21-761 Finite Difference Methods
Spring 2010

Yekaterina Epshteyn
notes by Brendan Sullivan

April 23, 2010

Contents

0 Introduction

Our main interest is in solving initial boundary value problems.

Example 0.1. Consider the *Poisson equation*, a standard example of an *elliptic PDE*

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= g(x) && \text{on } \Gamma \end{aligned}$$

where $x = (x_1, x_2, \ldots, x_n)$ and $\Gamma \subseteq \partial\Omega$. Recall that $\Delta u = \sum_j \frac{\partial^2 u}{\partial x_j^2}$. The second line above is the boundary data, and such a condition is known as *Dirichlet* boundary data. We will also consider *Neumann* boundary data

$$\frac{\partial u}{\partial \underline{n}} = g(x) \text{ on } \Gamma$$

where $\underline{n}$ is the outward unit normal, and so $\frac{\partial u}{\partial \underline{n}} = \nabla u \cdot \underline{n}$. We will also consider *Robin* boundary data

$$\frac{\partial u}{\partial \underline{n}} + \beta u(x) = g(x) \text{ on } \Gamma$$

More generally, an elliptic PDE is of the form $Au = f$ for some operator

$$Au := -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + \sum_{j=1}^{d} b_j(x)\frac{\partial u}{\partial x_j}$$

with $A(x) := (a_{ij}(x))_{i,j=1}^{d}$ a positive-definite matrix.

Example 0.2. Consider the *heat equation*

$$\begin{aligned} \tfrac{\partial u}{\partial t} - \Delta u &= f(x,t) \\ u(x,0) &= u_0(x) \end{aligned}$$

plus some boundary data. This is an example of a parabolic (linear) PDE.

Example 0.3. Consider the *wave equation*

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f(x,t)$$

This is an example of a hyperbolic PDE.

1 The Finite Difference Method

Consider a simple FDM on the ODE

$$\begin{aligned} u''(x) &= f(x) \quad \text{in } [0,1] \\ u(0) &= \alpha \\ u(1) &= \beta \end{aligned} \tag{1}$$

Divide $[0,1]$ into subintervals by identifying the nodes $x_j = jh$ with $h = \frac{1}{m+1}$, for $j = 0, 1, \ldots, m+1$. We seek to approximate $u_j := u(x_j)$. We replace u'' by using the *centered difference approximation*

$$u''(x_j) \approx \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f(x_j) \quad , j = 1, 2, \ldots, m$$

and $u_0 = u(0) = \alpha$ and $u_{m+1} = u(1) = \beta$. Then we define the solution vector $U = [u_1, u_2, \ldots, u_m]^T$ and we seek the solution to the linear system $AU = F$ where

$$A = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & & \\ 1 & -2 & 1 & 0 & \\ 0 & 1 & -2 & 1 & 0 \\ & & \ddots & & \\ & 0 & 1 & -2 & 1 \\ & & 0 & 1 & -2 \end{bmatrix}$$

and

$$F = \begin{bmatrix} f(x_1) - \frac{\alpha}{h^2} \\ f(x_2) \\ \vdots \\ f(x_m) - \frac{\beta}{h^2} \end{bmatrix}$$

1.1 Error Analysis

The global error of the approximation U we obtain is give by $E = U - u$, where u is the *true* solution vector. To quantify this error, we use various norms:

1. $\|E\|_\infty = \max_{1 \le j \le m} |E_j| = \max_{1 \le j \le m} |U_j - u_j|$
2. $\|E\|_1 = h \sum_{j=1}^m |E_j|$
3. $\|E\|_2 = \left(h \sum_{j=1}^m |E_j|^2 \right)^{1/2}$

The space $\mathbb{R}^d$ is finite-dimensional, so all of these norms are equivalent, topologically, but the constant of equivalence typically involves some power of h, so convergence *rates* may vary in different norms.

Definition 1.1. *We consider the* local truncation error $\tau = [\tau_j]_{j=1}^m$ *given by*

$$\tau_j := \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} - f(x_j) \quad , 1 \leq j \leq m$$

By Taylor's Theorem, we can write

$$u(x_{j-1}) = u(x_j) - hu'(x_j) + \frac{h^2}{2}u''(x_j) - \frac{h^3}{6}u'''(x_j) + \frac{h^4}{24}u^{(4)}(x_j) - \frac{h^5}{120}u^{(5)}(x_j) + O(h^6)$$

and

$$u(x_{j+1}) = u(x_j) + hu'(x_j) + \frac{h^2}{2}u''(x_j) + \frac{h^3}{6}u'''(x_j) + \frac{h^4}{24}u^{(4)}(x_j) + \frac{h^5}{120}u^{(5)}(x_j) + O(h^6)$$

so then

$$\tau_j = \frac{h^2}{12}u^{(4)}(x_j) + u''(x_j) + O(h^4) - f(x_j) = \frac{h^2}{12}u^{(4)}(x_j) + O(h^4)$$

and so $\tau_j = O(h^2)$ as $h \to 0$.

