

# Mat1062: Introductory Numerical Methods for PDE

Mary Pugh

January 22, 2009

## 1 Ownership

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

## 2 Stability for PDEs

### 2.1 Diffusion Equation with Dirichlet Boundary Conditions

Let's return to the IBVP

$$\begin{cases} u_t = Du_{xx} & x \in (0, L), t \in (0, \infty) \\ u(x, 0) = u_0(x) & x \in [0, L] \\ u(0, t) = u(L, t) = 0 & t > 0 \end{cases} \quad (1)$$

As in the January 13 notes, we discretize by fixing  $N$ , setting  $h = L/N$ , introducing meshpoints  $x_j = jh$ , and forming the  $N - 1$  by  $N - 1$  matrix  $M_1$

$$M_1 = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}$$

The PDE is then approximated by  $N - 1$  ODEs, written as

$$\frac{d}{dt} \mathbf{u} = D M_1 \mathbf{u} \quad (2)$$

which diagonalizes to

$$\frac{d}{dt}V = \Lambda V. \quad (3)$$

(Since  $M_1$  is a symmetric matrix, it is diagonalizable.) If we can find the eigenvalues of the matrix  $DM_1$  then this will then allow us to use the stability studies from the Jan 20 notes of the ODE  $dy/dt = -\sigma y$  to understand what happens when using the one-step timestepping schemes (such as Forward Euler, Backward Euler, Crank-Nicolson) on the problem (1).

Note: We're arguing that if the time-stepping is stable (or unstable) on the diagonal system (3) then this will be inherited by the time-stepping on the original system (2). Inherent in this logic is that doing Forward Euler, say, on the original system (2) is the same, after change of coordinates, to doing Forward Euler on the diagonalized system (3). This is something you should check.

We are reduced to finding the  $N - 1$  eigenvalues and eigenvectors of  $DM_1$ . If we were asked for the eigenvalues and eigenfunctions of the elliptic problem

$$\begin{cases} u_{xx} = \lambda u & x \in (0, L) \\ u(0) = u(L) = 0 \end{cases}$$

then our immediate answer would be, "There are infinitely many:

$$u_k(x) = \sin\left(k \frac{\pi}{L} x\right) \quad \lambda_k = -\left(\frac{k\pi}{L}\right)^2 \quad \text{where } k \in \mathbb{N}." \quad (4)$$

This knowledge gives a natural guess for what the eigenvectors of  $DM_1$  might be:

$$V_k(j) = \sin\left(k \frac{\pi}{L} x_j\right) = \sin\left(k \frac{\pi}{L} jh\right) = \sin\left(k \frac{\pi}{N} j\right) \quad j = 1 \dots N - 1.$$

This gives the  $j$ th component of the (proposed) eigenvector. Since the vector depends on the wave number  $k$ , we label the vector with the index  $k$ . Multiplying  $V_k$  by the matrix  $DM_1$ , we find that indeed it is an eigenvector and its eigenvalue is

$$\lambda_k = -D \frac{2(1 - \cos(k\pi/N))}{h^2}.$$

And so we seem to have done our job. Except that we've found infinitely many eigenvector, eigenvalue pairs for an  $N - 1$  by  $N - 1$  matrix!

In fact, there are only  $N - 1$ . This is because of aliasing — if one samples a trig function on a uniform mesh then one can only distinguish finitely many frequencies, not infinitely many. How can you see this? Fix  $k$ . If we increase this wave number by an even multiple of  $N$  then:

$$\begin{aligned} V_{k+2lN}(j) &= \sin\left(\left(k + 2lN\right) \frac{\pi}{N} j\right) = \sin\left(k \frac{\pi}{N} j + 2l \pi j\right) = \sin\left(k \frac{\pi}{N} j\right) \cos(2l\pi j) \\ &= \sin\left(k \frac{\pi}{N} j\right) = V_k(j). \end{aligned}$$

If we increase the wave number  $k$  by an odd multiple of  $N$  then

$$\begin{aligned} V_{k+(2l+1)N}(j) &= V_{k+N}(j) = \sin\left(\left(k + N\right) \frac{\pi}{N} j\right) \\ &= \sin\left(k \frac{\pi}{N} j + \pi j\right) = \sin\left(k \frac{\pi}{N} j\right) \cos(\pi j) = -\sin\left(\left(N - k\right) \frac{\pi}{N} j\right) = -V_{N-k}(j). \end{aligned}$$

In this way, we see that there are only  $N - 1$  eigenvectors — all the candidate eigenvectors for  $k \geq N$  are either the zero vector (and thus ineligible to be eigenvectors) or they're one of the first  $N - 1$  eigenvectors or they're a multiple of one of them.

