18.336—Numerical Methods for Partial Differential Equations, Spring 2005

Plamen Koev

September 6, 2012

Contents

1 Hyperbolic PDEs **5**
- 1.1 Consistency, Stability, Well-posedness, and Convergence . . . 5
- 1.2 Fourier Analysis . . . 6
- 1.3 Von Neumann Analysis . . . 7
- 1.4 The Leap-Frog Scheme . . . 9
- 1.5 Dissipation . . . 11
- 1.6 Dispersion . . . 13
- 1.7 Group Velocity and Propagation of Wave Packets . . . 15
- 1.8 Summary of Schemes for the Wave Equation . . . 18

2 Parabolic equations **19**
- 2.1 The Heat Equation . . . 19
- 2.2 The Du Fort–Frankel Scheme . . . 23
- 2.3 The Convection-Diffusion Equation . . . 24
- 2.4 Summary of Schemes for the Heat Equation . . . 25

3 Systems of PDEs in Higher Dimensions **27**
- 3.1 The Equation $u_t + Au_x = 0$ . . . 27
- 3.2 The Equation $u_t + Au_x = Bu$ . . . 28
- 3.3 The Equation $u_t + Au_x + Bu_y = 0$ . . . 28
- 3.4 The equation $u_t = b_1 u_{xx} + b_2 u_{yy}$ . . . 30
- 3.5 ADI methods . . . 31
- 3.6 Boundary conditions for ADI methods . . . 33

4 Elliptic Equations **35**
- 4.1 Steady-State Heat Equation . . . 35
- 4.2 Numerical methods for $u_{xx} + u_{yy} = f$ . . . 38
- 4.3 Jacobi, Gauss–Seidel, and SOR(ω) . . . 39

Chapter 1

Hyperbolic PDEs

We consider the hyperbolic equation

$$u_t + au_x = 0$$

for $t \geq 0$ and initial condition $u(0, x) = u_0(x)$. The unique solution to this problem is given by

$$u(t, x) = u_0(x - at),$$

i.e., the solution is wave traveling right if $a > 0$ and left if $a < 0$.

We will show (a) how to generate schemes for its numerical solution, (b) verify that these schemes are a good approximation to the differential equation (i.e., are consistent) and (c) that the numerical solution converges to the solution to the differential equation.

The idea in using finite differences to solve a PDE is to select a grid in time and space (with meshlengths k and h, respectively) and to approximate the values $u(mk, nh)$ for integer m, n. All all that follows u will denote the **exact solution** to the PDE and

$$v^{(\text{time})}_{(\text{space})} = v^n_m \approx u(mh, kn)$$

will denote the **approximate finite difference** solution.

We will approximate derivatives of a function f as follows:

$$\begin{aligned}
\delta_+ f(x) &= \frac{f(x+h) - f(x)}{h} && \text{forward difference} \\
\delta_- f(x) &= \frac{f(x) - f(x-h)}{h} && \text{backward difference} \\
\delta_0 f(x) &= \frac{f(x+h) - f(x-h)}{2h} && \text{centered difference}
\end{aligned}$$

For a grid function $v = (\ldots, v_{-2}, v_{-1}, v_0, v_1, v_2, \ldots)$ we have:

$$\begin{aligned}
\delta_+ v_m &= \frac{v_{m+1} - v_m}{h} && \text{forward difference} \\
\delta_- v_m &= \frac{v_m - v_{m-1}}{h} && \text{backward difference} \\
\delta_0 v_m &= \frac{v_{m+1} - v_{m-1}}{2h} && \text{centered difference}
\end{aligned}$$

1.1 Consistency, Stability, Well-posedness, and Convergence

Definition 1 (Consistency)**.** *We say that a finite difference scheme $P_{k,h}v = f$ is* consistent *with the PDE $Pu = f$* of order (r, s) *if for any smooth function ϕ*

$$P\phi - P_{k,h}\phi = O(k^r, h^s) \tag{1.1}$$

To verify consistency expand ϕ in Taylor series and make sure (1.1) holds.

Definition 2 (L^2 norm)**.** *For a function $w = (\ldots, w_{-2}, w_{-1}, w_0, w_1, w_2, \ldots)$ on a grid with step size h:*

$$\|w\| = \left(h \sum_{m=-\infty}^{\infty} |w_m|^2 \right)^{1/2}$$

For a function f on the real line:

$$\|f\| = \left(\int_{-\infty}^{\infty} |f(x)|^2 dx \right)^{1/2}$$

Definition 3 (Stability)**.** *A finite one-step difference scheme $P_{k,h} v_m^n = 0$ for a first-order PDE is* stable *if there exist numbers $k_0 > 0$ and $h_0 > 0$ such that for any $T > 0$ there exists a constant C_T such that*

$$\|v^n\| \le C_T \|v^0\|$$

for $0 \le nk \le T, 0 < h \le h_0, 0 < k \le k_0$.

Definition 4 (Well-posedness)**.** *The initial value problem for the first-order PDE $Pu = 0$ is* well-posed *if for any time $T \ge 0$, there is a constant C_T, such that any solution $u(t,x)$ satisfies*

$$\|u(t,x)\| \le C_T \|u(0,x)\|, \qquad \text{for} \quad 0 \le t \le T.$$

Definition 5 (Convergence)**.** *A one-step finite difference scheme approximating a PDE is* convergent *if for any solution to the PDE, $u(t,x)$, and solution to the finite difference scheme v_m^n, such that $v_m^0 \to u(0,x)$ as $mh \to x$, we have $v_m^n \to u(t,x)$ as $(nk, mh) \to (t,x)$ (as $h, k \to 0$).*

Theorem 1 (Lax)**.** *A consistent finite difference scheme for a PDE for which the initial value problem is well-posed is convergent if and only if it is stable.*

Theorem 2 (The Courant–Friedrichs–Lewy Condition)**.** *A necessary condition for stability of the explicit scheme for the hyperbolic equation $u_t + a u_x = 0$:*

$$v_m^{n+1} = \alpha v_{m-1}^n + \beta v_m^n + \gamma v_{m+1}^n$$

with $k/h = \lambda$ held constant is

$$|a\lambda| \le 1.$$

Proof. The solution is $u(t,x) = u_0(x - at)$ and $u(1,0) = u_0(-a)$. The finite difference scheme v_0^n depends only on v_m^0 for $|m| \le n$. Therefore $|hn| \ge |-a|$. Since $kn = 1$, we have $n = 1/k$ and $|n/k| \ge |a|$ or $|a\lambda| \le 1$. □

1.2 Fourier Analysis

Fourier Transform and Inversion formula

- For u defined on $\mathbb{R}$

$$\hat{u}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x} u(x) dx, \qquad u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega x} \hat{u}(\omega) d\omega$$

- For a grid function $v = (\ldots, v_{-2}, v_{-1}, v_0, v_1, v_2, \ldots)$ with grid spacing h (here $\xi \in [-\pi/h, \pi/h]$)

$$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-imh\xi} v_m h, \qquad v_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \hat{v}(\xi) d\xi$$

Theorem 3 (Parseval)**.**

$$\|u(x)\| = \|\hat{u}(\omega)\|, \quad \|\hat{v}\| = \|v\|.$$

(where $\|\hat{v}\|^2 = \int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 d\xi$.)

1.3 Von Neumann Analysis

Provides an uniform way of verifying if a finite difference scheme is stable.

Example 1. *Let's study the forward-time backward-space scheme*

$$\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_m^n-v_{m-1}^n}{h}=0.$$

Rewrite as

$$v_m^{n+1}=(1-a\lambda)v_m^n+a\lambda v_{m-1}^n,$$

where $\lambda=k/h$. *Use the Fourier inversion formula*

$$v_m^n=\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{imh\xi}\hat{v}^n(\xi)d\xi$$

and substitute to obtain

$$\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{imh\xi}\underbrace{\hat{v}^{n+1}(\xi)}_{*}d\xi=\frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h}e^{imh\xi}\underbrace{[(1-a\lambda)+a\lambda e^{-ih\xi}]\hat{v}^n(\xi)}_{**}d\xi.$$

The Fourier transform is unique, so (*) must equal (**):

$$\hat{v}^{n+1}(\xi)=[(1-a\lambda)+a\lambda e^{-ih\xi}]\hat{v}^n(\xi).$$

Denote $g(h\xi)\equiv(1-a\lambda)+a\lambda e^{-ih\xi}$, *called* **amplification factor**. *We have*

$$\hat{v}^n(\xi)=g(h\xi)\hat{v}^{n-1}(\xi)=\ldots=\big(g(h\xi)\big)^n\hat{v}^0(\xi).$$

Now

$$\|v^n\|^2=\int_{-\pi/h}^{\pi/h}|\hat{v}^n(\xi)|^2d\xi=\int_{-\pi/h}^{\pi/h}|g(h\xi)|^{2n}|\hat{v}^0(\xi)|^2d\xi.$$

Therefore $\|v^n\|\le\|v^0\|$ *(i.e., the scheme is stable), if* $|g(h\xi)|\le 1$. *Write* $\theta=h\xi$ *and evaluate* $|g(\theta)|^2$

$$\begin{aligned}|g(\theta)|^2 &= |(1-a\lambda)+a\lambda e^{-ih\xi}|^2\\ &= (1-a\lambda+a\lambda\cos\theta)^2+a^2\lambda^2\sin^2\theta\\ &= (1-2a\lambda\sin^2\tfrac{\theta}{2})^2+4a^2\lambda^2\sin^2\tfrac{\theta}{2}\cos^2\tfrac{\theta}{2}\\ &= 1-4a\lambda(1-a\lambda)\sin^2\tfrac{\theta}{2}.\end{aligned}$$

Thus $|g(\theta)|\le 1$ *if* $0\le a\lambda\le 1$. *Then* $\|v^n\|\le\|v^0\|$ *and the scheme is stable if* $0\le a\lambda\le 1$.

Theorem 4 (Stability condition)**.** *A one-step finite difference scheme is* stable *if and only if there exist positive constants* K,h_0,k_0 *such that*

$$|g(\theta,k,h)|\le 1+Kk$$

for all $\theta, 0<k\le k_0, 0<h\le h_0$. *If* g *is independent of* k, *then the condition is*

$$|g(\theta,k,h)|\le 1.$$

Proof. Assume $g(\theta,k,h)\le 1+Kk$ for some K.

$$\|v^n\|^2=\int_{-\pi/h}^{\pi/h}|g|^{2n}|\hat{v}^0(\xi)|^2d\xi\le(1+Kk)^{2n}\int_{-\pi/h}^{\pi/h}|\hat{v}^0(\xi)|^2d\xi\le(1+Kk)^{2n}\|v^0\|^2$$

Now $nk \le T$ and $(1+Kk)^n \le (1+Kk)^{T/k} \le e^{KT}$, meaning $\|v^n\| \le e^{KT}\|v^0\|$ and the scheme is stable.

Conversely, assume that for any C there exists an interval $[\theta_1, \theta_2]$ such that $|g| \ge 1+Ck$ for $\theta \in [\theta_1, \theta_2]$, $h \in (0, h_0]$, and $k \in (0, k_0]$. Let

$$\hat{v}^0(\xi) = \begin{cases} 0 & \text{if } h\xi \notin [\theta_1, \theta_2], \\ \sqrt{h(\theta_2 - \theta_1)^{-1}} & \text{if } h\xi \in [\theta_1, \theta_2]. \end{cases}$$

Now $\|\hat{v}^0\| = 1$ and

$$\|v^n\|^2 = \int_{-\pi/h}^{\pi/h} |g|^{2n} |\hat{v}^0(\xi)|^2 d\xi = \int_{\theta_1/h}^{\theta_2/h} |g|^{2n} \frac{h}{\theta_2 - \theta_1} d\xi \ge (1+Ck)^{2n} \ge \frac{1}{2} e^{2TC} \|v^0\|^2$$

for n near T/k. Therefore the scheme is unstable if C can be arbitrarily large. If g is independent of h and k, then $|g| \le 1 + Kk$ must hold for any $0 < k \le k_0$, therefore $|g| \le 1$. □

In practice to analyze a finite difference scheme we do not write integrals. Instead we replace v_m^n by $g^n e^{im\theta}$ and solve for g.

Example 2. *Forward-time centered space*

$$\frac{v_m^{n+1} - v_m^n}{k} + a \frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0.$$

$$0 = \frac{g^{n+1} e^{im\theta} - g^n e^{im\theta}}{k} + a \frac{g^n e^{i(m+1)\theta} - g^n e^{im-1\theta}}{2h} = g^n e^{im\theta} \left(\frac{g-1}{k} + a \frac{e^{i\theta} - e^{-i\theta}}{2h} \right).$$

So $g(\theta) = 1 - ia\lambda \sin\theta$, with $\lambda = k/h$. If λ is constant, then $|g(\theta)|^2 = 1 + a^2\lambda^2 \sin^2\theta > 1$ and the scheme is unstable.

Example 3. *The Lax–Wendroff scheme:*

$$v_m^{n+1} = v_m^n - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n) + \frac{a^2\lambda^2}{2}(v_{m+1}^n - 2v_m^n + v_{m-1}^n),$$

so the amplification factor is:

$$\begin{aligned} g(\theta) &= 1 - \frac{a\lambda}{2}(e^{i\theta} - e^{-i\theta}) + \frac{a^2\lambda^2}{2}(e^{i\theta} - 2 + e^{-i\theta}) \\ &= 1 - ia\lambda \sin\theta - a^2\lambda^2(1 - \cos\theta) \\ &= 1 - 2a^2\lambda^2 \sin^2\tfrac{\theta}{2} - ia\lambda \sin\theta \end{aligned}$$

Thus

$$\begin{aligned} |g(\theta)|^2 &= (1 - 2a^2\lambda^2 \sin^2\tfrac{\theta}{2})^2 + \left(2a\lambda \sin\tfrac{\theta}{2} \cos\tfrac{\theta}{2}\right)^2 \\ &= 1 - 4a^2\lambda^2(1 - a^2\lambda^2) \sin^4\tfrac{\theta}{2} \end{aligned}$$

The scheme is stable only when $|g(\theta)| \le 1$, i.e., when $|a\lambda| \le 1$.