Thinking of $\tau = Au - F$ and $AU = F$, then we can subtract and write the global error as $AE = -\tau$. Since each depends on h, we typically write $A^h E^h = -\tau^h$. Assuming that $(A^h)^{-1}$ exists, we can write the error as

$$E^h = -\left(A^h\right)^{-1} \tau^h$$

and so

$$\|E^h\| = \left\|\left(A^h\right)^{-1}\right\| \cdot \|\tau^h\|$$

in any norm. In order to have $\|E^h\| = O(\|\tau^h\|)$, we need $\left\|\left(A^h\right)^{-1}\right\| \leq C$ for all sufficiently small h. If this holds, then $\|E^h\| \leq C\|\tau^h\|$.

Definition 1.2. *Suppose a FDM for a linear BVP gives a sequence of matrix equations $A^h U^h = F^h$, where h is the width of the mesh. We say that the method is* stable *provided $\left(A^h\right)^{-1}$ exists $\forall h < h_0$ and if $\exists C$ independent of h such that $\left\|\left(A^h\right)^{-1}\right\| \leq C \ \forall h < h_0$.*

Remark 1.3. Note that stability in one norm $\Rightarrow$ stability in any equivalent norm.

Definition 1.4. *We say a method is* consistent *with the original ODE and BCs if* $\|\tau^h\| \to 0$ *as* $h \to 0$.

Definition 1.5. *We say a method is* convergent *if* $\|E^h\| \to 0$ *as* $h \to 0$.

Theorem 1.6 (Fundamental Theorem of FDs)**.** *If a scheme is consistent and stable, then it is convergent.*

Proof. By assumption, we may write

$$\|E^h\| \leq \left\| \left(A^h\right)^{-1} \right\| \cdot \|\tau^h\| \leq C\|\tau^h\| \to 0$$

as $h \to$), which shows convergence. □

In general, we can say that stability and order $O(h^p)$ of local truncation error implies order $O(h^p)$ of global error. Note that the order *does* depend on the choice of norm.

Note that in general there is a practical tradeoff between complexity of programming and scheme convergence rate. Also, it is vastly more inefficient (in Matlab) to use the command $U =$inv$(A) \cdot F$ to solve the linear system, since this uses at least n^2 operations, whereas $U = A \setminus F$ (or some command like that) is more efficient, since it produces the product without actually computing the inverse.

It is a good idea to check for stability at the continuum level, to see whether we have a chance of achieving stability for the discretized scheme. For this specific model problem, we can use the Poincaré and Cauchy-Schwarz inequalities to take $\Delta E = \tau$, multiply both sides by E, then integrate by parts and write

$$c \int E^2 \leq \int |\nabla E|^2 = -\int \tau E \leq \left(\int \tau^2 \right)^{1/2} \left(\int E^2 \right)^{1/2}$$

and then divide, yielding $\|E\|_2 \leq c\|\tau\|_2$. For the discretized scheme, we want to write

$$\|E^h\|_2 = \left\|(A^h)^{-1}\tau^h\right\|_2 \leq \left\|(A^h)^{-1}\right\| \left\|\tau^h\right\|_2$$

and the appropriate choice for the norm on $(A^h)^{-1}$ is the *operator norm*:

$$\|A\|^2 = \sup_{v \neq \vec{0}} \frac{\|Av\|_2^2}{\|v\|_2^2}$$

Since A is symmetric, for this specific problem, we know $\exists$ eigenvalues $\lambda_1, \ldots, \lambda_m$ and eigenvectors $v_1, \ldots, v_m$ such that $v = \sum_i \alpha_i v_i$, and $Av = \sum_i \alpha_i \lambda_i v_i$. Then,

$$\|A\|^2 = \sup_{v \neq \vec{0}} \frac{\sum \alpha_i^2 \lambda_i^2}{\sum_i \alpha_i^2} = \max_i \lambda_i^2$$

and so $\|A\| = \max_i |\lambda_i|$. We now introduce the *eigenfunctions* $u_i = \sin(i\pi x)$ and $u_j^p = \sin(\pi pjh)$ for $h = \frac{1}{m+1}$ and $j, p = 1, \ldots, m$. Then

$$\lambda p = \frac{2}{h^2} \left(\cos(\pi ph) - 1\right) \approx -\pi^2 + O(h^2)$$

since $\cos x \geq 1 - \frac{x^2}{2}$. Then the eigenvalues of $(A^h)^{-1}$ are

$$\left|\frac{1}{\lambda p}\right| \sim \left|\frac{-1}{p^2\pi^2}\right| \gtrsim \frac{1}{\pi^2}$$

and so

$$\|E^h\|_2 \leq \frac{1}{\pi^2}\|\tau^h\|$$

Now, what about Neumann boundary conditions? We investigate the problem

$$u'' = f \quad , \quad u'(0) = b \quad , \quad u(1) = \beta$$

We have three different ideas to try.