Now that we have our  $N - 1$  eigenvectors, we have the corresponding  $N - 1$  eigenvalues:

$$\lambda_k(h) = -D \frac{2(1 - \cos(k\pi/N))}{h^2} = -D \frac{2(1 - \cos(k\frac{\pi}{L}h))}{h^2} \quad 1 \leq k \leq N - 1. \quad (5)$$

The larger  $k$  is, the more negative  $\lambda_k$  is. And so the fastest rate of decay that needs to be resolved is determined by taking  $k = N - 1$ , resulting in

$$\sigma_{\max}(h) = D \frac{2(1 + \cos(h\pi/L))}{h^2} = 4 \frac{D}{h^2} - D \frac{\pi^2}{L^2} + \dots \quad (6)$$

As a result, if one is using Forward Euler, the timestep constraint is given by

$$k\sigma_{\max}(h) = kD \frac{2(1 + \cos(h\pi/L))}{h^2} \leq 2$$

and the critical timestep is  $k = 2/\sigma_{\max}$ . (I.e. the timestep at which the timestepper switches from being stable to being unstable.) When considering Forward Euler on  $h\mathbb{Z}$ , we found a timestep constraint  $Dk/h^2 \leq 1/2$ ; if this constraint was met then the maximum principle held. From this, the critical timestep is  $k = h^2/(2D)$ . From (6),  $\sigma_{\max}(h) < 4D/h^2$  and so the critical timestep for the problem on a bounded domain is *larger* than the critical timestep for the problem on  $h\mathbb{Z}$ . This shows us that the Dirichlet boundary conditions have a very slight stabilizing effect in that there's a small interval of timestep sizes which would result in unstable Forward Euler on the line but stable Forward Euler on the interval  $[0, L]$ .

## 2.2 Diffusion Equation on $\mathbb{R}$ and on $h\mathbb{Z}$

Because our PDEs are linear, their general solutions are linear combinations of simple elementary solutions. These solutions are of the form  $T(t)X(x)$  — they're separable with one factor providing the temporal behaviour and the other factor the spatial structure. The spatial structure  $X(x)$  is found by seeking an eigenfunction of the spatial part of the linear PDE. The resulting eigenvalue then enters into the ODE determining  $T(t)$ . If the PDE has been discretized, resulting in  $dU/dt = AU$  then the spatial structure  $X(j)$  is found by seeking eigenvectors of the (infinite or finite) matrix  $A$ .

For a constant-coefficient linear PDE on  $\mathbb{R}$ , the spatial functions  $X(x)$  are  $e^{i\xi x}$  where  $\xi$  is real. (If you don't like complex numbers, you can just think of  $\sin \xi x$  and  $\cos \xi x$  instead.) For  $u_t = Du_{xx}$ , these result in the solutions

$$u(x, t) = A(\xi, 0)e^{-\sigma(\xi)t}e^{i\xi x} = A(\xi, 0)e^{-D\xi^2 t}e^{i\xi x}, \quad \xi \in \mathbb{R}. \quad (7)$$

When we considered the problem (1) on the interval  $[0, L]$ , the boundary conditions restricted the allowable frequencies — rather than having a continuum of permissible frequencies ( $\xi \in \mathbb{R}$ ) there was an infinite sequence of permissible frequencies ( $\xi = k\pi/L$ ,  $k \in \mathbb{N}$ ).

We call

$$\sigma(\xi) = D\xi^2. \quad (8)$$

the *continuous dispersion relation*. In general, for a linear equation, the term “dispersion relation” refers to the relation giving the time behavior  $\sigma(\xi)$  in terms of the spatial wave number  $\xi$ . In this case it tells us that short-wavelength modes (large  $\xi$ ) decay rapidly (large positive  $\sigma$ ).