Example 4. *For the Crank–Nicolson scheme*

$$v_m^{n+1} = v_m^n - \frac{a\lambda}{4}(v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n)$$

we obtain

$$g(\theta) = \frac{1 - \frac{1}{2}ia\lambda \sin\theta}{1 + \frac{1}{2}ia\lambda \sin\theta} \qquad \textit{thus} \qquad |g(\theta)|^2 = \frac{1 + (\frac{1}{2}a\lambda \sin\theta)^2}{1 + (\frac{1}{2}a\lambda \sin\theta)^2} = 1$$

so this scheme is unconditionally stable.

1.4 The Leap-Frog Scheme

In this section we prove that Leap-frog is stable if and only if $|a\lambda| < 1$. The scheme is

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0, \quad \text{i.e.,} \quad v_m^{n+1} = v_m^{n-1} + a\lambda(v_{m+1}^n - v_{m-1}^n).$$

Write $\delta_0 v_m = \frac{v_{m+1} - v_{m-1}}{2h}$. Then

$$\begin{bmatrix} v_m^{n+1} \\ v_m^n \end{bmatrix} = \begin{bmatrix} -2ka\delta_0 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_m^n \\ v_m^{n-1} \end{bmatrix}.$$

Fourier transform for vectors = Fourier transform in each component:

$$\begin{aligned} \begin{bmatrix} v_m^{n+1} \\ v_m^n \end{bmatrix} &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} d\xi, \\ \begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} &= \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-imh\xi} \begin{bmatrix} v_m^{n+1} \\ v_m^n \end{bmatrix} h \end{aligned}$$

and Parseval for vectors ($|\cdot|$ means the 2-norm for vectors or matrices so we can tell it apart from the L^2-norm)

$$\left\| \begin{bmatrix} v^{n+1} \\ v^n \end{bmatrix} \right\|^2 = h \sum_{m=-\infty}^{\infty} \left| \begin{bmatrix} v_m^{n+1} \\ v_m^n \end{bmatrix} \right|^2 = \int_{-\pi/h}^{\pi/h} \left| \begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} \right|^2 d\xi = \left\| \begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} \right\|^2.$$

Now Fourier of Leap-Frog:

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} d\xi &= \begin{bmatrix} -2ka\delta_0 & 1 \\ 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \begin{bmatrix} \hat{v}^n(\xi) \\ \hat{v}^{n-1}(\xi) \end{bmatrix} d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \begin{bmatrix} -2ka\delta_0 e^{imh\xi}\hat{v}^n(\xi) + e^{imh\xi}\hat{v}^{n-1}(\xi) \\ e^{imh\xi}\hat{v}^n(\xi) \end{bmatrix} d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \begin{bmatrix} -2ka\frac{e^{ih\xi} - e^{-ih\xi}}{2h} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{v}^n(\xi) \\ \hat{v}^{n-1}(\xi) \end{bmatrix} d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \begin{bmatrix} -2ia\lambda\sin(h\xi) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \hat{v}^n(\xi) \\ \hat{v}^{n-1}(\xi) \end{bmatrix} d\xi \end{aligned}$$

Therefore

$$\begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} = \underbrace{\begin{bmatrix} -2ia\lambda\sin(h\xi) & 1 \\ 1 & 0 \end{bmatrix}}_{G(h\xi)} \begin{bmatrix} \hat{v}^n(\xi) \\ \hat{v}^{n-1}(\xi) \end{bmatrix} = G \cdot \begin{bmatrix} \hat{v}^n(\xi) \\ \hat{v}^{n-1}(\xi) \end{bmatrix} = \ldots = G^n \cdot \begin{bmatrix} \hat{v}^1(\xi) \\ \hat{v}^0(\xi) \end{bmatrix}.$$

Parseval now gives

$$\begin{aligned}
\left\| \begin{bmatrix} v^{n+1} \\ v^n \end{bmatrix} \right\|^2 &= \left\| \begin{bmatrix} \hat{v}^{n+1}(\xi) \\ \hat{v}^n(\xi) \end{bmatrix} \right\|^2 \\
&= \int_{-\pi/h}^{\pi/h} \left| (G(h\xi))^n \cdot \begin{bmatrix} \hat{v}^1(\xi) \\ \hat{v}^0(\xi) \end{bmatrix} \right|^2 d\xi \\
&\leq \int_{-\pi/h}^{\pi/h} |G(h\xi)|^n \cdot \left| \begin{bmatrix} \hat{v}^1(\xi) \\ \hat{v}^0(\xi) \end{bmatrix} \right|^2 d\xi \\
&\leq \max_{|h\xi| \leq \pi} |G(h\xi)|^{2n} \left\| \begin{bmatrix} \hat{v}^1(\xi) \\ \hat{v}^0(\xi) \end{bmatrix} \right\|^2 \\
&= \max_{|h\xi| \leq \pi} |G(h\xi)|^{2n} \left\| \begin{bmatrix} v^1 \\ v^0 \end{bmatrix} \right\|^2 .
\end{aligned}$$

Remains to see when the 2-norm of G is bounded. Jordan form: $G = T\Lambda T^{-1}$ and $G^n = T\Lambda^n T^{-1}$. Characteristic polynomial:

$$g^2 + 2ia\lambda \sin(h\xi) g - 1 = 0$$

Λ^n bounded only if the roots (not to be confused with λ): $|\lambda_{1,2}| \leq 1$, but $\lambda_1 \lambda_2 = -1$, so we must have $|\lambda_1| = |\lambda_2| = 1$. Eigenvalues (denote $s \equiv \sin(h\xi)$ for short):

$$\lambda_{1,2} = -ia\lambda s \pm \sqrt{1 - (a\lambda s)^2}.$$

If $|a\lambda| > 1$, then there exists ξ, s.t. $|a\lambda s| > 1$ and both λ_1 and λ_2 are purely imaginary and distinct, so one of them will be > 1 and the other < 1 in magnitude. So we must have $|a\lambda| \leq 1$.

When $|a\lambda| \leq 1$ we have $|\lambda_{1,2}|^2 = (a\lambda s)^2 + 1 - (a\lambda s)^2 = 1$. Therefore both λ_1 and λ_2 are on the unit circle.

If $\lambda_1 \neq \lambda_2$, then Jordan form of G is (exercise):

$$G = \begin{bmatrix} 1 & 1 \\ -\lambda_2 & -\lambda_1 \end{bmatrix} \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} \frac{\begin{bmatrix} -\lambda_1 & -1 \\ \lambda_2 & 1 \end{bmatrix}}{-\lambda_1 + \lambda_2}$$

Exercise: $|A| \leq \|A\|_F = (\sum_{i,j} |a_{ij}|^2)^{1/2}$ (Frobenius norm).

Now

$$|G^n| \leq |T\Lambda^n T^{-1}| \leq |T| \cdot 1 \cdot |T^{-1}| \leq \|T\|_F \|T^{-1}\|_F \leq \sqrt{4} \cdot \frac{\sqrt{4}}{|-2\sqrt{1-|a\lambda|^2}|} \leq \frac{2}{\sqrt{1-|a\lambda|^2}}$$

is nicely bounded. Going back to Parseval

$$\left\| \begin{bmatrix} v^{n+1} \\ v^n \end{bmatrix} \right\| \leq \frac{\sqrt{2}}{(1-|a\lambda|^2)^{1/4}} \left\| \begin{bmatrix} v^1 \\ v^0 \end{bmatrix} \right\|$$

and Leap-frog is stable.

Next case: $\lambda_1 = \lambda_2$. It occurs when $\sin(h\xi) = \pm 1$ and $|a\lambda| = 1$. Assume $a\lambda s = 1$, (the -1 case is analogous).

$$G = \begin{bmatrix} -2i & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ i & 0 \end{bmatrix} \begin{bmatrix} -i & -i \\ & -i \end{bmatrix} \begin{bmatrix} 0 & -i \\ 1 & i \end{bmatrix} = T\Lambda T^{-1}$$

A little unusual to write a Jordan block with $-i$ in position $(1,2)$ but legal and, in this case, convenient.

$$G^n = T(-i)^n \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^n T^{-1} = (-i)^n T \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} T^{-1},$$

i.e., $G(\pm\pi/2)$ will blow up. You'd think that there may be cancellation, but no:

$$n = \left|\begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}\right| = |i^n T^{-1} G^n T| \le |T| \cdot |G^n| \cdot |T|$$

and both $|T|$ and $|T^{-1}|$ are nicely bounded by (say) 2, so $|G^n| \ge n/4 = T/(4k) \to \infty$ as $k \to 0$.

The stability condition is therefore $|a\lambda| < 1$.

1.5 Dissipation

We would expect the wave equation to propagate the initial condition with a constant speed a, including all frequencies that make up that initial condition.

Unfortunately the discrete nature of our data means that instead of the initial condition $u_0(x)$ we have a discrete version of it—v_0^n.

The initial condition $u_0(x)$ is a superposition (in theory) of an infinite number of frequencies (think Fourier expansion), whereas v_0^n only inherits the frequencies $\xi \in [-\pi/h, \pi/h]$. All higher frequencies are ignored by our discrete initial condition. Recall that the Fourier transform $\hat{v}^n(\xi)$ of v^n is only defined for $\xi \in [-\pi/h, \pi/h]$.

Obviously different frequencies are treated differently and we would like to get a better understanding of that treatment. Example is the best way to go here. Consider Lax–Friedrichs:

$$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0,$$

equivalently,

$$v_m^{n+1} = \tfrac{1-a\lambda}{2} v_{m+1}^n + \tfrac{1+a\lambda}{2} v_{m-1}^n.$$

Von Neumann analysis implies

$$\begin{aligned} g(h\xi) &= \cos(h\xi) - ia\lambda \sin(h\xi), \\ |g(h\xi)|^2 &= \cos^2(h\xi) + (a\lambda)^2 \sin^2(h\xi). \end{aligned}$$

Let $\theta = h\xi$ as usual.

We see that $\theta = 0$ and $\theta = \pi$ are not dampened, but all other θ are. Let's observe closely. Pick $a\lambda$ to be (say) $1/2$. Then

$$v_m^{n+1} = \frac{1}{4} v_{m+1}^n + \frac{3}{4} v_{m-1}^n$$

- $\theta = \pi/2$. Then $e^{imh\xi} = e^{im\theta} = e^{im\pi/2} = \{\ldots, 1, 0, -1, 0, 1, 0, -1, 0, 1, \ldots\}$

$n=4$					$1/16$				
$n=3$				$1/8$	0	$-1/8$			
$n=2$			$1/4$	0	$-1/4$	0	$1/4$		
$n=1$		$1/2$	0	$-1/2$	0	$1/2$	0	$-1/2$	
$n=0$	1	0	-1	0	1	0	-1	0	1

- $\theta = \pi$, we have $e^{imh\xi} = e^{im\theta} = e^{im\pi} = \{\ldots, 1, -1, 1, -1, 1, \ldots\} = (-1)^m$.

  We can verify that $v_m^n = (-1)^{m+n}$ is a solution to Lax–Friedrichs, so $\theta = \pi$ is not dampened at all.

We don't really expect good results for wildly oscillating solutions, so we can expect that the higher frequencies will not be well-represented in our calculation. However it is unacceptable for higher frequencies to be less dampened than the middle-range ones.

Another example. Look at Lax–Wendroff: $|g|^2 = 1 - 4a^2\lambda^2(1 - a^2\lambda^2)\sin^4\frac{\theta}{2} \le 1 - \text{const} \cdot \sin^4\frac{\theta}{2}$.

This is very important—says that all frequencies, except $\xi = 0$ (then $\theta = 0$) are decreasing and the highest frequencies are suppressed the most. This is exactly what we want and will call schemes that have this property dissipative.

Definition 6 (Dissipative Scheme)**.** *A scheme is dissipative of order* $2r$ *if*

$$|g(\theta)| \leq 1 - c \cdot \sin^{2r} \tfrac{\theta}{2}.$$

The reason we like dissipative schemes is that if we are not doing a good job with the high frequencies anyway, why not kill them.

Remark 1. *A dissipative scheme is always stable.*

How can we make a non-dissipative scheme dissipative? This calls for another example. Crank–Nicolson, which is second order accurate.

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n}{4h} = 0$$

So adding a fourth derivative in there will not affect the order of accuracy of the approximation, since fourth derivatives get ignored anyway. When we do the Fourier analysis the fourth derivative will bring a $\sin^4 \frac{\theta}{2}$.