1. Try setting $u_0 = \alpha$ and $\frac{u_1 - u_0}{h} = b$. Then we know

   $$\hat{u}(x_1) - \hat{u}(x_0) = \hat{u}'(x_0)h + \frac{\hat{u}''(x_c)}{2}h^2$$

   where $\hat{u}$ is the *actual* solution and x_c is some point guaranteed by the Mean Value Theorem. Then we have

   $$\frac{\hat{u}(x_1) - \hat{u}(x_0}{h} - b = \frac{f(x_c)}{2}h$$

   which yields an error of order h, which is not as good as before.

2. To motivate the next idea, recall that

   $$\lim_{h\to 0}\left|\frac{f(x+h) - f(x)}{h} - f'(x)\right| = 0$$

   but

   $$\lim_{h\to 0}\left|\frac{1}{h}\left(\frac{f(x+h) - f(x-h)}{2h} - f'(x)\right)\right| = 0$$

   As such, we "introduce" u_{-1} and set

   $$\frac{u_{-1} - 2u_0 + u_1}{h^2} = f(x_0) \qquad \text{and} \qquad \frac{u_1 - u_{-1}}{2h} = b$$

   Really, we use the second equation to "find" u_{-1} and plug this value into the first equation.

3. The "best" idea yields a second-order accurate approximation. We set

   $$\frac{1}{h}\left(-\frac{3}{2}u_0 + 2u_1 - \frac{u_2}{2}\right) = b$$

Example 1.7. The problem

$$u'' \equiv 0 \quad , \quad u'(0) = 1 \quad , \quad u'(1) = 0$$

has no solution. Specifically, the matrix in the discrete scheme becomes *singular*. This illustrates how things can go wrong.

Recall the 3rd approach above. This is an illustration of the more general *method of undetermined coefficients*. Suppose that

$$u'(0) = au(0) + bu(h) + cu(2h)$$

Applying Taylor's Theorem, we have

$$u(h) = u(0) + hu'(0) + \frac{h^2}{2}u''(0) + \frac{h^3}{6}u'''(\xi_1) \quad \text{for some} \xi_1 \in (0, h) \tag{2}$$

and

$$u(2h) = u(0) + 2hu'(0) + 2h^2u''(0) + \frac{4h^3}{3}u'''(\xi_2) \quad \text{for some } \xi_2 \in (0, 2h) \tag{3}$$

Combining these, we have

$$\begin{aligned} u'(0) &\approx au(0) + b\left(u(0) + hu'(0) + \frac{h^2}{2}u''(0) + \frac{h^3}{6}u'''(\xi_1)\right) \\ &\quad + c\left(u(0) + 2hu'(0) + 2h^2u''(0) + \frac{4h^3}{3}u'''(\xi_2)\right) \\ &= (a+b+c)u(0) + (b+2c)hu'(0) + \left(\frac{b+4c}{2}\right)h^2u''(0) + O(h^3) \end{aligned}$$

so we require

$$a + b + c = 0 \quad , \quad b + 2c = \frac{1}{h} \quad , \quad \frac{b+4c}{2} = 0$$

to make $u'(0) \approx u'(0) + O(h^3)$. The solution is

$$a = -\frac{3}{2h} \quad , \quad b = \frac{2}{h} \quad , \quad c = -\frac{1}{2h}$$

This is precisely approach 3 from before. Note that $b + c \sim \frac{1}{h}$ which yields an $O(h^2)$ error overall.

1.2 Existence and Uniqueness

Example 1.8. Consider solving

$$u''(x) = f(x) \quad , \quad u'(0) = \sigma_0 \quad , \quad u'(1) = \sigma_1$$

This problem is not *well-posed*. A solution will be of the form

$$u(x) = \frac{x^2}{2} + c_1x + c_2$$

which can be obtained by integrating, with $c_1 = \sigma_0 = \sigma_1 - 1$. Therefore, we either have infinitely-many solutions or none. The discretized version of this

problem yields the matrix/vector equation

$$\frac{1}{h^2}\begin{bmatrix} -h & h & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & & & \\ & & & 1 & -2 & 1 \\ & & & & h & -h \end{bmatrix}\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_m \\ u_{m+1} \end{bmatrix} = \begin{bmatrix} -\sigma_0 + \frac{h}{2}f(0) \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ -\sigma_1 + \frac{h}{2}f(x_{m+1}) \end{bmatrix}$$

Notice that the matrix A_h is singular because $A_h[1, 1, \ldots, 1]^T = \vec{0}$.