Now consider the discretization of the PDE, resulting in the semi-discrete model of infinitely many coupled ODEs on the spatial grid  $h\mathbb{Z}$ :

$$\frac{dU_j}{dt} = \frac{D}{h^2}(U_{j+1} - 2U_j + U_{j-1}) \quad j \in \mathbb{Z} \quad (9)$$

We again analyze this system using Fourier analysis. Consider the spatial structure  $X(j) = e^{i\xi x_j} = e^{i\xi jh} = \omega^j$  where  $\omega = e^{i\xi h}$ . The solution of the PDE is bounded in space and so the discrete solution should be as well. And so, we need  $\xi$  real resulting in  $|\omega| = 1$  (otherwise, either  $|\omega| < 1$  or  $|\omega| > 1$ , and thus exponential growth in space as  $j$  tends to  $\pm\infty$ ).

If the exact solution is

$$U_j(t) = A_h(\xi, t) \omega^j$$

then

$$\frac{dU_j}{dt} = \frac{D}{h^2} (U_{j+1} - 2U_j + U_{j-1}) = -D \frac{2(1 - \cos \xi h)}{h^2} A_h(\xi, t) \omega^j,$$

yielding the special solutions

$$U_j(t) = A(\xi, 0) e^{-\sigma_h(\xi)t} e^{i\xi jh} \quad \text{where} \quad \sigma(h) = -D \frac{2(1 - \cos \xi h)}{h^2}. \quad (10)$$

That is, the *discrete dispersion relation* is

$$\sigma_h(\xi) = 2D \frac{1 - \cos \xi h}{h^2}. \quad (11)$$

This should be compared with the continuous version (8). For small  $\xi h$ ,

$$\sigma_h(\xi) = \sigma(\xi) \left(1 - \frac{1}{12}(\xi h)^2 + \dots\right), \quad \xi h \rightarrow 0.$$

On the line, there's a plane wave solution (7) for every  $\xi \in \mathbb{R}$ . We seem to have found analogous solutions (10) on  $h\mathbb{Z}$ . Is it possible to “see” a function like  $\cos(\xi x)$  for arbitrarily large  $\xi$ , if one can only sample the function at points separated by  $h$ ? No. Given a spacing  $h$ , this determines a range of frequencies that can be resolved. The smaller the value of  $h$  is, the larger this range of frequencies will be. But it will never be  $\mathbb{R}$ . How large can  $\xi$  be? The highest wavenumber that can be represented on a discrete grid has wavelength equal to twice the grid spacing, so that alternate grid points have values  $\pm 1$ . (Higher wavenumbers are mapped back to lower ones by *aliasing*.) Thus the maximum value is  $\xi h = \pi$  — this means that values on the grid  $h\mathbb{Z}$  can represent any frequency  $\xi \in [-\pi/h, \pi/h]$ .

For  $0 < \xi \leq \pi/h$ , the discrete decay rate  $\sigma_h(\xi)$  satisfies  $\sigma(\xi) < \sigma_h(\xi) < 0$  (see Figure 1). So in the semi-discrete model (9) all modes decay though not as rapidly as for the PDE. Since  $\sigma_h(\xi)$  is a decreasing function, it takes its most negative value at  $\xi = \pi/h$ :

$$\sigma_{\max}(h) = D \frac{2(1 - \cos(\pi))}{h^2} = \frac{4D}{h^2}. \quad (12)$$

We can now use our understanding of the ODE system to predict the stability of the discrete methods for the PDE  $u_t = Du_{xx}$ ; we simply use  $\sigma_{\max}$  for  $\sigma$ . The parameter that controls stability is

$$\lambda = \frac{Dk}{h^2} = \frac{1}{4}\sigma_{\max}k.$$

- The **forward Euler** method is stable for  $\sigma_{\max}k \leq 2$ ; that is, for  $\lambda \leq \frac{1}{2}$ . If  $\lambda > \frac{1}{2}$ , then the instability will be a “checkerboard” of alternate-grid-point oscillations in both space and time, growing exponentially in time. For  $1 < \sigma_{\max}k \leq 2$ , or  $\frac{1}{4} < \lambda \leq \frac{1}{2}$ , it will display oscillations on successive time steps, but these oscillations will decay in time.

