes obvious, sometimes not.

Explicit upwind 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

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

To study stability, we take the von Neumann approach and look for exact solutions in terms of Fourier modes $u_j^n = \omega^j \eta^n$ where ω is a complex number of unit length. This yields

$$\omega^j \nu^{n+1} = \omega^j \nu^n - \mu(\omega^j \nu^n - \omega^{j-1}\nu^n) \Longrightarrow \eta = 1 - \mu(1 - \bar{\omega}).$$

We find $|\eta| \leq 1$ if $0 \leq \mu \leq 1$. If $\mu < 0$ or $\mu > 1$ then there are unit-length ωs such that $|\nu| > 1$, causing instability. In Figure 1 you can see what explicit upwinding does to a shock wave.

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

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

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 $h, k \to 0$ 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:*[1] the domain of dependence of the PDE must be included in the domain of dependence of the discrete scheme: if the analytic domain of dependence is not contained in the domain of dependence of the discrete scheme then the discrete scheme will be unstable.

Note: the CFL condition *does not* guarantee stability if the analytic domain of dependence is contained in the discrete domain of dependence. Also, while the above is a rigorous definition and works fine for explicit methods applied to linear equations, if you google "CFL condition" you will see that it's short-hand for a broader, vaguer concept. One which reduces to the above when considering explicit schemes and linear hyperbolic equations.

For the Explicit Upwind method, 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

[1] This is introduced in the work by Courant, Friedrichs, and Lewy "On the partial difference equations of mathematical physics", IBM Journal, March 1967, pp. 215-234, English translation of the 1928 German original.

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 nth 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 $\mu \leq 1 \Longrightarrow k \leq h/a$. 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 this time-step constraint is fairly mild, implicit methods are much less popular.

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

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

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

$$\eta \;=\; 1 \;-\; i\mu \,\mathrm{Im}\,\omega$$

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

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

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.

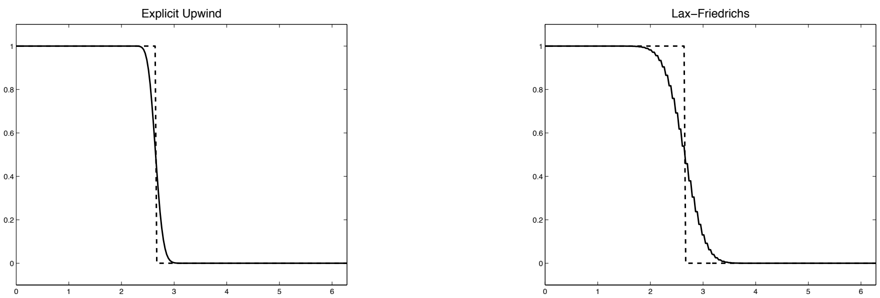

Figure 1: The advection equation $u_t + au_x = 0$ on $[0, 2\pi]$. The advection speed is $a = 1/2$. There are 200 subintervals in space ($h = \pi/100$). The initial data is a step function that jumps at $x = \pi - 1$. The time-step is $k = 1/100$. The approximate solutions (solid lines) and exact solutions (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.

The dispersion relation from the von Neumann stability analysis is

$$\eta \;=\; \mathrm{Re}\,\omega \;-\; i\mu\,\mathrm{Im}\,\omega$$

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

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

and so the trucation error is $\mathcal{O}(k^2, h^2)$. This means that if one takes the "refinement path" of keeping μ constant as $h, k \to 0$ 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.

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.

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 $u^n_{j\pm1}$ and u^n_j. The quadratic function $q(x)$ that takes values $q(\pm h) = u_{j\pm1}$ and

$q(0) = u_j$ is $q(x) = \alpha x^2 + \beta x + \gamma$, with

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

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

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

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 from the von Neumann stability analysis is

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

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 local truncation error is

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

and so the trucation 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 $h, k \to 0$ 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 not only does the Lax-Wendroff scheme not require knowledge of the sign of a but it is more accurate than Explicit Upwind, Implicit Upwind, or Lax-Friedrichs.

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:

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

The truncation error is

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

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.

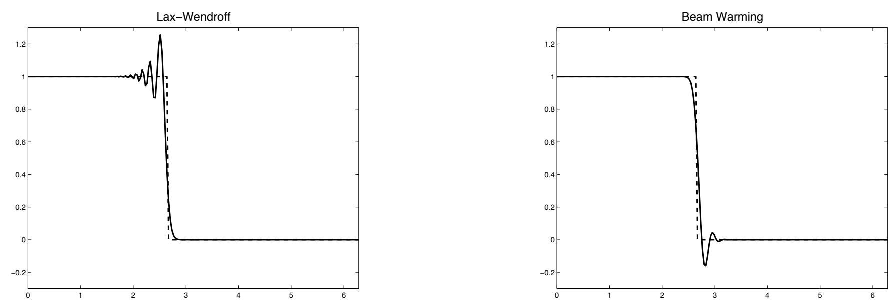

Figure 2: The same as in Figure 1 except that here the left plot shows the discrete solution as computed by Lax-Wendroff and the right plot shows the discrete solution as computed by the Beam-Warming method. Note that both methods have dispersive effects near the shock. Lax-Wendroff has dispersive ripples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

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.