1.3 Elliptic Problems in 2-D

Consider the general form of an elliptic problem

$$a_1 u_{xx} + a_2 u_{xy} + a_3 u_{yy} + a_4 u_x + a_5 u_y + a_6 u = f(x, y)$$

where $a_2^2 - 4a_1 a_3 < 0$ (the ellipticity condition). An example is Laplace's equation $u_{xx} + u_{yy} = 0$, or something like $(ku_x)_x + (ku_y)_y = f(x, y)$.

Let's examine Laplace's equation on a square. Set

$$\Omega = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\}$$

and consider solving

$$\Delta u := u_{xx} + u_{yy} = f(x, y) \text{ in } \Omega \quad , \quad \text{with } u \restriction_{\partial\Omega} = g(x, y)$$

Note the Dirichlet boundary condition. We set $x_i = i\Delta x$ and $y_j = j\Delta y$ and write $u(x_i, y_j) \approx u_{i,j}$. We use the approximations

$$u_{xx} \approx \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{(\Delta x)^2}$$

and

$$u_{yy} \approx \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{(\Delta y)^2}$$

Note that j is fixed in the first approximation, so it works just like the 1-D approximation, and likewise for i fixed in the second one. Combining these, we have

$$\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{(\Delta x)^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{(\Delta y)^2} = f_{i,j} \approx f(x_i, y_j)$$

If $\Delta x = \Delta y = h$, then

$$f_{i,j} = \frac{u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j}}{h^2}$$

Since $i = 1, \ldots, m$ and $j = 1, \ldots, m$ then we have m^2 grid points, so the unknown solution "vector" is $u = u_{11}, u_{12}, \ldots, u_{21}, u_{22}, \ldots, u_{mm}$ with m^2 entries.

We still write $A_h U = F$, and we can think of U and F as m^2-length vectors or $m \times m$ matrices, essentially. However, there is nothing to stop us from switching the order of the entries in U, which leads to different forms of the matrix A_h. For example, if we order U as

$$U = \begin{bmatrix} u_{1,1} & u_{2,1} & \cdots & u_{m,1} \\ u_{1,2} & u_{2,2} & \cdots & u_{m,2} \\ & \vdots & & \\ u_{1,m} & u_{2,m} & \cdots u_{m,m} & \end{bmatrix}$$

then one can show (as an exercise) that the $m^2 \times m^2$ matrix A_h will have the form

$$A_h = \frac{1}{h^2} \begin{bmatrix} T & I & & & \\ I & T & I & & \\ & I & T & I & \\ & & \ddots & & \\ & & I & T & I \\ & & & I & T \end{bmatrix}$$

where I is the $m \times m$ identity matrix and T is the $m \times m$ matrix given by

$$T = \begin{bmatrix} -4 & 1 & & & \\ 1 & -4 & 1 & & \\ & & \ddots & & \\ & & 1 & -4 & 1 \\ & & & 1 & -4 \end{bmatrix}$$

This form of A_h is good for linear solvers, in practice.

1.3.1 Accuracy and Stability

This is similar to the 1-D case. We define the local truncation error

$$\tau_{i,j} := \frac{1}{h^2} \Big(u(x_{i-1}, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) + \\ + u(x_i, y_{j+1}) - 4u(x_i, y_j) \Big) - f(x_i, y_j)$$

Splitting the difference in the x and y directions, one can show that

$$\tau_{i,j} = \frac{1}{12} h^2 \left(u_{xxxx} + u_{yyyy} \right) + O(h^4)$$

We then define

$$E_{i,j} := u_{i,j} - u(x_i, y_j)$$

to be the *global* error at (x_i, y_j), and so $A_h E = -\tau$, as before. Since the matrix A_h is symmetric, we can write

$$\|A_h\|_2 = \rho(A_h) = \max_{(p,q)} \left| \lambda^h_{p,q} \right|$$

where ρ indicates the *spectral radius*, so $\lambda_{p,q}^h$ are eigenvalues of A_h. It can be shown that

$$\left\|A_h^{-1}\right\|_2 = \left(\min_{1\le p,q\le m} \left|\lambda_{p,q}^h\right|\right)^{-1} \approx \frac{1}{2\pi^2}$$

for sufficiently small h, which implies $\left\|A_h^{-1}\right\| \le C \ \forall h < h_0$ small enough. We can then write

$$A_h E = -\tau \Rightarrow \|E\| \le \left\|A_h^{-1}\right\|_2 \|\tau\|_2 \le C \|\tau\|_2$$

so the method is stable with respect to $\|\cdot\|_2$, and $\|E\|_2 = O(h^2)$.