Recall that we had identified  $\lambda = \frac{1}{2}$  as the threshold value at which the solution ceased to be positive. This was done by considering special initial data, see the January 8 notes. Now we see that in fact the whole scheme is completely *unstable* for  $\lambda > \frac{1}{2}$ . Such instabilities will certainly cause the maximum principle to be violated, no matter what the initial data.

- The **backwards Euler** method is stable for all ratios of  $k$  and  $h^2$ . There will never be any oscillation in time.
- The **Crank-Nicolson** (trapezoid) method is stable for all ratios of  $k$  and  $h^2$ . For  $\sigma_{\max}k > 2$ , or  $\lambda > \frac{1}{2}$ , it will display oscillations on alternate time steps, but these oscillations will decay in time. (They are especially visible near a discontinuity in the initial data.)

For the forward Euler method for the ODE, we could always make the scheme stable by decreasing  $k$  until  $\sigma k \leq 2$ . For the semi-discrete model (9) one has  $k\sigma_{\max}(h) \leq 2$  and so whether or not the timestepping is stable depends on *how* we decrease  $h$  and  $k$ . If we fix  $\lambda$  and decrease  $h$  while determining  $k$  from  $k = \lambda h^2/D$ , then we will always be stable or always be unstable depending on whether  $\lambda \leq \frac{1}{2}$  or  $\lambda > \frac{1}{2}$ . If we choose  $k = \mu h$  for some fixed  $\mu$ , then the scheme will always be unstable when  $k$  and  $h$  are small, since  $Dk/h^2 = D\mu/h \rightarrow \infty$  as  $h \rightarrow 0$ , no matter what value  $\mu$  takes.

For  $\theta$  methods with  $\theta \geq \frac{1}{2}$  (such as Backward Euler or Crank-Nicolson) the scheme will be stable for *any* chosen relationship between  $k$  and  $h$ ; the linear scaling  $k = \mu h$  is often a convenient choice.

Figure 1 presents the four decay rates we’ve discussed: the discrete decay rates arising from the IBVP problem (1) on a bounded interval and the continuum of decay rates arising from the PDE on  $\mathbb{R}$  and the infinite system of ODEs on  $h\mathbb{Z}$ . Note that for  $\xi h \ll 1$ , the discrete decay rates are close to the decay rates for the PDE.

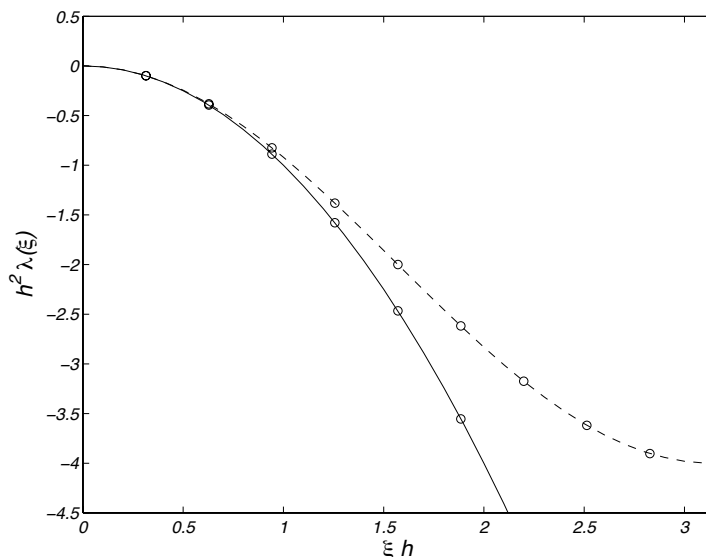


Figure 1: Continuous and discrete decay rates  $\sigma(\xi)$  and  $\sigma_h(\xi)$ . The  $x$ -axis is scaled by  $h$  and the  $y$ -axis is scaled by  $h^2$ . The diffusion constant  $D$  is taken to be 1. The solid line is the graph of  $-h^2\sigma(\xi)$ ; the dashed line is the graph of  $-h^2\sigma_h(\xi)$ . The open circles are the decay rates for the IBVP on a bounded interval for the value  $N = 10$ . There are  $N - 1$  circles on the discrete decay rate curve, corresponding to the 9 values given by (4) and (5). Note that what's plotted is  $-h^2\sigma(\xi)$  and  $-h^2\sigma_h(\xi)$ .