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n}{4h} + C\epsilon\frac{v_{m-2}^n - 4v_{m-1}^n + 6v_m^n - 4v_{m+1}^n + v_{m+2}^n}{h^4} = 0$$

Now select C appropriately so that the fourth derivative only brings $\sin^4 \frac{\theta}{2}$ into the picture, without any weird powers of k and h: $C = \frac{h^4}{16k}$. Then after some simplification

$$g(\theta) = \frac{1 - \epsilon \sin^4 \frac{\theta}{2} - i\frac{a\lambda}{2}\sin\theta}{1 + i\frac{a\lambda}{2}\sin\theta}$$

implying

$$\begin{aligned}
|g(\theta)|^2 &= \frac{1 + \left(\frac{a\lambda}{2}\sin\theta\right)^2 - 2\epsilon\sin^4\frac{\theta}{2} + \epsilon^2\sin^8\frac{\theta}{2}}{1 + \left(\frac{a\lambda}{2}\sin\theta\right)^2} \\
&= 1 - \frac{\epsilon\sin^4\frac{\theta}{2} - \overbrace{\epsilon\sin^4\frac{\theta}{2}(1 - \epsilon\sin^4\frac{\theta}{2})}^{>0 \text{ for } \epsilon<1}}{1 + \left(\frac{a\lambda}{2}\sin\theta\right)^2} \\
&\leq 1 - \frac{\epsilon\sin^4\frac{\theta}{2}}{1 + \left(\frac{a\lambda}{2}\sin\theta\right)^2}.
\end{aligned}$$

If we now restrict $|a\lambda|$ (say) ≤ 10, then $1 + \left(\frac{a\lambda}{2}\sin\theta\right)^2 \leq 26$, and

$$|g|^2 \leq 1 - \frac{\epsilon}{26}\sin^4\tfrac{\theta}{2}.$$

We want a bound on $|g|$, not $|g|^2$:

$$|g|^2 \leq 1 - \frac{\epsilon}{26}\sin^4\tfrac{\theta}{2} \leq |g|^2 \leq 1 - \frac{\epsilon}{26}\sin^4\tfrac{\theta}{2} + \left(\frac{\epsilon}{52}\right)^2\sin^8\tfrac{\theta}{2} = \left(1 - \frac{\epsilon}{52}\sin^4\tfrac{\theta}{2}\right)^2,$$

so

$$|g| \leq 1 - \frac{\epsilon}{52}\sin^4\tfrac{\theta}{2}.$$

The scheme is all of a sudden dissipative of order 4 (since $2r = 4$). Although Crank–Nicolson is stable for all $a\lambda$ we cannot make it dissipative without restricting $a\lambda$.

The exact same trick works for Leap-frog.

1.6 Dispersion

In this section we investigate whether in the numerical solution of $u_t + au_x = 0$ different frequencies travel with the same speed a as they should. They, of course, do not and we will see that in fact travel with speed $\alpha(h\xi) \approx a$.

Look for a solution to

$$u_t + au_x = 0, \qquad u(0,x) = f(x)$$

(which has a unique solution $u(t,x) = f(x-at)$) using separation of variables

$$u(t,x) = g(t)e^{ix\xi},$$

assuming $u(0,x) = g(0)e^{ix\xi} = e^{ix\xi}$ =periodic wave (here we assume $g(0) = 1$). Then

$$u_t + au_x = g'(t)e^{ix\xi} + ag(t)i\xi e^{ix\xi} = (g'(t) + ai\xi g(t))e^{ix\xi} = 0.$$

Since $|e^{ix\xi}| = 1$ we have $g'(t) + ai\xi g(t) = 0$, which implies $g(t) = e^{-iat\xi}g(0)$. Insert back and get

$$u(t,x) = g(0)e^{-iat\xi}e^{ix\xi} = e^{i(x-at)\xi},$$

since $g(0) = 1$.

Therefore the initial condition is translated with speed a for all ξ.

Example 5. *Same Fourier analysis can be used for other equation also to study the speed of different frequency waves, e.g.,*

$$u_t + au_x + u_{xxx} = 0.$$

For $u(t,x) = g(t)e^{ix\xi}$ we get

$$u_t + au_x + u_{xxx} = (g' + i\xi(a - \xi^2)g)e^{ix\xi} = 0,$$

thus

$$g' + i\xi(a-\xi^2)g = 0, \Rightarrow g(t) = e^{-i\xi(a-\xi^2)t}g(0),$$

so the solution is

$$u(t,x) = e^{i(x-(a-\xi^2)t)\xi}g(0).$$

now the speed of the waves depends on ξ.

Definition 7 (Dispersion)**.** *The phenomenon of waves with different frequencies moving with different speeds is called* **dispersion**.

Return now to the solution of the difference equation. Take Lax–Friedrichs:

$$v_m^{n+1} = \frac{1}{2}(v_{m+1}^n + v_{m-1}^n) - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n).$$

Separation of variables: $v_m^n = g^n e^{imh\xi}$ and substitute above to get

$$g = \cos(h\xi) - ia\lambda\sin(h\xi),$$

so the solution is

$$v_m^n = (\cos(h\xi) - ia\lambda\sin(h\xi))^n e^{imh\xi},$$

which looks nothing like $e^{i(x-at)\xi}$. Let

$$g(h\xi) \equiv \rho e^{-i\omega} = \rho\cos\omega - \rho i\sin\omega = \cos(h\xi) - ia\lambda\sin(h\xi).$$

Therefore $\tan\omega = a\lambda\tan(h\xi)$, $\quad \rho^2 = \cos^2(h\xi) + a^2\lambda^2\sin^2(h\xi)$. For $|h\xi| \le \pi/2$ we have

$$v_m^n = (\rho e^{-i\omega})^n e^{imh\xi} = \rho^n e^{imh\xi - i\omega n} = \rho^n e^{i(mh - \omega n/\xi)\xi} = \rho^n e^{i(x - \frac{\omega n}{\xi}\frac{t}{nk})\xi} = \rho^n e^{i(x - \frac{\omega}{k\xi}t)\xi} = \rho^n e^{i(x - \alpha(h\xi)t)\xi},$$

where $x = mh, t = nk$ and

$$\alpha(h\xi) \equiv \frac{\omega}{k\xi} = \frac{\arctan(a\lambda\tan(h\xi))}{\lambda h\xi}$$

(Recall $\tan\epsilon \approx \epsilon$ and $\arctan\epsilon \approx \epsilon$ so $\alpha \approx \frac{\arctan(a\lambda\tan(h\xi))}{\lambda h\xi} \approx a$.)

We have

$$v_m^n = |g(h\xi)|^n \cdot e^{i(x - \alpha(h\xi)t)\xi}.$$

Definition 8 (Phase speed)**.** *The quantity $\alpha(h\xi)$ is called* **phase speed***, and is the speed at which waves of frequency ξ are propagated by the difference scheme.*

Once again, waves with different frequencies travel with different speeds. Thus we say that the scheme is *dispersive*. We want the scheme to be dispersive as little as possible (i.e., $\alpha(h\xi) \approx a$), so that the numerical solution looks like the exact solution.

Time to study the Taylor series for $\alpha(h\xi)$ to obtain a better estimate of the closeness to a.

$$\begin{aligned} \tan z &= z + \frac{1}{3}z^3 + O(z^5) \\ \arctan z &= z - \frac{1}{3}z^3 + O(z^5) \end{aligned}$$

Let $z = h\xi$

$$\begin{aligned} \alpha(z) &= \frac{\arctan(a\lambda\tan z)}{\lambda z} \\ &= \frac{a\lambda\tan z}{\lambda z} - \frac{(a\lambda\tan z)^3}{3\lambda z} + \dots \\ &= a\cdot\frac{z + z^3/3 + \dots}{z} - \frac{a^3\lambda^3}{\lambda}\frac{z^3}{3z} + \dots \\ &= a\left(1 + (1 - a^2\lambda^2)\frac{(h\xi)^2}{3} + \dots\right) \end{aligned}$$

So, if ξ is given and h is small, then the wave speed is slightly higher than a, and the high frequencies travel fastest. Let's look at some special cases.

Take $h\xi = \pi/2$. Then $\omega = \pi/2$ and $\rho = |a\lambda|$, so

$$v_m^n = |a\lambda|^n e^{imh\xi}\cdot e^{-in\pi/2} = |a\lambda|^n e^{i(x\xi - \pi n/2)} = |a\lambda|^n e^{i(x - t/\lambda)\xi}$$

(since $n\pi/2 = (t/k)(h\xi) = t\xi/\lambda$). So the speeds can be quite different. Exact $= a$; Computed $= 1/\lambda = \frac{a}{a\lambda}$, so it is not a good idea to take $a\lambda$ small. The closer to the stability limit (i.e., the closer to 1) the better.

1.7 Group Velocity and Propagation of Wave Packets

Consider the numerical solution to $u_t + au_x = 0$ with initial data $u(0,x) = e^{i\xi_0 x}p(x)$, where $p(x)$ decays rapidly at ∞. The function $u(0,x)$ is called a *wave packet*—see Figure 1.1.

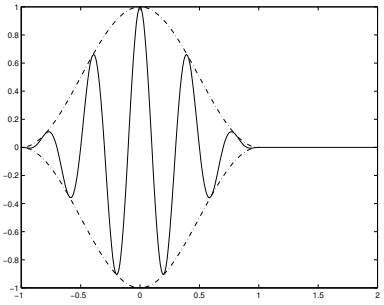

Figure 1.1: Wave packet $\cos 5\pi \cos^2 \frac{\pi x}{2}$ on $[-1,1]$

The exact solution is $u(t,x) = e^{i\xi_0(x-at)}p(x-at)$.

Proposition 1. *A finite difference scheme will have a solution that is approximately*

$$v^*(t,x) = e^{i\xi_0(x-\alpha(h\xi_0)t)}p(x-\gamma(h\xi_0)t),$$

where $\alpha(h\xi_0)$ *is the* phase speed*, and* $\gamma(h\xi_0)$ *is the* **group velocity***, given by*

$$\gamma(\theta) = \alpha(\theta) + \theta\alpha'(\theta).$$

The rest of this section is devoted to proving Proposition 1 and may be skipped on a first reading. Consider a class of one-step schemes with the property

$$\hat{v}^n(\xi) = g(h\xi)\hat{v}^{n-1}(\xi) = \ldots = (g(h\xi))^n\hat{v}^0(\xi).$$

In addition, for simplicity, assume $|g(h\xi)| = 1$. The numerical method will give

$$v_m^n = \frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h} e^{i\xi x_m}(g(h\xi))^n\hat{v}^0(\xi)d\xi.$$

On the other side

$$u(0,mh) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{imh\xi}\hat{u}(0,\xi)d\xi.$$

Split up the interval $(-\infty,\infty) = \bigcup_{l=-\infty}^{\infty}[l\frac{2\pi}{h} - \frac{\pi}{h}, l\frac{2\pi}{h} + \frac{\pi}{h})$ to obtain

$$u(0,mh) = \frac{1}{\sqrt{2\pi}}\sum_l \int_{l\frac{2\pi}{h}-\frac{\pi}{h}}^{l\frac{2\pi}{h}+\frac{\pi}{h}} e^{imh\xi}\hat{u}(0,\xi)d\xi.$$

Set $\xi = l\frac{2\pi}{h} + \xi'$, meaning $d\xi = d\xi'$ and

$$\begin{aligned} u(0,mh) &= \frac{1}{\sqrt{2\pi}}\sum_l \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{imh(l\frac{2\pi}{h}+\xi')}\hat{u}(0,l\frac{2\pi}{h}+\xi')d\xi' \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{imh\xi'}\sum_{l=-\infty}^{\infty}\hat{u}(0,\xi'+l\frac{2\pi}{h})d\xi'. \end{aligned}$$

This gives a formula for

$$\hat{v}^0(\xi) = \sum_{l=-\infty}^{\infty} \hat{u}(0, \xi + l\frac{2\pi}{h}).$$

If u is smooth, then its Fourier transform decays rapidly, and only the middle $l = 0$ term really matters

$$\hat{v}^0(\xi) \sim \hat{u}(0, \xi),$$

with the error bounded by h to some high power depending on the smoothness of $u(0, x)$.

Consider

$$\begin{aligned}
\hat{u}(0,\xi) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ix\xi} u(0,x)dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ix\xi} e^{i\xi_0 x} p(x)dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-ix(\xi-\xi_0)} p(x)dx \\
&= \hat{p}(\xi - \xi_0)
\end{aligned}$$

Let's recall how we handle the phase speed

$$g(h\xi) = |g(h\xi)|e^{-i\omega} = e^{-i\omega} \Rightarrow (g(h\xi))^n = e^{-i\omega n} = e^{-i\frac{\omega n k}{\frac{k}{n} h\xi}\xi} = e^{-i\frac{\omega}{\lambda h\xi}t\xi} = e^{-i\alpha(h\xi)t\xi}.$$

We can return to

$$v_m^n = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{i\xi x_m} \cdot e^{-i\alpha(h\xi)t_n\xi} \cdot \hat{v}^0(\xi)d\xi.$$

Exercise 1. *Verify that $e^{i\xi x_m}, \hat{v}^0(\xi), g(h\xi)$, and ω are periodic functions of ξ with period $\frac{2\pi}{h}$.*

Since $\hat{v}^0(\xi) \sim \hat{u}(0, \xi) = \hat{p}(\xi - \xi_0)$, we change the variables $\phi = \xi - \xi_0$. Then $\xi = \phi + \xi_0$, and

$$v_m^n = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h+\xi_0}^{\pi/h+\xi_0} e^{i(\phi+\xi_0)x_m} \cdot e^{-i\alpha(h(\phi+\xi_0))t_n(\phi+\xi_0)} \cdot \hat{v}^0(\phi + \xi_0)d\phi.$$

Since all functions are periodic with period $\frac{2\pi}{h}$ we can shift the interval of integration back and get

$$v_m^n = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{i(\phi+\xi_0)x_m} \cdot e^{-i\alpha(h(\phi+\xi_0))t_n(\phi+\xi_0)} \cdot \hat{v}^0(\phi + \xi_0)d\phi.$$

This begins to look right. Factor out the phase speed

$$v_m^n = e^{i\xi_0(x_m - \alpha(h\xi_0)t_n)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{i\phi x_m} \cdot e^{-i[(\phi+\xi_0)\alpha(h\phi+h\xi_0)-\xi_0\alpha(h\xi_0)]t_n} \cdot \hat{v}^0(\phi + \xi_0)d\phi.$$