2 Iterative Methods for Sparse Linear Systems

We will examine the Jacobi, Gauss-Seidel, and Successive Over-relaxation methods.

2.1 Jacobi and Gauss-Seidel Methods

Define the tolerance $\varepsilon = 10^{-5}$. The Jacobi iteration is defined as

$$u_{i,j}^{(k=1)} = \frac{1}{4}\left(u_{i-1,j}^{(k)} + u_{i+1,j}^{(k)} + u_{i,j+1}^{(k)} + u_{i,j-1}^{(k)}\right) - \frac{h^2}{4} f_{i,j} \tag{4}$$

The method is as follows:

1. For $k = 0$, pick an initial guess $U^{(0)} = [u_{1,1}^{(0)}, \ldots, u_{m,m}^{(0)}]^T$.
2. Plug $U^{(0)}$ into (4) to obtain $U^{(1)}$.
3. Compute $\left\|u^{(k)} - u^{(k+1)}\right\|_2$. If $< \varepsilon$ then stop. Otherwise ...
4. Repeat steps (1)-(3) with $U^{(1)}$ instead of $U^{(0)}$.

Remark 2.1. It can be shown that the Jacobi iterative method will converge for *any* initial guess (but very slowly!). See page 70 in the text [Leveque] for Matlab code for the Jacobi method.

The Gauss-Seidel method is quite similar, but we take advantage of some of the updates in our matrix. Specifically, we follow these steps:

1. For $k = 0$, pick an initial guess $u^{(0)}$ and define a tolerance ε.
2. Obtain $u^{(1)}$ by plugging into

$$u_{ij}^{(k+1)} = \frac{1}{4}\left(u_{i-1,j}^{(k+1)} + u_{i+1,j}^{(k)} + u_{i,j-1}^{(k+1)} + u_{i,j+1}^{(k)}\right) - \frac{h^2}{4} f_{ij} \tag{5}$$

3. Compute the error $E = \left\|u^{(k+1)} - u^{(k)}\right\|$. If $\le \varepsilon$ then stop. Otherwise ...

4. Repeat steps (1)-(3) with $u^{(1)}$ instead of $u^{(0)}$.

Example 2.2. Consider the usual ODE $u''(x) = f(x), u(0) = \alpha, u(1) = \beta$. When we solve using the Finite Difference Method, we obtain a tridiagonal system, due to the relation

$$\frac{1}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) = f_i$$

Applying Jacobi, we have

$$u_i^{(k+1)} = \frac{1}{2}\left(u_{i-1}^{(k)} + u_{i+1}^{(k)} - h^2 f_i\right)$$

and applying Gauss-Seidel, we have

$$u_i^{(k+1)} = \frac{1}{2}\left(u_{i-1}^{(k+1)} + u_{i+1}^{(k)} - h^2 f_i\right)$$

This yields a system $Au = f$. The goal of this method will be to write $A = M - N$ for some choice of matrices M, N, so that $Mu = Nu + f$, which we will write as

$$Mu^{(k+1)} = Nu^{(k)} + f$$

The idea is to make M as simple as possible but still retain as much of the info from A as required. Specifically, for the Jacobi method, we take $M = \text{diag}(A)$. Then we have

$$M = -\frac{2}{h^2}I \quad , \quad N = M - A = \frac{-1}{h^2}\begin{bmatrix} 0 & 1 & & \\ 1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & 0 \end{bmatrix}$$

For Gauss-Seidel, however, we use $M = L(A)$ the lower-triangular part of A, so that

$$M = \begin{bmatrix} -2 & & & \\ 1 & -2 & & \\ & \ddots & \ddots & \\ & & 1 & -2 \end{bmatrix} \quad , \quad N = M - A = \begin{bmatrix} 0 & -1 & & \\ & \ddots & \ddots & \\ & & -1 & 0 \end{bmatrix}$$

In either case, we have $Mu^{(k+1)} = Nu^{(k)} + f$. Notice that M is invertible in both cases! Thus,

$$u^{(k+1)} = M^{-1}\left(Nu^{(k)} + f\right) = (\underbrace{M^{-1}N}_{:=G})u^{(k)} + \underbrace{M^{-1}f}_{:=C}$$

which we write as

$$u^{(k+1)} = Gu^{(k)} + C \tag{6}$$

Suppose u_T is the true solution of the original system, so $Au_T = f$. Then we take Equation (6) and take a limit as $k \to \infty$: $u_T = Gu_T + C$; i.e. u_T is the fixed point for Equation (6).