### 2.3 von Neumann Analysis

What do those plane wave solutions (7) have to do with anything? In this section, we revisit Section 2.2's stability analysis for the problem on  $\mathbb{R}$  and  $h\mathbb{Z}$ . There, we analysed the problem in real space — all our functions were considered as functions of  $x$  and  $t$ . We will now analyse the problem in Fourier space — by applying the Fourier transform we consider the problem in terms of functions of  $\xi$  and  $t$ . This approach was discovered by John von Neumann and it is significantly easier in terms of computations.

We start with a quick review of Fourier analysis. Recall that for a real- or complex-valued function  $u(x)$  defined on  $\mathbb{R}$  we can define the Fourier

transform which is another function on  $\mathbb{R}$

$$\hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u(x) e^{-i\xi x} dx.$$

This is defined for a wide class of functions  $u$ , for simplicity's sake let's assume all our functions have finite  $L^1$  norm:  $\int |u(x)| dx < \infty$ . The Fourier inversion formula states

$$u(x) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \hat{u}(\xi) e^{i\xi x} d\xi.$$

Similarly, if one has a function  $\{v_j\}$  defined on  $\mathbb{Z}$  then its Fourier transform is

$$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} v_j e^{-i\xi j}$$

which is defined for all  $\xi$  in  $[-\pi, \pi]$  The Fourier inversion formula is

$$v_j = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \hat{v}(\xi) e^{i\xi j} d\xi$$

and Parseval's theorem is

$$\sum_{j=-\infty}^{\infty} |v_j|^2 = \int_{-\pi}^{\pi} |\hat{v}(\xi)|^2 d\xi.$$

If the function is defined on  $h\mathbb{Z}$  then by a change of variables,

$$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{j=-\infty}^{\infty} v_j e^{-i\xi j h},$$

for  $\xi \in [-\pi/h, \pi/h]$  and the inversion formula is

$$v_j = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \hat{v}(\xi) e^{i\xi j h} d\xi$$

and Parseval's theorem is

$$\|\hat{v}\|_h^2 = \int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 d\xi = \sum_{j=-\infty}^{\infty} |v_j|^2 h = \|v\|_h^2$$

Note that if  $\omega \in \mathbb{R}$  then  $\omega = \xi + N2\pi/h$  for some  $\xi \in [-\pi/h, \pi/h]$  and some  $N \in \mathbb{Z}$ . If we then sample the function  $\exp(i\omega x)$  on  $h\mathbb{Z}$  we find

$$e^{i\omega j h} = e^{i(\xi + N2\pi/h) j h} = e^{i\xi j h}.$$

In short, the grid  $h\mathbb{Z}$  cannot “see” high frequencies  $\omega$  such that  $|\omega| > \pi/h$ . (This is exactly what we observed before.)

Let’s return to the Forward Euler scheme for  $u_t = Du_{xx}$ :

$$u_j^{n+1} = \lambda u_{j+1}^n + (1 - 2\lambda)u_j^n + \lambda u_{j-1}^n \quad (13)$$

where  $\lambda = Dk/h^2$ . We write each of the terms in terms of its Fourier transform

$$\begin{aligned} u_j^{n+1} &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \widehat{u}^{n+1}(\xi) e^{i\xi j h} d\xi \\ u_j^n &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \widehat{u}^n(\xi) e^{i\xi j h} d\xi \\ u_{j+1}^n &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \widehat{u}^n(\xi) e^{i\xi(j+1)h} d\xi \\ u_{j-1}^n &= \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \widehat{u}^n(\xi) e^{i\xi(j-1)h} d\xi \end{aligned}$$

hence the Forward Euler scheme (13) becomes

$$\int_{-\pi/h}^{\pi/h} \widehat{u}^{n+1}(\xi) e^{i\xi j h} d\xi = \int_{-\pi/h}^{\pi/h} \left( \lambda e^{-i h \xi} + (1 - 2\lambda) + \lambda e^{i h \xi} \right) \widehat{u}^n(\xi) e^{i\xi j h} d\xi$$

from which we conclude

$$\widehat{u}^{n+1}(\xi) = \left( \lambda e^{-i h \xi} + (1 - 2\lambda) + \lambda e^{i h \xi} \right) \widehat{u}^n(\xi) = \sigma_h(\xi) \widehat{u}^n(\xi),$$

for  $-\pi \leq h\xi \leq \pi$ . And so we see that advancing the solution of the finite difference scheme by one step is the same as multiplying its Fourier transform by the factor  $\sigma_h(\xi)$ . Going all the way back to the initial data

$$\widehat{u}^n(\xi) = \sigma_h(\xi)^n \widehat{u}_0(\xi), \quad \text{for } -\pi \leq h\xi \leq \pi.$$