This begins to look like a Fourier transform

$$\sim e^{i\xi_0(x-\alpha(h\xi_0)t)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{i\phi\left(x - \frac{(\phi+\xi_0)\alpha(h\phi+h\xi_0)-\xi_0\alpha(h\xi_0)}{\phi}t\right)} \cdot \hat{v}^0(\phi + \xi_0)d\phi.$$

Since $\hat{v}^0(\xi) \sim \hat{p}(\xi - \xi_0)$ we have $\hat{v}^0(\phi + \xi_0) \sim \hat{p}(\phi + \xi_0 - \xi_0) = \hat{p}(\phi)$. The next step is to replace $\hat{v}^0(\phi + \xi)$ by $\hat{p}(\phi)$. Also since $\hat{p}(\phi)$ goes to zero rapidly as $\phi \to \infty$ we may as well let the integral go to infinity. Hence

$$v_m^n \sim e^{i\xi_0(x-\alpha(h\xi_0)t)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\phi(x-\tilde{\gamma}t)} \cdot \hat{p}(\phi)d\phi = e^{i\xi_0(x-\alpha(h\xi_0)t)} \cdot p(x - \tilde{\gamma}t).$$

The last step is wrong because $\tilde{\gamma}$ depends on ϕ, but it does tell us where we are going. We have

$$\begin{aligned}
\tilde{\gamma} &\equiv \frac{(\phi+\xi_0)\alpha(h\phi+h\xi_0)-\xi_0\alpha(h\xi_0)}{\phi} \\
&= \frac{(h\phi+\xi_0)\alpha(h\phi+h\xi_0)-h\xi_0\alpha(h\xi_0)}{h\phi} \\
&= \frac{(\theta+\theta_0)\alpha(\theta+\theta_0)-\theta_0\alpha(\theta_0)}{\theta} \\
&= \frac{G(\theta+\theta_0)-G(\theta_0)}{\theta} \\
&= G'(\theta_0)+\tfrac{\theta}{2}G''(\theta^*),
\end{aligned}$$

where $\theta \equiv h\phi$, $\theta_0 \equiv h\xi_0$ and $G(\theta) \equiv \theta\alpha(\theta)$ and θ^* is between θ_0 and θ. The beauty of the above expression is that $G'(\theta_0)$ does not depend on ϕ, but only on $h\xi_0 = \theta_0$.

Let's go back to

$$v_m^n \sim e^{i\xi_0(x_m-\alpha(h\xi_0)t_n)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{i\phi(x_m-G'(h\xi_0))t_n} \cdot e^{-i\phi\frac{h\phi}{2}G''(\theta^*)t}\hat{p}(\phi)d\phi.$$

The idea is to replace $e^{-i\phi\frac{h\phi}{2}G''(\theta^*)t}$ by one. By doing so we are making an error bounded by

$$\int_{-\infty}^{\infty} \left| e^{-i\phi\frac{h\phi}{2}G''(\theta^*)t} - 1 \right| \cdot |\hat{p}(\phi)| d\phi.$$

We will now show that this error is at most $O(h)$.

Let's first bound $|\hat{p}(\phi)|$

$$\begin{aligned}
\hat{p}(\phi) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\phi x} p(x) dx \\
\phi^4\hat{p}(\phi) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-i\phi)^4 e^{-i\phi x} p(x) dx; \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\partial^4}{\partial x^4}(e^{-i\phi x}) \cdot p(x) dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\phi x} \cdot p''''(x) dx
\end{aligned}$$

Thus

$$|\hat{p}(\phi)| \le \frac{1}{\sqrt{2\pi}} \cdot \frac{\int_{-\infty}^{\infty} |p''''(x)| dx}{|\phi|^4} = \frac{C}{|\phi|^4}.$$

Also $|e^{iz}-1|^2 = 4\sin^2\frac{z}{2} \le |z|^2$, so $|e^{iz}-1| \le |z|$. By using this estimate we have

$$\begin{aligned}
\int_{-\infty}^{\infty} \left| e^{-i\phi\frac{h\phi}{2}G''(\theta^*)t} - 1 \right| \cdot |\hat{p}(\phi)| d\phi &\le \int_{-\infty}^{\infty} \left| h\phi^2 \frac{1}{2} G''(\theta^*)t \right| \cdot |\hat{p}(\phi)| d\phi \\
&\le h \cdot \text{const} \cdot \int_{-\infty}^{\infty} |\phi^2\hat{p}(\phi)| d\phi \\
&\le h \cdot \text{const} \cdot \int_{-\infty}^{\infty} \left| \frac{1}{\phi^2} \right| d\phi \\
&\le h \cdot \text{const}.
\end{aligned}$$

If we work in L^2 we can bound the error by h^2 in norm, but not pointwise.

Either way we have shown that

$$v_m^n = e^{i\xi_0(x-\alpha(h\xi_0)t)} p(x - G'(h\xi)t) + \text{small terms}.$$

Definition 9. *The quantity* $\gamma(\theta) = G'(\theta) = \alpha(\theta) + \theta \cdot \alpha'(\theta)$ *is called* **group velocity**.

We have $\alpha(\theta) \to a$ as $h \to 0$. So the phase speed is different from the group velocity, but both tend to the correct speed a as $h \to 0$. Otherwise the numerical method will not converge.

1.8 Summary of Schemes for the Wave Equation $u_t + au_x = 0$

Notation: $\lambda = \frac{k}{h}$, $\theta = h\xi$.

Name	Scheme	$g(\theta)$	Stable	dissipative	$\alpha(\theta)/a$
Forward time forward space	$\frac{v_m^{n+1}-v_m^n}{k} + a\frac{v_{m+1}^n - v_m^n}{h} = 0$				
Forward time backward space	$\frac{v_m^{n+1}-v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = 0$	$1 - a\lambda + a\lambda e^{-i\theta}$	$0 \le a\lambda \le 1$	order 2, if $0 < a\lambda < 1$	
Forward-time centered space	$\frac{v_m^{n+1}-v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$	$1 - ia\lambda \sin\theta$	No	No	
Backward-time centered space	$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$				$1 - \frac{\theta^2}{6}(1 + 2a^2\lambda^2)$
Lax–Wendroff	$v_m^{n+1} = v_m^n - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n) + \frac{a^2\lambda^2}{2}(v_{m+1}^n - 2v_m^n + v_{m-1}^n)$	$1 - 2a^2\lambda^2 \sin^2\frac{\theta}{2} - ia\lambda\sin\theta$	$\lvert a\lambda\rvert \le 1$	order 4	$1 - \frac{1}{6}\theta^2(1 - a^2\lambda^2)$
Lax–Friedrichs	$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} + a\frac{m_{m+1}^n - v_{m-1}^n}{2h} = 0$	$\cos\theta - ia\lambda\sin\theta$	$\lvert a\lambda\rvert \le 1$	No	$1 + (1 - a^2\lambda^2)\frac{\theta^2}{3}$
Leap-Frog	$\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$	$-ia\lambda\sin\theta \pm \sqrt{1 - (a\lambda\sin\theta)^2}$	$\lvert a\lambda\rvert < 1$	No	$\frac{\operatorname{atan}\frac{a\lambda\sin\theta}{\sqrt{1-(a\lambda\sin\theta)^2}}}{a}$
Crank–Nicolson	$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n}{4h} = 0$	$\frac{1 - \frac{1}{2}ia\lambda\sin\theta}{1 + \frac{1}{2}ia\lambda\sin\theta}$	Always	No	$1 - \frac{\theta^2}{6}(1 + \frac{1}{2}a^2\lambda^2)$

Chapter 2

Parabolic equations

2.1 The Heat Equation

$$u_t = bu_{xx}$$

Schemes:

- Lax–Friedrichs
$$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

- Lax–Wendroff
$$\begin{aligned} u(t+k) &= u + ku_t + \frac{k^2}{2}u_{tt} \\ &= u + kbu_{xx} + \frac{k^2b^2}{2}u_{xxxx} \\ v_m^{n+1} &= v_m^n + kb\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + \frac{k^2b^2}{2}\cdot\frac{v_{m+2}^n - 4v_{m+1}^n + 6v_m^n - 4v_{m-1}^n + v_{m-2}^n}{h^4} \end{aligned}$$

- Forward in time, centered in space
$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

- Backward in time, centered in space
$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2}$$

- Leap-Frog
$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

- Du Fort–Frankel
$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - (v_m^{n+1} + v_m^{n-1}) + v_{m-1}^n}{h^2}$$

- Crank–Nicolson
$$\frac{v_m^{n+1} - v_m^n}{k} = \frac{b}{2}\left(\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} + \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}\right).$$

For parabolic equations it appears natural to have a "new" λ, which we will call μ.

Definition 10 (μ)**.**

$$\mu \equiv \frac{k}{h^2}.$$

Von Neumann analysis works the same way: $v_m^n = g^n e^{imh\xi}$.

Example 6. *Forward in time, centered in space.*

$$g(h\xi) = 1 - 4b\mu \sin^2 \tfrac{h\xi}{2}.$$

Stability requires $|g| \le 1$*, i.e.,*

$$0 \le 4b\mu \sin^2 \tfrac{h\xi}{2} \le 2 \quad \textit{for all } |h\xi| \le \pi, \quad \textit{meaning } b\mu \le \tfrac{1}{2}.$$

The scheme is dissipative of order 2 as long as $b\mu$ *is strictly less than* $\frac{1}{2}$ *(check!). For* $b\mu = \frac{1}{2}$ *we have* $g = 1 - 2\sin^2 \frac{h\xi}{2}$ *and the frequency* $\xi = \frac{\pi}{h}$ *is not damped at all:* $v_m^0 = e^{imh\xi} = e^{im\pi} = (-1)^m$ *remains unchanged by the scheme.*

Definition 11. *Let*

$$\|u(t,x)\|_x \equiv \left(\int_{-\infty}^{\infty} |u(t,x)|^2 dx \right)^{\frac{1}{2}}$$

mean the L^2 *norm of* $u(t,x)$ *with respect to* x *for a fixed* t*.*

Remark 2. *Let* u *be a solution to* $u_t = bu_{xx}$*. Then the overall energy* $E(t) \equiv \|u(t,x)\|_x^2$ decreases *with time*

$$\|u(t,x)\|_x \le \|u(s,x)\|_x, \ \textit{when } t \ge s$$

and the solution becomes smoother with time

$$\|u_x(t,x)\|_x^2 \le \frac{1}{2bt} \|u(0,x)\|^2.$$

The dissipative schemes possess the same qualities

$$\|v^{n+1}\| = \|\hat{v}^{n+1}\| = \|g(h\xi)\hat{v}^n\| \le \|\hat{v}^n\| = \|v^n\|$$

and

$$\|\delta_+ v^n\|^2 \le \frac{4\mu}{Ct} \|v^0\|^2,$$

where $\delta_+ v_m \equiv \frac{v_{m+1} - v_m}{h}$*, and* C *is a constant.*

Proof. We will show that $E'(t) \le 0$ meaning $E(t)$ is decreasing:

$$E'(t) \quad = \quad \int_{-\infty}^{\infty} 2uu_t dx = \int_{-\infty}^{\infty} 2buu_{xx} dx = 2buu_x \Big|_{-\infty}^{\infty} - 2b \int_{-\infty}^{\infty} u_x^2 dx = -2b\|u_x\|_x^2 \le 0,$$

because $u(t,x) \to 0$ as $x \to \pm\infty$ for $\|u(t,x)\|_x$ to exist.

The above implies (after integrating from 0 to t):

$$E(t) - E(0) = -2b \int_0^t \|u_x(\tau,x)\|_x^2 d\tau \quad \Rightarrow \quad E(0) \ge 2b \int_0^t \|u_x(\tau,x)\|_x^2 d\tau$$

The derivative $u_x = \frac{\partial}{\partial x} u(t,x)$ is also a solution to $u_t = bu_{xx}$ because $(u_x)_t = (u_t)_x = bu_{xxx} = b(u_x)_{xx}$, therefore $\|u_x(t,x)\|_x \le \|u_x(s,x)\|_x$ for $t \ge s$. Now we get

$$E(0) \ge 2b \int_0^t \|u_x(\tau,x)\|_x^2 d\tau \ge 2bt \|u_x(t,x)\|_x^2,$$

meaning

$$\|u_x(t,x)\|_x^2 \le \frac{1}{2bt}\|u(0,x)\|_x^2,$$

i.e., the solution get smoother and smoother as $t \to \infty$.