Will this method converge for all u_0? To answer this question, we define the error

$$e^{(k+1)} := u^{(k+1)} - u_T \equiv Ge^{(k)}$$

Iterating, we have $e^{(k)} = G^k e^{(0)}$, so for convergence we need $G^k \to 0$ in some sense. We know that $G^k \to 0$ if $\sigma(G) < 1$ (the spectral radius σ); if this holds, the we obtain convergence for any initial guess u_0! Note:

$$\sigma\left(G_{\text{Jacobi}}\right) = |\cos(\pi h)| < 1$$

for $h < 1$, and

$$\sigma\left(G_{\text{Gauss-Seidel}}\right) \approx 1 - \pi^2 h^2 + O(h^4)$$

if h is small enough (i.e. $\forall 0 < h < h_0$). Note that as $h \to 0$, $\sigma(G) \to 1$, so convergence rate gets slower, but the convergence is still guaranteed for any guess u_0.

2.2 Successive Over-Relaxation Method

Recall

$$u_i^{\text{GS}} = \frac{1}{2}\left(u_{i-1}^{(k+1)} + u_{i+1}^{(k)} - h^2 f_i\right)$$

is the $k+1$-th iterate for the Gauss-Seidel method. We use this to define

$$u_i^{(k+1)} = u_i^{(k)} + \omega\left(u_i^{\text{GS}} - u_i^{(k)}\right) \quad , \quad 0 < \omega < 2$$

which can be written as

$$u_i^{(k+1)} = \frac{\omega}{2}\left(u_{i-1}^{(k+1)} + u_{i+1}^{(k)} - h^2 f_i\right) + (1-\omega)u_i^{(k)}$$

Notice that $\omega = 1$ corresponds to the Gauss-Seidel method.

3 Initial Value Problems for ODEs

The goal of this section is to prepare for the study of elliptic and parabolic PDEs. We consider the ODE

$$\begin{aligned} u' + au &= f(t) \quad , \quad t > 0 \\ u(0) &= \gamma \end{aligned} \tag{7}$$

To solve this, we introduce the *integrating factor* e^{at} and multiply both sides of Equation (7) by e^{at} to get

$$u'\mathrm{e}^{at} + a\mathrm{e}^{at}u = f(t)\mathrm{e}^{at}$$

Then, we recognize the LHS as a derivative to write

$$\frac{d}{dt}\left(\mathrm{e}^{at}u\right) = \mathrm{e}^{at}f(t)$$

and then integrate both sides with respect to t to get

$$\mathrm{e}^{at}u = \mathrm{e}^{at}u \restriction_{t=0} + \int_0^t \mathrm{e}^{at}f(s)\,ds = \gamma + \int_0^t \mathrm{e}^{at}f(s)\,ds$$

This gives us the solution

$$u(t) = \mathrm{e}^{-at}\gamma + \int_0^t \exp(-a(t-s))f(s)\,ds \tag{8}$$

If $a > 0$, then we have the estimate

$$|u(t)| \leq |\gamma| + \int_0^t |f(s)|\,ds \quad , \quad t \geq 0$$

If we can obtain an estimate like this, then we say that $u' + au = f(t)$ is a *stable* ODE (i.e. the solution is bounded by initial data and the RHS). Why is this the definition of stability? To see why, consider the two problems

$$\begin{aligned} u_1' + au_1 &= f_1(t) \quad , \quad t > 0 \\ u_1(0) &= \gamma_1 \end{aligned}$$

and

$$\begin{aligned} u_2' + au_2 &= f_2(t) \quad , \quad t > 0 \\ u_2(0) &= \gamma_2 \end{aligned}$$

and assume both are stable. Then we can consider the (difference) problem

$$\begin{aligned} (u_1 - u_2)' + a(u_1 - u_2) &= f_1(t) - f_2(t) \quad , \quad t > 0 \\ (u_1 - u_2)(0) &= \gamma_1 - \gamma_2 \end{aligned}$$

and we know that we have the stability estimate

$$|u_1(t) - u_2(t)| \leq |\gamma_1 - \gamma_2| + \int_0^t |f_1(s) - f_2(s)|\,ds$$

This implies that small perturbations to initial data and/or the RHS do not produce large changes in the solution. Specifically, consider $f_2 = f_1 + \varepsilon_1$ and $\gamma_2 = \gamma_1 + \varepsilon_2$ for some small $\varepsilon_1, \varepsilon_2$.