Using Parseval’s theorem, we can relate the size of the solution at the  $n$ th timestep to what we’ve learnt from time-stepping in Fourier space:

$$\|u^n\|_h^2 = \sum_{j=-\infty}^{\infty} |u_j^n|^2 h = \int_{-\pi/h}^{\pi/h} |\sigma_h(\xi)^{2n} |\widehat{u}_0(\xi)|^2 d\xi.$$

From this, it’s clear that if there’s an interval in  $[-\pi/h, \pi/h]$  on which  $|\sigma_h(\xi)| > 1$  then the solution  $u^n$  will grow exponentially with  $n$ .

We now analyse the factor

$$\sigma_h(\xi) = \lambda e^{-i h \xi} + (1 - 2\lambda) + \lambda e^{i h \xi} = 2\lambda \cos(h\xi) + 1 - 2\lambda.$$

If  $|\sigma_h(\xi)| > 1$  then the scheme is not stable. Note that  $\sigma_h(\xi)$  is an even function in  $\xi$ , it equals 1 at  $\xi = 0$ , and is decreasing in  $\xi$  on  $[0, \pi/h]$ . Its minimum value is at  $h\xi = \pm\pi$

$$\max_{\xi \in [-\pi/h, \pi/h]} \sigma_h(\xi) = 1, \quad \min_{\xi \in [-\pi/h, \pi/h]} \sigma_h(\xi) = 1 - 4\lambda$$

It follows immediately that if  $1 - 4\lambda < -1$  (that is,  $\lambda > 1/2$ ) then the scheme is unstable.

We now consider the Backward Euler scheme

$$\frac{u_j^{n+1} - u_j^n}{k} = D \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} \implies \sigma_h(\xi) = \frac{1}{1 + 2\lambda - 2\lambda \cos(h\xi)}$$

Again,  $\sigma_h(\xi)$  is an even function in  $\xi$ , it equals 1 at  $\xi = 0$ , and is decreasing in  $\xi$  on  $[0, \pi/h]$ . Its minimum value is at  $h\xi = \pm\pi$

$$\max_{\xi \in [-\pi/h, \pi/h]} \sigma_h(\xi) = 1, \quad \min_{\xi \in [-\pi/h, \pi/h]} \sigma_h(\xi) = \frac{1}{1 + 4\lambda} > 0 > -1$$

In this way, we see that the scheme is stable for all  $\lambda$  and the factor  $\sigma_h(\xi) > 0$  for all  $\xi \in [-\pi/h, \pi/h]$ .

Finally, we consider the Crank-Nicolson scheme

$$\begin{aligned} \frac{u_j^{n+1} - u_j^n}{k} &= D \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{2h^2} + D \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2h^2} \\ &\implies \sigma_h(\xi) = \frac{1 + \lambda \cos(h\xi) - \lambda}{1 + \lambda - \lambda \cos(h\xi)} \end{aligned}$$

As above,

$$\max_{\xi \in [-\pi/h, \pi/h]} \sigma_h(\xi) = 1, \quad \min_{\xi \in [-\pi/h, \pi/h]} \sigma_h(\xi) = \frac{1 - 2\lambda}{1 + 2\lambda} > -1$$

In this way, we see that the scheme is stable for all  $\lambda$ .

If you do the stability analysis of the  $\theta$ -scheme

$$\frac{u_j^{n+1} - u_j^n}{k} = \theta D \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} + (1 - \theta) D \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}$$

you will find that it's stable for all  $\lambda$  if  $\theta \geq 1/2$ . And that it's stable for  $\lambda \leq \lambda_\theta$  otherwise.

Above, the analysis was applied to the problem on  $h\mathbb{Z}$ . What does one do if you have the problem on a bounded domain? One can do something quite similar except the interval of frequencies  $[-\pi/h, \pi/h]$  will be replaced by a finite collection of frequencies. We'll discuss this more when we consider spectral methods.