Now repeat the same analysis for a difference scheme that is dissipative of order (say) 2:

$$\hat{v}^{n+1}(\xi) = g(h\xi)\hat{v}^n(\xi), \quad \text{where} \quad |g(h\xi)| \le 1 - C\sin^2\tfrac{h\xi}{2},$$

which implies

$$\|\hat{v}^{n+1}(\xi)\|^2 \le \left\|\left(1 - C\sin^2\tfrac{h\xi}{2}\right)\hat{v}^n(\xi)\right\|^2$$

and (after some major reworking):

$$\|\hat{v}^{n+1}\|^2 + C\left\|\sin\tfrac{h\xi}{2}\cdot\hat{v}^n\right\|^2 \le \|\hat{v}^n\|^2$$

(these are, of course, the discrete L^2 norms). Now comes the big moment,

$$\sin(h\xi/2)\cdot\hat{v}^n(\xi) = \frac{e^{ih\xi/2} - e^{-ih\xi/2}}{2i}\cdot\hat{v}^n(\xi) = \frac{e^{-ih\xi/2}}{2i}(e^{ih\xi} - 1)\hat{v}^n(\xi).$$

Next, observe that

$$\delta_+ v_m^n \equiv \frac{1}{h}(v_{m+1}^n - v_m^n) = \frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h} e^{imh\xi}\frac{e^{ih\xi}-1}{h}\hat{v}^n(\xi)d\xi.$$

On the other side,

$$\delta_+ v_m^n = \frac{1}{\sqrt{2\pi}}\int_{-\pi/h}^{\pi/h} e^{imh\xi}\widehat{\delta_+ v^n}(\xi)d\xi,$$

so

$$\frac{e^{ih\xi}-1}{h}\hat{v}^n(\xi) = \widehat{\delta_+ v^n}(\xi),$$

and inserting we get

$$\|\sin(h\xi/2)\cdot\hat{v}^n\| = \tfrac{1}{2}\|(e^{ih\xi}-1)\hat{v}^n\| = \tfrac{1}{2}\cdot h\cdot\|\widehat{\delta_+ v^n}\|.$$

We can therefore simplify our inequality

$$\|\hat{v}^{n+1}\|^2 + C\frac{h^2}{4}\|\widehat{\delta_+ v^n}\|^2 \le \|\hat{v}^n\|^2.$$

Parseval says "hats = no hats", so

$$\|v^{n+1}\|^2 + \frac{Ck}{4\mu}\|\delta_+ v^n\|^2 \le \|v^n\|^2, \tag{2.1}$$

where, again, $\mu = k/h^2$. In particular,

$$\|v^{n+1}\| \le \|v^n\|.$$

Next, we prove that $\|\delta_+ v^{n+1}\| \le \|\delta_+ v^n\|$. The only property we used was that v_m^n was solution to

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}.$$

Now try

$$\frac{v_{m+1}^{n+1} - v_{m+1}^n}{k} = b\frac{v_{m+2}^n - 2v_{m+1}^n + v_m^n}{h^2}.$$

Subtract first from second, divide by h and get

$$\frac{\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}-\frac{v_{m+1}^n-v_m^n}{h}}{k}=b\frac{\frac{v_{m+2}^n-v_{m+1}^n}{h}-2\frac{v_{m+1}^n-v_m^n}{h}+\frac{v_m^n-v_{m-1}^n}{h}}{h^2}.$$

Or in a simpler language

$$\frac{(\delta_+v)_m^{n+1}-(\delta_+v^n)_m}{k}=b\frac{(\delta_+v)_{m+1}^n-2(\delta_+v)_m^n+(\delta_+v)_{m-1}^n}{h^2},$$

meaning $\|\delta_+v^{n+1}\|\leq\|\delta_+v^n\|$.

The inequality (2.1) works for all time steps

$$\begin{aligned}
\|v^{n+1}\|^2+\frac{Ck}{4\mu}\|\delta_+v^n\|^2 &\leq \|v^n\|^2\\
\|v^n\|^2+\frac{Ck}{4\mu}\|\delta_+v^{n-1}\|^2 &\leq \|v^{n-1}\|^2\\
&\vdots\\
\|v^1\|^2+\frac{Ck}{4\mu}\|\delta_+v^0\|^2 &\leq \|v^0\|^2,
\end{aligned}$$

which we sum up and cancel common terms to obtain

$$\begin{aligned}
\|v^{n+1}\|^2+\frac{Ck}{4\mu}\sum_{k=0}^{n}\|\delta_+v^k\|^2 &\leq \|v^0\|^2\Rightarrow\\
\|v^{n+1}\|^2+\frac{Ck(n+1)}{4\mu}\|\delta_+v^{n+1}\|^2 &\leq \|v^0\|^2\Rightarrow\\
\|\delta_+v^{n+1}\|^2 &\leq \frac{4\mu}{Ct}\|v^0\|^2.
\end{aligned}$$

This means that the numerical solution will smooth out—as long as the scheme is dissipative. □

2.2 The Du Fort–Frankel Scheme

This is an example of an *explicit* and *unconditionally stable* scheme for $u_t = bu_{xx}$.

The problem with schemes like forward time, centered space is that they are stable for $b\mu = bk/h^2 \leq \frac{1}{2}$, which puts a terrible restriction $k \leq \frac{h^2}{2b}$ on the timestep. The Du Fort–Frankel scheme,

$$v_m^{n+1} - v_m^{n-1} = 2b\mu(v_{m+1}^n - (v_m^{n+1} + v_m^{n-1}) + v_{m-1}^n),$$

is a slight modification of the *unstable* Leap–Frog scheme. We rewrite the Du Fort–Frankel scheme as

$$(1 + 2b\mu)v_m^{n+1} - (1 - 2b\mu)v_m^{n-1} = 2b\mu(v_{m+1}^n + v_{m-1}^n).$$

To study the stability, we substitute $v_m^n = g^n e^{imh\xi}$ to get

$$(1 + 2b\mu)g^2 - (1 - 2b\mu) = 2b\mu(e^{ih\xi} + e^{-ih\xi})g,$$

which implies

$$g_\pm = \frac{2b\mu\cos(h\xi) \pm \sqrt{1 - 4b^2\mu^2\sin^2(h\xi)}}{1 + 2b\mu}.$$

The scheme is not dissipative since $g_-(\pi) = -1$. To determine stability we consider two cases:

- $1 - 4b^2\mu^2 \geq 0 \Rightarrow |g_\pm| \leq \frac{2b\mu|\cos(h\xi)|+\sqrt{1}}{1+2b\mu} \leq \frac{2b\mu+1}{1+2b\mu} = 1.$
- $1 - 4b^2\mu^2 < 0 \Rightarrow |g_\pm|^2 = \frac{(2b\mu\cos(h\xi))^2+4b^2\mu^2\cos^2(h\xi)-1}{1+2b\mu} = \frac{4b^2\mu^2-1}{(1+2b\mu)^2} = \frac{2b\mu-1}{1+2b\mu} \leq 1.$

In addition, we do not want double roots on the unit circle. Double root occurs when $1 - 4b^2\mu^2 = 0$, but then $|g_\pm| \leq \frac{2b\mu|\cos(h\xi)|}{1+2b\mu} < 1$.

So we have stability for any value of μ. But how is that possible? The catch is in the consistency. In order for the scheme to be consistent we must have $k/h \to 0$, as we will now demonstrate.

Rewrite Du Fort–Frankel as

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} - b\frac{v_m^{n+1} + v_m^{n-1} - 2v_m^n}{h^2}$$

then expand in Taylor to see that it approximates

$$u_t + \frac{k^2}{6}u_{ttt} = b\big(u_{xx} + \frac{h^2}{12}u_{xxxx}\big) - b\big(\frac{k^2}{h^2}u_{tt} + \frac{k^4}{12h^2}u_{tttt}\big).$$

Now think numerically. For hyperbolic systems we could (at best) hope for $\frac{k}{h} \approx 1$. However, if we use Du Fort–Frankel with such a timestep, $k = h$, the solution will not converge to the solution of $u_t = bu_{xx}$, but instead to the solution of $bu_{tt} + u_t = bu_{xx}$ (i.e., the solution to a wave equation). This was not the purpose of the exercise. So the scheme will only converge to the solution of $u_t = bu_{xx}$ if $\frac{k}{h} \to 0$. Even so the truncation error will be dominated by $b\frac{k^2}{h^2}u_{tt}$, which is not small unless $\frac{k}{h^2}$ is constant, but then we are back where we started—with the same restrictions as the ones for forward in time centered in space.

We, of course have two explicit schemes—backward in time, centered in space (which is $O(k + h^2)$ and dissipative) and Crank–Nicolson (which is $O(k^2 + h^2)$ and not dissipative if $\frac{k}{h}$ is constant.).

2.3 The Convection-Diffusion Equation: $u_t + au_x = bu_{xx}$

If $a = 0$, we have the heat equation, and if $b = 0$, we have the wave equation. Define a new function w such that $w(t, x - at) = u(t, x)$, i.e., $w(t, x) = u(t, x + at)$. Then

$$bw_{xx} = bu_{xx} = u_t + au_x = w_t - aw_x + aw_x = w_t.$$

So u is simply the solution to the heat equation translated with speed a. The problem occurs when the viscosity coefficient b is very small compared to a. Then the obvious numerical methods trick you. We have a choice of

$$\begin{aligned}
\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} &= b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}, && (2.2)\\
\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_m^n}{h} &= b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}, &&\\
\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} &= b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}. && (2.3)
\end{aligned}$$

We expand in Taylor series to see that the orders of accuracy are

$$O(k + h^2), \quad O(k + h), \quad \text{and} \quad O(k + h),$$

respectively, which speaks strongly in favor of (2.2).

The heat equation has one nice property. The maximum at a later time is less than the maximum at an earlier time. We will copy that. Rewrite (2.2) as

$$\begin{aligned}
v_m^{n+1} &= \left(b\mu - \frac{a\lambda}{2}\right) v_{m+1}^n + (1 - 2b\mu)v_m^n + \left(b\mu + \frac{a\lambda}{2}\right) v_{m-1}^n\\
&= b\mu\left(1 - \frac{a\lambda}{2b\mu}\right) v_{m+1}^n + (1 - 2b\mu)v_m^n + b\mu\left(1 + \frac{a\lambda}{2b\mu}\right) v_{m-1}^n.
\end{aligned}$$

Let $\alpha \equiv \frac{a\lambda}{2b\mu} = \frac{ah}{2b}$. Now if all the coefficients are positive (i.e., $|\alpha| = |\frac{a\lambda}{2b\mu}| < 1$), then

$$\begin{aligned}
|v_m^{n+1}| &\le b\mu(1 - \alpha)\max_m |v_m^n| + (1 - 2b\mu)\max_m |v_m^n| + b\mu(1 + \alpha)\max_m |v_m^n|\\
&\le [b\mu(1 - \alpha) + (1 - 2b\mu) + b\mu(1 + \alpha)]\max_m |v_m^n|\\
&\le \max_m |v_m^n|.
\end{aligned}$$

The maximum is a decreasing function of time if $|\alpha| \le 1$, i.e., if

$$\frac{|a|}{b} \cdot \frac{h}{2} \le 1, \quad \text{which is the same as} \quad h \le \frac{2b}{|a|}. \tag{2.4}$$

This will, of course, be satisfied eventually as $h \to 0$, but who can wait that long? Say $a = 10, b = 10^{-2} \Rightarrow h \approx 10^{-3}$ is needed. And remember, for stability we must have $1/2 \ge b\mu = 10^{-2} \cdot k/(10^{-3})^2$. This implies $k \approx 10^{-4}/2$, which is terribly small.

Now look at (2.3) instead:

$$v_m^{n+1} - v_m^n + a\lambda(v_m^n - v_{m-1}^n) = b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n),$$

which we rewrite as

$$\begin{aligned}
v_m^{n+1} &= (b\mu + a\lambda)v_{m-1}^n + (1 - 2b\mu - a\lambda)v_m^n + b\mu v_{m+1}^n\\
&= b\mu(1 + \frac{a\lambda}{b\mu})v_{m-1}^n + (1 - 2b\mu(1 + \frac{a\lambda}{2b\mu}))v_m^n + b\mu v_{m+1}^n\\
&= b\mu(1 + 2\alpha)v_{m-1}^n + (1 - 2b\mu(1 + \alpha))v_m^n + b\mu v_{m+1}^n.
\end{aligned}$$

Say $a > 0$, so $\alpha > 0$. Now the requirement for max-norm stability becomes

$$2b\mu(1+\alpha) < 1 \quad \text{or} \quad 2b\mu + a\lambda < 1,$$

which is a lot less restrictive than (2.4). We can pick h to be 10^{-2} instead of 10^{-3}, i.e., 10 times larger. We try $k = \frac{10^{-3}}{5}$, $h = 10^{-2}$, $a = 10, b = 10^{-2}$. Then

$$2b\frac{k}{h^2} + a\frac{k}{h} = 2 \cdot 10^{-2} \cdot \frac{\frac{10^{-3}}{5}}{(10^{-2})^2} = \frac{2}{50} + \frac{1}{5} < 1.$$

So we increased the timestep by a factor of 10. But at what price?

We can rewrite (2.3) as

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = \left(b + \frac{ah}{2}\right)\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}.$$

We introduced an artificial viscosity $\frac{ah}{2} = b\alpha$. This artificial viscosity comes from our numerical method.

Therefore solving $u_t + au_x = b(1+\alpha)u_{xx}$ using (2.2) is equivalent to solving $u_t + au_x = bu_{xx}$ using (2.3).

2.4 Summary of Schemes for the Heat Equation $u_t = bu_{xx}$

Notation: $\mu \equiv \frac{k}{h^2}$, $\theta = h\xi$.