3.1 Numerical Solution of ODEs

Consider the ODE in Equation (7), and apply the Forward Euler method, given by

$$U'(t_n) \approx \frac{U^n - U^{n-1}}{\Delta t}$$

where $U^n \approx u(t_n)$, $t_n = n\Delta t$, $\Delta = \frac{t}{N}$. The Forward Euler (FE) approximation of the ODE becomes

$$\begin{array}{rl} \frac{U^n - U^{n-1}}{\Delta t} + aU^{n-1} & = f\left(t_n, U^{n-1}\right) \quad , \quad n \geq 1 \\ U^0 & = \gamma \end{array} \tag{9}$$

To solve this, we apply the following steps:

1. For, $n = 1$ we use
$$\frac{U^1 - U^0}{\Delta t} + aU^0 = f\left(t, U^0\right)$$
to find U^1:
$$U^1 = \Delta t \cdot f\left(t, U^0\right) + (1 - a\Delta t)U^0$$
We write $U^1 \approx u(t_1) = u(\Delta t)$.

2. Next, set $U^0 = U^1$.

3. Repeat steps 1 and 2 to obtain U^2. Continue until U^N.

This method is an *explicit* time-stepping algorithm, and be coded via a loop over time steps $\Delta t = \frac{T}{N}$ to solve the ODE on $[0, T]$ with N subintervals.

Let's consider a specific model problem where $f(T, U) \equiv 0$, and investigate the accuracy and stability of the method.

$$\begin{array}{rl} \frac{U^n - U^{n-1}}{\Delta t} + aU^{n-1} & = 0 \quad , \quad n \geq 1 \\ U^0 & = \gamma \end{array} \tag{10}$$

We have

$$\begin{array}{rrl} U^n = (1 - a\Delta t)U^{n-1} & , \quad U^{n-1} = (1 - a\Delta t)U^{n-2} & , \cdots , \\ & U^2 = (1 - a\Delta t)U^1 & , \quad U^1 = (1 - a\Delta t)U^0 \end{array}$$

and this allows us to write a recursive formula

$$u(t_n) \approx U^n = (1 - a\Delta t)^n \cdot \gamma$$

For $t = t_n$ fixed, we have

$$\lim_{n\to\infty} U^n = \lim_{n\to\infty} \left(1 - a\frac{t}{n}\right)^n = \gamma \exp(-at)$$

so that $\Delta t \to 0 \;\Rightarrow\; U^n \to u(t_n)$. Q: How fast is this convergence? That is, what is the error $|U^n - u(t_n)|$ and how does it change with n?

Assume $a > 0$, and let's take Δt such that $1 \geq 1 - a\Delta t \geq -1$; i.e. $\Delta t \leq \frac{2}{a}$. Then

$$U^n = (1 - a\Delta t)^n \cdot \gamma \;\Rightarrow\; |U^n| \leq |\gamma|$$

However, if $\Delta t > \frac{2}{a}$, then U^n will grow with n; i.e. $|U^n| \to \infty$ as $n \to \infty$. That is, the scheme will *not* be stable.

3.1.1 Error Analysis

Observe that we can write the error as

$$\begin{aligned}
U^n - u(t_n) &= (1-a\Delta t)^n \gamma - \exp(-at_n)\gamma = (1-a\Delta t)^n \gamma - \exp(-an\Delta t)\gamma \\
&= (1-a\Delta t)^n \gamma - (\exp(-a\Delta t))^n \gamma \\
&= \gamma\left((1-a\Delta t)^n - (\exp(-a\Delta t))^n\right) \\
&= \gamma\left(1-a\Delta t - \exp(-a\Delta t)\right)\cdot\Big((1-a\Delta t)^{n-1} \\
&\qquad +(1-a\Delta t)^{n-2}\exp(-a\Delta t) + \cdots + (\exp(-a\Delta t))^{n-1}\Big) \\
&= (1-a\Delta t - \exp(-a\Delta t))\sum_{k=0}^{n-1}(1-a\Delta t)^j \exp(-(n-1-j)a\Delta t)
\end{aligned}$$

where we have applied the identity

$$a^n - b^n = (a-b)\cdot\left(a^{n-1} + ab^{n-2} + a^2b^{n-3} + \cdots + a^{n-2}b + b^{n-1}\right)$$

Thus,

$$\begin{aligned}
|U^n - u(t_n)| &\le |1-a\Delta t - \exp(-a\Delta t)|\cdot \\
&\qquad \cdot\left|\sum_{j=0}^{n-1}(1-a\Delta t)^j \exp(-(n-1-j)a\Delta t)\right|\cdot|\gamma| \\
&\le \frac{a^2(\Delta t)^2}{2}\sum_{j=0}^{n-1}|\gamma| = \frac{a^2(\Delta t)^2}{2}\cdot n|\gamma| \\
&= \frac{a^2\Delta t}{2}t_n|\gamma| = O(\Delta t)
\end{aligned}$$

since

$$|1-a\Delta t - \exp(-a\Delta t)| \le \frac{1}{2}a^2(\Delta t)^2$$

and each term in the sum above satisfies

$$\left|(1-a\Delta t)^j\right|\cdot|\exp(-(n-1-j)a\Delta t)| \le 1$$

Thus, $a\Delta t \le 2$ is the so-called *stability restriction.*

****** insert picture *****

Now, let's examine the same ODE in Equation (7) using the Backward Euler (BE) method (an *implicit* time scheme). That is, we write

$$\begin{aligned}
\tfrac{U^n - U^{n-1}}{\Delta t} + aU^n &= f\left(t_n, U^n\right) \quad , \quad n \ge 1 \\
U^0 &= \gamma
\end{aligned} \tag{11}$$