## 2.4 Summary

Ideally, once you've implemented the spatial discretization and the boundary conditions for the problem you're interested in, you would find the eigenvalues and eigenvectors of the resulting matrix and use these in the stability analysis of your timestepper. This is what's done in Subsection 2.1. However, this can be quite difficult to do in practice. As a result, what one often does is to use Fourier methods to analyse the discretized problem in the unbounded domain, as is done in Subsections 2.2 and 2.3. It's not a perfect approach but is often very useful.

NOTE: in the von Neumann analysis we were in Fourier space and so we needed to know if the multipliers  $\sigma_h(\xi)$  had magnitude less than one. In the plane wave analysis of Section 2.2, we were in real space and we needed to know if  $\sigma_h(\xi)$  was positive. Both sections had a  $\sigma_h(\xi)$  but they were doing slightly different things.

NOTE: we concerned ourselves with the diffusion equation, for which we knew the dispersion relation  $\sigma(\xi) = -D\xi^2$ . If one had some other linear PDE (wave equation, advection equation, etc.) then one would seek timesteppers whose dispersion relations respected the corresponding  $\sigma(\xi)$ .

## 3 Convergence, Consistency, and Stability

We wish to time-step the ODE:

$$\frac{d\mathbf{U}}{dt} = f(\mathbf{U}(t), t) \quad \text{for } t \in [0, T], \quad \text{with } \mathbf{U}(0) = \mathbf{U}_0.$$

Choose an integer  $M$  and a final time  $T$ , thus determining the time step  $k = T/M > 0$ . Define time levels  $t_n = nk$ , for  $n = 0, 1, \dots, M$ . We want to construct a sequence of points in  $\mathbb{R}^N$ :  $u^0, u^1, \dots, u^M$ , so that  $u^n \approx \mathbf{U}(nk)$  when  $k$  is small. Here  $u^n$  is the solution of the fully discrete model, and  $\mathbf{U}(nk)$  is the exact solution of the ODE, evaluated at the discrete time points.

We need a scheme for constructing the discrete approximation; such a scheme must consist of two things. First, we need a rule for picking the initial value  $u^0$ . Since  $\mathbf{U}(t)$  and  $u^n$  are members of the same space, we can usually just take  $u^0 = \mathbf{U}(0)$ . Second, we need a rule of the form  $u^{n+1} = F(u^n, \dots, u^0; k)$  for computing each step of the approximation in terms of earlier steps. In fact,  $F$  could well be different for each step  $n$ .

A rich class of schemes are the “linear multistep formulas,” for which  $u^{n+1}$  is a linear function of  $u^n, \dots, u^0, f^n, \dots, f^0$ , and possibly  $f^{n+1}$ , where  $f^n = f(u^n, t_n)$ :

$$u^{n+1} = \sum_{j=0}^p a_j u^{n-j} + k \sum_{j=-1}^p b_j f^{n-j}$$

where  $p > 0$  and  $a_0, \dots, a_p$  and  $b_{-1}, \dots, b_p$  are constants. If either  $a_p$  or  $b_p$  is nonzero this is called a  $p + 1$ -multistep method. If  $b_{-1} = 0$  the scheme is *explicit*. Otherwise, it is an *implicit* scheme.

Note that a 1-step scheme needs only one initial value: given  $u^0$  one can define  $u^1$  which one can then use to define  $u^2$  which one can then use to define  $u^3$  and so on. For a  $p$ -step scheme with  $p > 1$  one needs  $p$  values ( $u^0, \dots, u^{p-1}$ ) to get the process started. One usually does this with a sequence of  $q$ -step regimes where  $q$  increases from 1 to  $p - 1$ . For example, if one had a 3-step scheme, say, what one usually does is: given  $u^0$  one uses a 1-step scheme to define  $u^1$ . One then uses a 2-step scheme to define  $u^2$  from  $u^0$  and  $u^1$  and one can then use the 3-step scheme from then on.

Given a time-stepping scheme for an ODE there are three immediate questions: is the scheme consistent? is it convergent? is it stable? For multistep schemes for nonlinear ODEs, these questions are rather tricky. They are much simpler for 1-step methods applied to linear PDEs, and so we now restrict our attention to this case. Much of the following comes from sections 1.4 and 1.5 of Strikwerda’s “Finite Difference Schemes and Partial Differential Equations” which is on reserve.