Name	Scheme	$g(h\xi)$	Stable	Dissipative
Lax–Friedrichs	$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$			
Lax–Wendroff	$v_m^{n+1} = v_m^n + kb\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + \frac{k^2b^2}{2} \cdot \frac{v_{m+2}^n - 4v_{m+1}^n + 6v_m^n - 4v_{m-1}^n + v_{m-2}^n}{h^4}$			
Forward time centered space	$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$	$1 - 4b\mu\sin^2\frac{h\xi}{2}$	$b\mu \le \frac{1}{2}$	2, if $b\mu < \frac{1}{2}$
Backward time centered space	$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2}$		Always	if $\mu \ge c$, $c > 0$
Leap-Frog	$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$			No
Du Fort–Frankel	$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - (v_m^{n+1} + v_m^{n-1}) + v_{m-1}^n}{h^2}$	$\frac{2b\mu\cos\theta \pm \sqrt{1 - 4b^2\mu^2\sin^2\theta}}{1+2b\mu}$	Always	No
Crank–Nicolson	$\frac{v_m^{n+1} - v_m^n}{k} = \frac{b}{2}\left(\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} + \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}\right)$		Always	order 2

Chapter 3

Systems of PDEs in Higher Dimensions

3.1 The Equation $u_t + Au_x = 0$

We consider $u_t + Au_x = 0$, where u is a d-vector of functions and A and B are d-by-d matrices. Initial conditions $u(0,x) = f(x)$ (all d-vectors). If A is symmetric $A = Q\Lambda Q^T$, where $Q^TQ = I$, then

$$u_t + Q\Lambda Q^T u_x = 0 \Rightarrow Q^T u_t + \Lambda Q^T u_x = 0 \Rightarrow w_t + \Lambda w_x = 0,$$

where $w = Q^T u$. The matrix Λ is diagonal, so the problem falls apart to d independent equations.

$$(w_i)_t + \lambda_i (w_i)_x = 0.$$

The solution is $w_i(t,x) = g_i(x - \lambda_i t)$, where $g(x) = Q^T f(x)$. In general

$$u(t,x) = Qw(t,x) = Q \cdot (g_1(x - \lambda_1 t), \ldots, g_d(x - \lambda_d t))^T.$$

How do we generate and analyze a numerical scheme?
Take Lax–Wendroff:

$$u(t+k) = u + ku_t + \frac{k^2}{2} u_{tt} = u - kAu_x + \frac{k^2}{2} A^2 u_{xx}.$$

Therefore

$$\begin{aligned} v_m^{n+1} &= v_m^n - kA\frac{v_{m+1}^n - v_{m-1}^n}{2h} + \frac{k^2}{2}A^2 \frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} \\ &= v_m^n - \frac{\lambda}{2}A(v_{m+1}^n - v_{m-1}^n) + \frac{\lambda^2}{2}A^2(v_{m+1}^n - 2v_m^n + v_{m-1}^n). \end{aligned}$$

Here v_m^n is a d-vector. We can use the same Fourier analysis as before. Fourier transform on a vector means Fourier transform in each component, therefore, as before

$$\hat{v}^{n+1}(\xi) = \left[I - i\lambda A \sin(h\xi) - 2\lambda^2 A^2 \sin^2 \frac{h\xi}{2} \right] \hat{v}^n(\xi) = G(h\xi)\hat{v}^n(\xi) = \ldots = (G(h\xi))^n v^0(\xi)$$

We call G the *amplification matrix*. Stability will mean

$$\|v^{n+1}\| \le (1 + Ck)\|v^n\|, \quad \text{or} \quad \|v^{n+1}\| \le \text{Const} \cdot \|v^0\|,$$

when C does not depend on h, k. Let's analyze $(G(h\xi))^n$. Assume $A = T\Lambda T^{-1}$. Then (after some very simple math)

$$G^n(h\xi) = T\left[I - i\lambda\Lambda\sin(h\xi) - 2\lambda^2\Lambda^2\sin^2\frac{h\xi}{2}\right]^n T^{-1}.$$

Therefore $|G^n(h\xi)|$ is bounded if $[I - i\lambda\Lambda\sin(h\xi) - 2\lambda^2\Lambda^2\sin^2\frac{h\xi}{2}]^n$ is. The latter is a diagonal matrix, meaning each element on the diagonal needs to be bounded. We have reduced the problem to the one-dimensional case (which implies $\lambda|\lambda_j(A)| < 1$) , therefore the stability condition is

$$\lambda \cdot \max_{1\le j\le d}|\lambda_j(A)| \le 1.$$

3.2 The Equation $u_t + Au_x = Bu$

Take Lax–Wendroff for $u_t + Au_x = 0$ and just add the lower order term

$$v_m^{n+1} = v_m^n - \frac{\lambda}{2}A(v_{m+1}^n - v_{m-1}^n) + \frac{\lambda^2}{2}A^2(v_{m+1}^n - 2v_m^n + v_{m-1}^n) + kBv_m^n$$

The amplification factor for the scheme becomes

$$\hat{v}^{n+1}(\xi) = (G + kB)\hat{v}^n(\xi),$$

so we end up with the question when $|(G + kB)^n|$ is bounded by a constant, where $|G^n(h\xi)| \le C$. We use Strang's idea:

$$\begin{aligned}(G+kB)^n &= G^n + k\left(G^{n-1}B + G^{n-2}BG + \ldots BG^{n-1}\right)\\ &\quad + k^2\left(G^{n-2}B^2 + G^{n-3}BGB + B^{n-4}BG^2B + \ldots + B^2G^{n-2}\right)\\ &\quad + \ldots\\ &\quad + k^n(G^0BG^0B\ldots G^0BG^0).\end{aligned}$$

Now estimate upward using $|G^n| \le C$ to obtain

$$\begin{aligned}|(G+kB)^n| &\le C + \binom{n}{1}kC^2|B| + \binom{n}{2}k^2C^3|B|^2 + \ldots\\ &\le C\left(1 + \binom{n}{1}kC\cdot|B| + \binom{n}{2}k^2C^2\cdot|B|^2 + \ldots\right)\\ &\le C(1+k|B|C)^n \le Ce^{nk|B|C} \le Ce^{T|B|C}\end{aligned}$$

for all $nk \le T$.

Therefore the condition for stability is once again $|G^n| \le C$.

3.3 The Equation $u_t + Au_x + Bu_y = 0$

The equation $u_t + Au_x + Bu_y = 0$, where A and B are d-by-d matrices, is a baby problem for an underlying three dimensional problem, which linearized and reduced leads to the above equation, namely Euler's equation in fluid mechanics. For a numerical scheme pick

$$\frac{v_{ml}^{n+1} - v_{ml}^{n-1}}{2k} + A\frac{v_{m+1,l}^n - v_{m-1,l}^n}{2h} + B\frac{v_{m,l+1}^n - v_{m,l-1}^n}{2h} = 0.$$

Using Fourier transforms in two dimensions we get

$$\hat{v}^{n+1}(\xi_1,\xi_2) - \hat{v}^n(\xi_1,\xi_2) = 2i\lambda[A\sin(h\xi_1) + B\sin(h\xi_2)]\hat{v}^n(\xi_1,\xi_2),$$

which is the same as

$$\begin{bmatrix} \hat{v}^{n+1} \\ \hat{v}^n \end{bmatrix} = \begin{bmatrix} 2i\lambda(A\sin\theta_1 + B\sin\theta_2) & I \\ I & 0 \end{bmatrix} \begin{bmatrix} \hat{v}^n \\ \hat{v}^{n-1} \end{bmatrix} = G(\theta_1, \theta_2) \begin{bmatrix} \hat{v}^n \\ \hat{v}^{n-1} \end{bmatrix}.$$

So we have to study $|G^n(\theta_1, \theta_2)|$ and see when it is bounded. Depending on the assumptions one makes this can be easy or hard. If we make the simplifying assumptions that A and B are simultaneously diagonalizable, $A = P\Lambda' P^{-1}$, $B = P\Lambda'' P^{-1}$. Then

$$A\sin(h\xi_1) + B\sin(h\xi_2) = P\,\mathrm{diag}(\lambda_i' \sin(h\xi_1) + \lambda_i'' \sin(h\xi_2))P^{-1}$$

and we can write

$$G = \begin{bmatrix} P & \\ & P \end{bmatrix} \begin{bmatrix} 2i\lambda(\Lambda'\sin\theta_1 + \Lambda''\sin\theta_2) & I \\ I & 0 \end{bmatrix} \begin{bmatrix} P^{-1} & \\ & P^{-1} \end{bmatrix},$$

so the whole analysis breaks down to the analysis of the scalar case. Without repeating the previous analysis we want

$$|\lambda(\lambda_i' \sin\theta_1 + \lambda_i'' \sin\theta_2)| < 1 - \epsilon$$

for all $\theta_1 = h\xi_1, \theta_2 = h\xi_2$ and $i = 1, 2, \ldots, d$. Our condition for stability is then

$$\lambda \max_i(|\lambda_i'| + |\lambda_i''|) < 1 - \epsilon.$$

In general if the assumption is that A and B are only symmetric, then the stability criterion is the same, but the proof is a bit more complicated.

3.4 The equation $u_t = b_1 u_{xx} + b_2 u_{yy}$

Introduce a notation

$$A_1 u = b_1 \frac{\partial^2}{\partial x^2} u, \qquad A_2 u = b_2 \frac{\partial^2}{\partial y^2} u.$$

The discrete versions of these operators are

$$A_{1h} v^n_{ml} = b_1 \frac{v^n_{m+1,l} - 2v^n_{ml} + v^n_{m-1,l}}{h^2} \quad \text{and} \quad A_{2h} v^n_{ml} = b_2 \frac{v^n_{m,l+1} - 2v^n_{ml} + v^n_{m,l-1}}{h^2},$$

where $u(t,x,y) = u(nk, mh, lh) \sim v^n_{ml}$ (assume $\Delta x = \Delta y = h$ and $\Delta t = k$). Next we attack $u_t = b_1 u_{xx} + b_2 u_{yy} = A_1 u + A_2 u$ using the Crank–Nicolson idea

$$\frac{v^{n+1}_{ml} - v^n_{ml}}{k} = (A_{1h} + A_{2h}) \frac{v^{n+1}_{ml} + v^n_{ml}}{2},$$

which is equivalent to

$$\left(I - \frac{k}{2}(A_{1h} + A_{2h})\right) v^{n+1} = \left(I + \frac{k}{2}(A_{1h} + A_{2h})\right) v^n,$$

and also equivalent to

$$\left(I - \frac{k}{2}A_{1h}\right)\left(I - \frac{k}{2}A_{2h}\right) v^{n+1} = \left(I + \frac{k}{2}A_{1h}\right)\left(I + \frac{k}{2}A_{2h}\right) v^n + A_{1h}A_{2h} \underbrace{\frac{k^2}{4}(v^{n+1} - v^n)}_{(*)}$$

The expression (*) is $O(k^3)$ (since $v^{n+1} - v^n = O(k)$), so dropping it does not affect the order of the accuracy of the scheme.

Our scheme thus becomes

$$\left(I - \frac{k}{2}A_{1h}\right)\left(I - \frac{k}{2}A_{2h}\right) v^{n+1} = \left(I + \frac{k}{2}A_{1h}\right)\left(I + \frac{k}{2}A_{2h}\right) v^n.$$

3.5 ADI methods

Solve the above as

$$\begin{aligned}\left(I-\frac{k}{2}A_{1h}\right)v^{n+1/2} &= \left(I+\frac{k}{2}A_{2h}\right)v^{n}\\ \left(I-\frac{k}{2}A_{2h}\right)v^{n+1} &= \left(I+\frac{k}{2}A_{1h}\right)v^{n+1/2}\end{aligned}$$

This is the **Peaceman–Rachford** Algorithm, which is an *ADI method—alternating direction implicit method*. Meaning that the two-dimensional problem has been reduced to two one-dimensional implicit problems by factoring the scheme.

Let's now perform stability analysis of this scheme.

$$\begin{aligned}\left(I-\frac{k}{2}A_{2h}\right)v^{n+1} &= \left(I+\frac{k}{2}A_{1h}\right)v^{n+1/2}\\ \left(1-b_2\frac{k}{2h^2}(e^{ih\xi_2}-2+e^{-ih\xi_2})\right)\hat{v}^{n+1} &= \left(1+b_1\frac{k}{2h^2}(e^{ih\xi_1}-2+e^{-ih\xi_1})\right)\hat{v}^{n+1/2}\\ \left(1+2\mu b_2\sin^2(h\xi_2/2)\right)\hat{v}^{n+1} &= \left(1-2\mu b_1\sin^2(h\xi_1/2)\right)\hat{v}^{n+1/2}\\ \hat{v}^{n+1} &= \frac{1-2\mu b_1\sin^2(h\xi_1/2)}{1+2\mu b_2\sin^2(h\xi_2/2)}\hat{v}^{n+1/2}.\\ \text{Similarly}\qquad \hat{v}^{n+1/2} &= \frac{1-2\mu b_2\sin^2(h\xi_2/2)}{1+2\mu b_1\sin^2(h\xi_1/2)}\hat{v}^{n}\\ \text{Meaning}\qquad \hat{v}^{n+1} &= \frac{1-2\mu b_1\sin^2(h\xi_1/2)}{1+2\mu b_1\sin^2(h\xi_1/2)}\cdot\frac{1-2\mu b_2\sin^2(h\xi_2/2)}{1+2\mu b_2\sin^2(h\xi_2/2)}\cdot\hat{v}^{n}.\end{aligned}$$

Since $\left|\frac{1-x}{1+x}\right|\le 1$ for any $x\ge 0$, we conclude $|\hat{v}^{n+1}|\le|\hat{v}^{n}|$.

Now using Parseval

$$\begin{aligned}\sum_{m,l}|v_{ml}^{n+1}|^2\cdot h^2 &= \int_{-\pi/h}^{\pi/h}\int_{-\pi/h}^{\pi/h}|\hat{v}^{n+1}(\xi_1,\xi_2)|^2d\xi_1d\xi_2\\ &\le \int_{-\pi/h}^{\pi/h}\int_{-\pi/h}^{\pi/h}|\hat{v}^{n}(\xi_1,\xi_2)|^2d\xi_1d\xi_2\\ &= \sum_{m,l}|v_{ml}^{n}|^2\cdot h^2.\end{aligned}$$

Therefore we have stability for all values of m and l, and the order of accuracy of the scheme is $O(k^2+h^2)$. Thus we can take $k=h$ and the scheme is efficient and accurate.