3 Implicit Upwind

Let us consider a natural generalization of explicit upwind — the implicit upwind scheme:

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

The dispersion relation from the von Neumann stability analysis is

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

and $|\eta| \leq 1$ for any $\mu \geq 0$ **and** for any $\mu \leq -1$. However, if $-1 < \mu < 0$ then there are unit-length ωs which result in $|\eta| > 1$.

Okay this is weird. The $\mu \geq 0$ condition isn't too strange — it has to do with the speed being positive $a > 0$ and is what we understood before with upwind schemes. But how can this upwind scheme be stable some $a < 0$?

To understand this, we look at how the scheme would be implemented on the interval $[0, L]$. For the advection equation to be well-posed on a bounded interval, we need to include boundary conditions. If $a > 0$ then we specify $u(0,t)$ for all $t > 0$ and do not specify $u(L,t)$. If $a < 0$ then we specify $u(L,t)$ for all $t > 0$ and do not specify $u(0,t)$.

First, consider the $a > 0$ case. In this case, u_0^n is specified for all $n \geq 0$ and at time level $n+1$ we need to solve for $u_1^{n+1}, u_2^{n+1}, \ldots u_N^{n+1}$. The corresponding N equations are

$$\begin{aligned} &(1+\mu)u_1^{n+1} = u_1^n + \mu u_0^{n+1} \\ &-\mu u_{j-1}^{n+1} + (1+\mu)u_j^{n+1} = u_j^n \qquad 2 \leq j \leq N \end{aligned}$$

In matrix notation this is $AX = B$ where

$$\begin{pmatrix} 1+\mu & & & \\ -\mu & 1+\mu & & \\ & \ddots & \ddots & \\ & & -\mu & 1+\mu \end{pmatrix} \begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_N^{n+1} \end{pmatrix} = \begin{pmatrix} u_1^n + \mu u_0^{n+1} \\ u_2^n \\ \vdots \\ u_N^n \end{pmatrix}.$$

To solve for u^{n+1} we invert A. The inverse of A is a full, lower-triangular matrix. And so u_j^{n+1} depends on $u_0^n, u_1^n, \ldots u_j^n$.

To understand the stability of the scheme, we need to understand the eigenvalues of A^{-1}. If these eigenvalues have magnitude less than or equal to 1 then the scheme is stable. And so the scheme will be stable if the eigenvalues of A have magnitude greater than or equal to 1. Because A is lower triangular, we see its eigenvalues immediately — they all equal $1+\mu$. And so the scheme is stable if $\mu \geq 0$.

Now, consider the $a < 0$ case. In this case, u_N^n is specified for all $n \geq 0$ and at time level $n+1$ we need to solve for $u_0^{n+1}, u_1^{n+1}, \dots u_{N-1}^{n+1}$. The corresponding N equations are

$$\begin{aligned} &-\mu u_{j-1}^{n+1} + (1+\mu) u_j^{n+1} = u_j^n \qquad 1 \leq j \leq N-1 \\ &-\mu u_{N-1}^{n+1} = u_N^n - (1+\mu) u_N^{n+1} \end{aligned}$$

In matrix notation this is $AX = B$ where

$$\begin{pmatrix} -\mu & 1+\mu & & \\ & -\mu & 1+\mu & \\ & & \ddots & \ddots \\ & & & -\mu \end{pmatrix} \begin{pmatrix} u_0^{n+1} \\ u_1^{n+1} \\ \vdots \\ u_{N-1}^{n+1} \end{pmatrix} = \begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_N^n - (1+\mu) u_N^{n+1} \end{pmatrix}.$$

Because A is upper triangular, we see its eigenvalues immediately — they all equal $-\mu$. And so the scheme is stable if $|\mu| \geq 1$; since $a < 0$ this corresponds to $\mu \leq -1$.

To solve for u^{n+1}, we invert A. The inverse of A is a full, upper-triangular matrix. And so u_j^{n+1} depends on $u_{j+1}^n, u_{j+2}^n, \dots u_N^n$. We revisit the CFL condition: the scheme is unstable if the analytic domain of dependence is not in the numerical domain of dependence. That is, the scheme is unstable if

$$x_j - ak < x_{j+1} \Longleftrightarrow -ak < h \Longleftrightarrow \mu > -1.$$

And so the CFL condition $(-1 < \mu < 0)$ would have predicted the instability without considering the matrix A and its eigenvalues.

Note: a more natural implicit scheme would be one based on a centered difference in space.