### 3.1 Convergence

Consider a linear partial differential equation of the form

$$P(\partial_t, \partial_x)u(x, t) = f(x, t) \tag{14}$$

which is of first order in the derivative with respect to  $t$ . Examples would include:

$$\begin{aligned} u_t - bu_{xx} + au_x &= 0 \\ u_t - cu_{txx} + bu_{xxxx} &= 0 \\ u_t + cu_{tx} + au_x &= 0 \end{aligned}$$

Consider a one-step finite difference scheme for the equation

$$P_{k,h}u = f. \quad (15)$$

Given initial data  $\{u_j^0\}$  where  $j \in \mathbb{Z}$  we denote the solution of the scheme (15) by  $u_j^n$  where  $j$  ranges over  $\mathbb{Z}$  and  $n$  ranges over  $\mathbb{N}$ .

*Let  $u(x, t)$  be the solution of the PDE (14) with initial data  $u_0(x)$ . Given a spatial grid spacing  $h$ , generate approximate initial data  $\{u_j^0\}$  in such a way that as  $jh$  converges to  $x$  the approximate initial data  $u_j^0$  converges to  $u_0(x)$ . For each fixed  $h$  and  $k$ , let  $u_j^n$  be the resulting solution of the one-step finite difference scheme (15). The scheme is convergent if  $u_j^n$  converges to  $u(x, t)$  as  $(jh, nk)$  converges to  $(x, t)$  as  $h$  and  $k$  converge to 0.*

Usually, the approximate initial data is taken to be  $u_0(mh)$  — just sample the function at the grid points. Note that this “definition” of convergence is somewhat squishy because I haven’t rigorously specified what I mean by a function that’s defined on a grid ( $u_j^n$ ) to converge to a function that’s defined on a continuous region ( $u(x, t)$ ).

### 3.2 Consistency

In practice, proving that a scheme is convergent is not easy. However, if the Lax-Richtmyer theorem holds then if your scheme is “consistent” and “stable” then it will be convergent. And checking consistency and stability isn’t hard.

We start by defining what “consistency” means.

*Given a partial differential equation,  $Pu = f$ , and a finite difference scheme,  $P_{k,h}u = f$ , we say the finite difference scheme is consistent with the partial differential equation if for any smooth  $\phi(x, t)$*

$$P\phi - P_{k,h}\phi \rightarrow 0, \quad \text{as } k, h \rightarrow 0,$$

the convergence being pointwise convergence at each grid point. Note that unlike the the definition of local truncation error, we are not requiring that  $\phi(x, t)$  be a solution of the PDE.

### 3.3 Stability

We now define what stability means for a *homogeneous* finite difference scheme to be stable

A finite difference scheme  $P_{k,h}u = 0$  for a first-order equation is stable if there is an integer  $J$  and positive numbers  $h_0$  and  $k_0$  such that for any positive time  $T$  there is a constant  $C_T$  such that

$$h \sum_{j=-\infty}^{\infty} |u_j^n|^2 \leq \tilde{C}_T \sum_{l=0}^J h \sum_{j=-\infty}^{\infty} |u_j^l|^2$$

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

Note that if we introduce the discrete analogue of the  $L^2(\mathbb{R})$  norm

$$\|w\|_h = \left( h \sum_{j=-\infty}^{\infty} |w_j|^2 \right)^{1/2}$$

then the stability constraint can be written as

$$\|u^n\|_h^2 \leq \tilde{C}_T \sum_{l=0}^J \|u^l\|_h^2$$

and hence

$$\|u^n\|_h \leq C_T \sum_{l=0}^J \|u^l\|_h \tag{16}$$

for some  $C_T$ .

The stability inequality (16) states that the  $L^2$  norm of the discrete solution at time level  $n$  is bounded by some constant times the sum of the  $L^2$  norms of the discrete solution at the first  $J + 1$  levels. For one-step schemes, we will see that one can take  $J = 0$ : the scheme is stable if there is a constant  $\tilde{C}_T$  so that

$$h \sum_{j=-\infty}^{\infty} |u_j^n|^2 \leq \tilde{C}_T h \sum_{j=-\infty}^{\infty} |u_j^0|^2.$$