The **Douglas–Rachford** method starts with the backward-time, central-space scheme for $u_t=A_1u+A_2u$

$$(I-kA_{1h}-kA_{2h})v_{ml}^{n+1}=v_{ml}^{n},$$

to obtain (after dropping an $O(k^3)$ term)

$$(I-kA_{1h})(I-kA_{2h})v_{ml}^{n+1}=(I+k^2A_{1h}A_{2h})v_{ml}^{n}.$$

The method is

$$\begin{aligned}\left(I-kA_{1h}\right)v^{n+1/2} &= \left(I+kA_{2h}\right)v^{n}\\ \left(I-kA_{2h}\right)v^{n+1} &= v^{n+1/2}-kA_{2h}v^{n}.\end{aligned}$$

The **Mitchell–Fairweather** method is second-order accurate in time and fourth-order accurate in space. Recall the operator δ^2:

$$\delta^2 f(x) \equiv \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

and

$$\delta_x^2 v_{ml}^n = \frac{v_{m+1,l}^n - 2v_{ml}^n + v_{m-1,l}^n}{h^2} \quad \text{and} \quad \delta_y^2 v_{ml}^n = \frac{v_{m,l+1}^n - 2v_{ml}^n + v_{m,l-1}^n}{h^2}.$$

Then

$$\delta^2 f = f'' + \frac{h^2}{12} f'''' + O(h^4) = f'' + O(h^2),$$

so $f'' = \delta^2 f + O(h^2)$. Then

$$\begin{aligned}
\delta^2 f &= f'' + \frac{h^2}{12} f'''' + O(h^4) \\
&= f'' + \frac{h^2}{12}(f'')'' + O(h^4) \\
&= f'' + \frac{h^2}{12}(\delta^2 f'' + O(h^2)) + O(h^4) \\
&= f'' + \frac{h^2}{12}\delta^2 f'' + O(h^4) \\
&= \delta^2 f = \left(1 + \frac{h^2}{12}\delta^2\right) \frac{d^2}{dx^2} f + O(h^4). \qquad (3.1)
\end{aligned}$$

Now start with Peaceman–Rachford idea for $u_t = A_1 u + A_2 u$:

$$\left(I - \frac{k}{2}A_1\right)\left(I - \frac{k}{2}A_2\right) u^{n+1} = \left(I + \frac{k}{2}A_1\right)\left(I + \frac{k}{2}A_2\right) u^n + O(k^3)$$

Multiply both sides by

$$\left(1 + \frac{h^2}{12}\delta_x^2\right)\left(1 + \frac{h^2}{12}\delta_y^2\right)$$

and replace

$$\left(1 + \frac{h^2}{12}\delta_x^2\right) \frac{\partial^2}{\partial x^2}$$

by $\delta_x^2 + O(h^4)$ (see (3.1)). Similar changes are made for the derivatives with respect to y. The result is

$$\begin{aligned}
&\left(1 + \frac{h^2}{12}\delta_x^2 - \frac{k}{2}b_1\delta_x^2\right)\left(1 + \frac{h^2}{12}\delta_y^2 - \frac{k}{2}b_2\delta_y^2\right) u^{n+1} \\
&\qquad = \left(1 + \frac{h^2}{12}\delta_x^2 + \frac{k}{2}b_1\delta_x^2\right)\left(1 + \frac{h^2}{12}\delta_y^2 + \frac{k}{2}b_2\delta_y^2\right) u^n + O(k^3) + O(kh^4).
\end{aligned}$$

We obtain the Mitchell–Fairweather scheme:

$$\begin{aligned}
\left[1 - \frac{1}{2}\left(b_1\mu_1 - \frac{1}{6}\right)h_1^2\delta_x^2\right] v^{n+1/2} &= \left[1 + \frac{1}{2}\left(b_2\mu_2 + \frac{1}{6}\right)h_2^2\delta_y^2\right] v^n \\
\left[1 - \frac{1}{2}\left(b_2\mu_2 - \frac{1}{6}\right)h_2^2\delta_y^2\right] v^{n+1} &= \left[1 + \frac{1}{2}\left(b_1\mu_1 + \frac{1}{6}\right)h_1^2\delta_x^2\right] v^{n+1/2}.
\end{aligned}$$

3.6 Boundary conditions for ADI methods

One can obtain the boundary conditions for the intermediate step $v^{n+1/2}$ by solving for it using the boundary conditions at steps n and $n+1$.

For example by subtracting the equations of the Peaceman–Rachford method

$$\begin{aligned}\left(I - \frac{k}{2}A_1\right) u^{n+1/2} &= \left(I + \frac{k}{2}A_2\right) u^n \\ \left(I - \frac{k}{2}A_2\right) u^{n+1} &= \left(I + \frac{k}{2}A_1\right) u^{n+1/2}\end{aligned}$$

(note that we use u's and not v's and keep everything is operator form for the moment) we get

$$u^{n+1/2} = \frac{1}{2}\left(I + \frac{k}{2}A_2\right) u^n + \frac{1}{2}\left(I - \frac{k}{2}A_2\right) u^{n+1}.$$

Now if the indices i and j in v_{ij}^n range from 1 to m, the desired boundary conditions at $v_{1i}^{n+1/2}$ and $v_{mi}^{n+1/2}$ for $i = 2, 3, \ldots, m-1$ are computed as

$$\begin{aligned}v_{1i}^{n+1/2} &= \frac{1}{2}\left(1 + \frac{k}{2}A_2\right) v_{1i}^n + \frac{1}{2}\left(1 - \frac{k}{2}A_2\right) v_{1i}^{n+1} \\ &= \frac{b_2\mu v_{1,i-1}^n + 2(1-b_2\mu)v_{1i}^n + b_2\mu v_{1,i+1}^n}{4} + \frac{-b_2\mu v_{1,i-1}^{n+1} + 2(1+b_2\mu)v_{1i}^{n+1} - b_2\mu v_{1,i+1}^{n+1}}{4}.\end{aligned}$$

And similarly for $v_{mi}^{n+1/2}$ for $i = 2, 3, \ldots, m-1$.

The boundary condition

$$v_{ij}^{n+1/2} = u(t_{n+1/2}, x_i, y_j)$$

is only first order accurate and if used with the Peaceman–Rachford method (or other similar second order accuracy) will result in the overall accuracy being only first order.

Boundary conditions for the Mitchell–Fairweather scheme are obtained as follows. First we eliminate $\delta_x^2 v^{n+1/2}$ terms by multiplying the first equation by $1 + \frac{1}{2}(b_1\mu_1 + \frac{1}{6})h_1^2\delta_x^2$ and the second by $1 - \frac{1}{2}(b_1\mu_1 - \frac{1}{6})h_1^2\delta_x^2$ to obtain:

$$v_{1i}^{n+1/2} = \frac{\left(b_1\mu_1 + \frac{1}{6}\right)\left(1 + \frac{1}{2}(b_2\mu_2 + \frac{1}{6})h_2^2\delta_y^2\right) v_{1i}^n + \left(b_1\mu_1 - \frac{1}{6}\right)\left(1 - \frac{1}{2}(b_2\mu_2 - \frac{1}{6})h_2^2\delta_y^2\right) v_{1i}^{n+1}}{2b_1\mu_1}$$

for $i = 2, 3, \ldots, m-1$, and similarly for the other boundary.

Chapter 4

Elliptic Equations

4.1 Steady-State Heat Equation

The **steady-state heat equation** is

$$u_{xx} + u_{yy} = f(x, y). \tag{4.1}$$

We solve it numerically by introducing a rectangular grid on a finite domain Ω (which need not be rectangular).

The numerical scheme is then

$$\frac{v_{m+1,l} - 2v_{ml} + v_{m-1,l}}{h^2} + \frac{v_{m,l+1} - 2v_{ml} + v_{m,l-1}}{h^2} = f_{ml}, \tag{4.2}$$

with boundary conditions v_{ml} specified on the boundary of Ω. We write Ω_h for the set of grid points in Ω, $\partial\Omega$ for the boundary of Ω and $\partial\Omega_h$ for the set of grid points on the boundary of Ω.

Existence and uniqueness of the solution of (4.2)

Theorem 5. *The equation* (4.2) *has a unique solution.*

Proof. Assume there is another solution to (4.2); call it w_{ml}, and let $e_{ml} = v_{ml} - w_{ml}$. Subtract (4.2) for v and w to obtain

$$\frac{e_{m+1,l} - 2e_{ml} + e_{m-1,l}}{h^2} + \frac{e_{m,l+1} - 2e_{ml} + e_{m,l-1}}{h^2} = 0,$$

with $e_{ml} = 0$ on the boundary. Assume now that e_{ml} attains its maximum at an inner point (m, l) on Ω. Then

$$e_{ml} = \frac{e_{m+1,l} + e_{m-1,l} + e_{m,l+1} + e_{m,l-1}}{4}.$$

But $e_{ml} \geq e_{m\pm1,l\pm1}$, so equality is only possible if $e_{ml} = e_{m\pm1,l\pm1}$. Continuing that way we observe that the maximum must also occur on the boundary, where $e_{ml} = 0$. Thus the maximum of e_{ml} is zero and occurs on the boundary. Similarly for the minimum. Thus $e_{ml} \equiv 0$ for all m, l, which implies $v_{ml} = w_{ml}$ for all m, l.

The equations (4.2) represent a linear system for v_{ml}, which we can write as $Av = f$. Since $Ae = 0$ has the unique solution $e = 0$, we conclude that $\det A \neq 0$ and thus $Av = f$ has an unique solution. □

Convergence

Next we study convergence, namely does $v_{ml} \to u(x_m, y_l)$ as $h \to 0$? Let

$$\Delta_h v \equiv \frac{v_{m+1,l} - 2v_{ml} + v_{m-1,l}}{h^2} + \frac{v_{m,l+1} - 2v_{ml} + v_{m,l-1}}{h^2}.$$

To prove convergence we need two results. The first is analogous to the argument above:

$$\Delta_h v \geq 0 \Rightarrow \max_{\Omega_h} v_{ml} = \max_{\partial\Omega_h} v_{ml}$$

and is called **The Discrete Maximum Principle**. Second, we establish the inequality

$$\|v\|_{\infty,\Omega_h} \leq \frac{1}{8}\|\Delta_h v\|_{\infty,\Omega_h},$$

where Ω is the unit square, $v_{ml} = 0$ on $\partial\Omega_h$ and $\|v\|_{\infty,\Omega_h} \equiv \max_{(m,l)\in\Omega_h} |v_{ml}|$. The constant $\frac{1}{8}$ is connected with the shape of the domain. Start with the obvious

$$-\|f\|_{\infty,\Omega_h} \leq f_{ml} \leq \|f\|_{\infty,\Omega_h}. \tag{4.3}$$

Define

$$w_{ml} = \frac{1}{4}\left[\left(x_m - \tfrac{1}{2}\right)^2 + \left(y_l - \tfrac{1}{2}\right)^2\right].$$

Check that (a) $\Delta_h w_{ml} = 1$, and (b) $w_{ml} \leq \frac{1}{8}$ on $\partial\Omega_h$ (done in class).
Rewrite (4.3) as

$$-\|f\|_{\infty,\Omega_h}\Delta_h w_{ml} \leq f_{ml} \leq \|f\|_{\infty,\Omega_h}\Delta_h w_{ml}.$$

Therefore

$$\begin{aligned} \Delta_h(v + \|f\|_{\infty,\Omega_h} w) &\geq 0 & (4.4)\\ \Delta_h(\|f\|_{\infty,\Omega_h} w - v) &\geq 0. & (4.5) \end{aligned}$$

Now from the discrete maximum principle we have

$$\max_{\Omega_h}(v_{ml} + \|f\|_{\infty,\Omega_h} w_{ml}) \leq \max_{\partial\Omega_h}(v_{ml} + \|f\|_{\infty,\Omega_h} w_{ml}).$$

Then

$$\begin{aligned} v_{ml} &\leq v_{ml} + \|f\|_{\infty,\Omega_h} w_{ml} && \text{since } w \geq 0\\ &\leq \max_{\Omega_h}(v_{ml} + \|f\|_{\infty,\Omega_h} w_{ml})\\ &\leq \max_{\partial\Omega_h}(v_{ml} + \|f\|_{\infty,\Omega_h} w_{ml})\\ &\leq \frac{1}{8}\|f\|_{\infty,\Omega_h} && \text{since } v_{ml} = 0 \text{ on } \partial\Omega_h \text{ and } w_{ml} \leq \frac{1}{8} \text{ on } \partial\Omega_h. \end{aligned}$$

From (4.5) (in class) we get $-\frac{1}{8}\|f\|_{\infty,\Omega_h} \leq v_{ml}$, therefore $|v_{ml}| \leq \frac{1}{8}\|f\|_{\infty,\Omega_h}$ and we have

$$\max_{\Omega_h} |v_{ml}| \leq \frac{1}{8}\|f\|_{\infty,\Omega_h} = \frac{1}{8}\|\Delta_h v\|_{\infty,\Omega_h}.$$

Talyor series implies

$$\Delta_h u_{ml} = (\Delta u)_{x_m,y_l} + \frac{h^2}{12}(\partial_x^4 u(x_m + \theta_{ml}h, y_l) + \partial_y^4 u(x_m, y_l + \theta'_{ml}h) = f_{ml} + \underbrace{\frac{h^2}{12}(\partial_x^4 u + \partial_y^4 u)}_{\text{truncation error}}.$$

As usual define e_{ml} =computed $-$ exact $= v_{ml} - u_{ml}$, then

$$\begin{aligned} \Delta_h e_{ml} &= \Delta_h v_{ml} - \Delta_h u_{ml}\\ &= f_{ml} - \left[f_{ml} + \frac{h^2}{12}(\partial_x^4 u + \partial_y^4 u)\right]\\ &= -\frac{h^2}{12}(\partial_x^4 u + \partial_y^4 u). \end{aligned}$$

Note that $e_{ml} = 0$ on $\partial\Omega_h$, thus

$$\begin{aligned}\max_{\Omega_h}|e_{ml}| &\leq \frac{1}{8}\max_{\Omega_h}|\Delta_h e_{ml}| \\ &\leq \frac{h^2}{8\cdot 12}\cdot\max_{\Omega_h}|\partial_x^4 u + \partial_y^4 u| \\ &\leq \frac{h^2}{48}\cdot\max_{\Omega}(|\partial_x^4 u(x,y)|, |\partial_y^4 u(x,y)|).\end{aligned}$$

So the error goes to 0 as $h \to 0$. Remember that the constant $\frac{1}{8}$ was married to the unit square.

The continuous case

We will now prove that the equation $u_{xx} + u_{yy} = 0$ with $u(x,y) = 0$ on $\partial\Omega$ has a unique solution $u \equiv 0$. We start with the theorem of Gauss

$$\int_\Omega u_{x_i} dx = \int_{\partial\Omega} u\nu_i ds,$$

where ν_i is the normal in the ith direction. Since $uu_{xx} + uu_{yy} = 0$ we have $(uu_x)_x + (uu_y)_y - u_x^2 - u_y^2 = 0$, which we integrate over Ω using Gauss' theorem to obtain

$$\int_{\partial\Omega}(uu_x\nu_1 + uu_y\nu_2) - \int_\Omega (u_x^2 + u_y^2) = 0.$$

The first integral is 0, since $u = 0$ on $\partial\Omega$, therefore $u_x = u_y = 0$ on Ω, which along with $u = 0$ on $\partial\Omega$ implies $u \equiv 0$ on Ω.

4.2 Numerical methods for $u_{xx} + u_{yy} = f$

We have

$$v_{ml} - \frac{1}{4}(v_{m-1,l} + v_{m+1,l} + v_{m,l-1} + v_{m,l+1}) = -\frac{h^2}{4} f_{ml}.$$

Where $0 \le m, l \le N$. The above represents a linear system with $(N-1)^2$ unknowns v_{ml}, $1 \le m, l \le N-1$. Write it as $Av = b$. Solving $Av = b$ directly using Gaussian elimination would cost $O(((N-1)^2)^3) = O(N^6)$, which is prohibitive. We'd better use the banded structure of the system. For example when $N = 4$ with zero boundary conditions we have

$$\left[\begin{array}{ccc|ccc|ccc} 1 & -1/4 & & -1/4 & & & & & \\ -1/4 & 1 & -1/4 & & -1/4 & & & & \\ & -1/4 & 1 & & & -1/4 & & & \\ \hline -1/4 & & & 1 & -1/4 & & -1/4 & & \\ & -1/4 & & -1/4 & 1 & -1/4 & & -1/4 & \\ & & -1/4 & & -1/4 & 1 & & & -1/4 \\ \hline & & & -1/4 & & & 1 & -1/4 & \\ & & & & -1/4 & & -1/4 & 1 & -1/4 \\ & & & & & -1/4 & & -1/4 & 1 \end{array}\right] \cdot \left[\begin{array}{c} v_{11} \\ v_{12} \\ v_{13} \\ \hline v_{21} \\ v_{22} \\ v_{23} \\ \hline v_{31} \\ v_{32} \\ v_{33} \end{array}\right] = -\frac{h^2}{4} \cdot \left[\begin{array}{c} f_{11} \\ f_{12} \\ f_{13} \\ \hline f_{21} \\ f_{22} \\ f_{23} \\ \hline f_{31} \\ f_{32} \\ f_{33} \end{array}\right] \tag{4.6}$$

or

$$Av = b.$$

The idea in Jacobi, Gauss–Seidel, and SOR is to split $A = B + C$ and solve $(B + C)v = b$ iteratively as $v^{n+1} = B^{-1}(b - Cv^n)$, where n is the iteration, not the time. The error $e = v^n - v$ (= computed − exact) satisfies

$$e^{n+1} = B^{-1} C e^n$$

or $e^{n+1} = Fe^n$ for short.

Lemma 1. *Say $e^{n+1} = Fe^n$ for $n = 1, 2 \ldots$. Then $e^n \to 0$ if and only if $\rho(F) < 1$, where*

$$\rho(F) = \max_j |\lambda_j(F)|$$

is the spectral radius *of F.*

Proof. If $\rho(F) \ge 1$ and (say) $|\lambda_1| = \rho(F)$, then by picking e^1 as the eigenvector of F corresponding to λ_1 we have $e^{n+1} = \lambda_1^n e^1$ which will never converge to 0.

Now assume $\rho(F) < 1$ and let $F = T^{-1}JT$ be the eigenvalue decomposition of F, where

$$J = \begin{bmatrix} \lambda_1 & c_1 & & \\ & \lambda_2 & c_2 & \\ & & \ddots & \ddots \\ & & & \lambda_k \end{bmatrix}$$

and $c_i = 0$ or 1 depending on whether there is a Jordan block. The choice of ones as superdiagonal elements of Jordan blocks is is a matter of convention, we can put any positive number there, since if $S = \mathrm{diag}(\epsilon, \ldots, \epsilon^k)$,

then

$$\begin{aligned}
\tilde{J} &= S^{-1}JS \\
&= \begin{bmatrix} \epsilon^{-1} & & & \\ & \epsilon^{-2} & & \\ & & \ddots & \\ & & & \epsilon^{-k} \end{bmatrix} \begin{bmatrix} \lambda_1 & c_1 & & \\ & \lambda_2 & c_2 & \\ & & \ddots & \ddots \\ & & & \lambda_k \end{bmatrix} \begin{bmatrix} \epsilon^{1} & & & \\ & \epsilon^{2} & & \\ & & \ddots & \\ & & & \epsilon^{k} \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 & \epsilon c_1 & & \\ & \lambda_2 & \epsilon c_2 & \\ & & \ddots & \ddots \\ & & & \lambda_k \end{bmatrix}.
\end{aligned}$$

Selecting $\epsilon = |1 - \rho(F)|/2$ guarantees that

$$\max_j(|\lambda_j| + \epsilon|c_j|) < 1 - \epsilon < 1,$$

i.e., $\|\tilde{J}\|_\infty < 1 - \epsilon$. Now $F = TS\tilde{J}(TS)^{-1}$ and $F^n = TS\tilde{J}^n(TS)^{-1}$, which implies

$$\begin{aligned}
\|e^{n+1}\|_\infty &= \|TS\tilde{J}^n(TS)^{-1}e^1\|_\infty \\
&\le \|TS\|_\infty \cdot \|\tilde{J}\|_\infty^n \cdot \|(TS)^{-1}\|_\infty \cdot \|e^1\|_\infty \\
&\le \|TS\|_\infty \cdot (1-\epsilon)^n \cdot \|(TS)^{-1}\|_\infty \cdot \|e^1\|_\infty \to 0 \ \text{ as } n \to \infty.
\end{aligned}$$

□

4.3 Jacobi, Gauss–Seidel, and SOR(ω)

All start with an initial guess and iterate

$$\begin{aligned}
\text{Jacobi} \quad v_{ml}^{n+1} &= \frac{1}{4}(v_{m+1,l}^n + v_{m-1,l}^n + v_{m,l+1}^n + v_{m,l-1}^n) - \frac{h^2}{4}f_{ml}; \\
\text{Gauss–Seidel} \quad v_{ml}^{n+1} &= \frac{1}{4}(v_{m+1,l}^n + v_{m-1,l}^{n+1} + v_{m,l+1}^n + v_{m,l-1}^{n+1}) - \frac{h^2}{4}f_{ml}; \\
\text{SOR}(\omega) \quad v_{ml}^{n+1} &= (1-\omega)v_{ml}^n + \frac{\omega}{4}(v_{m+1,l}^n + v_{m-1,l}^{n+1} + v_{m,l+1}^n + v_{m,l-1}^{n+1}) - \frac{\omega h^2}{4}f_{ml}.
\end{aligned}$$

In matrix form we write the system (4.6) $Av = b$ as

$$(I - L - U)v = b,$$

where L (U) is strictly lower (upper) triangular. Then

$$\begin{aligned}
\text{Jacobi} \quad v^{n+1} &= (L+U)v^n + b \\
\text{Gauss–Seidel} \quad v^{n+1} &= (I-L)^{-1}(Uv^n + b) \\
\text{SOR}(\omega) \quad v^{n+1} &= \left(\tfrac{1}{\omega}I - L\right)^{-1}\left((\tfrac{1}{\omega}I - I + U)v^n + b\right).
\end{aligned}$$

The convergence of each method will be determined by looking at

$$\begin{aligned}
&\max_j |\lambda_j(L+U)| < 1 \\
&\max_j |\lambda_j((I-L)^{-1}U)| < 1 \\
&\max_j \left|\lambda_j\left((\tfrac{1}{\omega}I - L)^{-1}(\tfrac{1}{\omega} - I + U)\right)\right| < 1
\end{aligned}$$

Convergence of Jacobi's method

We must find the eigenvalues of $L+U$, i.e., solve

$$v_{m+1,l} + v_{m-1,l} + v_{m,l+1} + v_{m,l-1} = 4\lambda v_{ml}.$$

A direct verification (exercise in manipulating trigonometric functions) confirms that

$$v_{ml} = \sin\frac{am\pi}{N} \cdot \sin\frac{bl\pi}{N}$$

satisfy the above equation for any $a, b = 1, 2, \ldots, N-1$. Thus we have found the $(N-1)^2$ orthogonal eigenvectors. The eigenvalues are

$$\lambda^{ab} = \frac{1}{2}\left(\cos\frac{a\pi}{N} + \cos\frac{b\pi}{N}\right).$$

The spectral radius is now

$$\rho(L+U) = \cos\frac{\pi}{N} \approx 1 - \frac{1}{2}\left(\frac{\pi}{N}\right)^2.$$

Convergence of Gauss–Seidel method

We must find the eigenvalues of $(I-L)^{-1}U$, i.e., solve

$$v_{m+1,l} + \lambda v_{m-1,l} + v_{m,l+1} + \lambda v_{m,l-1} = 4\lambda v_{ml}.$$

The trick is to now set $v_{ml} = \lambda^{(m+l)/2} w_{ml}$ (we may lose a zero eigenvalue this way, but a zero eigenvalue will not hinder convergence; it will turn out later there were no zero eigenvalues). Then

$$w_{m+1,l} + w_{m-1,l} + w_{m,l+1} + w_{m,l-1} = 4\sqrt{\lambda} v_{ml}.$$

We already solved this problem for Jacobi. Therefore

$$\rho((I-L)^{-1}U) = \cos^2\frac{\pi}{N} \approx 1 - \frac{\pi^2}{N^2}.$$

Gauss–Seidel converges about twice as fast as Jacobi. To compare convergence look at the number of steps it will take to decrease the error by a factor of e (think, e.g., $e = 10^2$ or $e = 10^4$, etc.). For Jacobi, $(1 - \frac{\pi^2}{2N^2})^j = e^{-1}$ implies $j \log(1 - \frac{\pi^2}{2N^2}) = \log e^{-1}$. Since $\log(1-\rho) \approx -\rho$ for small ρ, we have $j \approx \frac{2N^2}{\pi^2} \cdot \log e$, whereas for Gauss–Seidel $(1 - \frac{\pi^2}{N^2})^g = e^{-1}$ implies $g \approx \frac{N^2}{\pi^2} \cdot \log e$, i.e., Gauss–Seidel converges twice as fast as Jacobi.

Convergence of SOR(ω)

Using the same substitution as in Gauss–Seidel we obtain

$$w_{m+1,l} + w_{m-1,l} + w_{m,l+1} + w_{m,l-1} = 4\frac{\lambda + \omega - 1}{\omega\sqrt{\lambda}} w_{ml},$$

i.e., $\mu = \frac{\lambda+\omega-1}{\sqrt{\lambda}\omega}$ is an eigenvalue of $L+U$. (Do not confuse ω with w_{ml}). Selecting ω to minimize $\rho\left((\frac{1}{\omega}I - L)^{-1}(\frac{1}{\omega}I - I + U)\right)$ we obtain

$$\begin{aligned} \omega_{opt} &= \frac{2}{1+\sqrt{1-\mu^2}} = \frac{2}{1+\sin\frac{\pi}{N}} \\ \rho\left((\tfrac{1}{\omega_{opt}}I - L)^{-1}(\tfrac{1}{\omega_{opt}}I - I + U)\right) &= \frac{\cos^2\frac{\pi}{N}}{(1+\sin\frac{\pi}{N})^2} \approx 1 - \frac{2\pi}{N}, \end{aligned}$$

i.e., SOR(ω) is N times faster than Gauss–Seidel.

Index

L^2 norm, 6

ADI methods, 31
    boundary conditions, 33
amplification factor, 7

consistency, 5
convergence, 6

dispersion, 13
dissipation, 11
Douglas–Rachford, 31

Fourier Analysis, 6

Gauss–Seidel method, 39
    convergence, 40
group velocity, 15

Jacobi's method, 39
    convergence, 40

Mitchell–Fairweather, 32

Parseval identity, 6
phase speed, 14

scheme
    heat equation
        Crank–Nicolson, 25
        Du Fort–Frankel, 23, 25
        Lax–Friedrichs, 25
        Lax–Wendroff, 25
        Leap-frog, 25
    wave equation
        forward time centered space, 8
        Lax-Wendroff, 8
        Leap-frog, 9
SOR(ω), 39
    convergence, 40
stability, 6

Von Neumann analysis, 7

wave packet, 15
well-posedness, 6