We solve this by performing the following steps:

1. Input U^0, $\Delta t = \frac{t}{n}$.

2. Solve the equation in (11) (which may be nonlinear, in which case we can use Newton's Method) for U^n.

3. Let $U^0 = U^n$ and find U^{n+1} by repeating Step 2.

In the model problem where $f \equiv 0$, then the BE method yields

$$U^n = \frac{1}{1 + a\Delta t} U^{n-1}$$

and iterating tells us

$$U^n = \left(\frac{1}{1 + a\Delta t}\right)^n \gamma$$

Supposing $a \geq 0$, we have $|U^n| \leq |\gamma|$.

Stability of BE: Notice that $|U^n| \leq |\gamma| \ \forall n \geq 0$ independent of the size of Δt and a.

For the following, we rewrite the ODE $u' + au = f(u,t)$ as $u' = g(u,t)$ and write

$$u^n = u^{n-1} + g(u^{n-1}, t)$$

In general, the Local Error (for time step Δt) is

$$\tau = |u^n - u\left((n+1)\Delta t\right)|$$

For the current model problem $u' + au = 0$ using Backwards Euler, the Local Erorr is

$$\tau = \left| \frac{1}{1 + a\Delta t} u^{n-1} - \exp(-a\Delta t) u^{n-1} \right|$$

where the second term comes from the solution to

$$u' + au = 0 \quad , \quad u\left((n-1)\Delta t\right) = u^{n-1} \ \Rightarrow \ u(n\Delta t) = \exp(-a\Delta t) u^{n-1}$$

We can then write

$$\begin{aligned} \tau &= \left| \frac{1}{1 + a\Delta t} - \exp(-a\Delta t) \right| \cdot \left| u^{n-1} \right| \\ &= \left| 1 - a\Delta t + a^2\Delta t^2 - \cdots - \left(1 - a\Delta t + \frac{a^2\Delta t^2}{2}\right) \right| \cdot \left| u^{n-1} \right| \\ &\leq C(\Delta t)^2 \cdot \left| u^{n-1} \right| \end{aligned}$$

which indicates a *first-order* scheme.

3.1.2 Crank-Nicholson scheme

Consider the scheme,

$$u^n = u^{n-1} + \Delta t \left(-a \left(\frac{u^{n-1} + u^n}{2} \right) + f\left(\frac{u^{n-1} + u^n}{2}, t + \frac{\Delta t}{2} \right) \right) \tag{12}$$

which, when applied to the model problem where $f \equiv 0$, yields

$$u^n = u^{n-1} - a\Delta t \left(\frac{u^{n-1} + u^n}{2} \right)$$

which can be written as

$$u^n = \frac{1 - \frac{a\Delta t}{2}}{1 + \frac{a\Delta t}{2}} u^{n-1}$$

We claim this is a *second-order* scheme in this case.

To approach this issue, we first do some analysis:

$$\begin{aligned} \left| \frac{1 - \frac{z}{2}}{1 + \frac{z}{2}} \mathrm{e}^{-z} \right| &= \left| \left(1 - \frac{z}{2}\right) - \left(1 + \frac{z}{2}\right) \cdot \left(1 - z + \frac{z^2}{2} + O(z^3)\right) \right| \\ &= \left| 1 - \frac{z}{2} - 1 - \frac{z}{2} + z + \frac{z^2}{2} + \frac{z^3}{4} + O(z^3) \right| \leq Cz^3 \end{aligned}$$

Thus, the Local Error must satisfy

$$\tau \leq Ca^3 (\Delta t)^3 \left| u^{n-1} \right|$$

which indicates a second-order scheme, locally. We now argue that this yields a *global* second-order scheme, as well. To investigate the global error, we observe that

$$|u^n - w^n| < C(\Delta t)^{1+\alpha} \left| u^{n-1} \right|$$

where w^n satisfies

$$(w^n)' = g(w^n, t) \quad , \quad w^n(n\Delta t) = u^{n-1}$$

We can look at the magnitude of the error that has propagated to time $T = n\Delta t$ and see that

$$e_1 = \left| w^1(T) - w^2(T) \right| \quad , \quad e_2 = \left| w^2(T) - w^3(T) \right| \quad , \quad \ldots$$

and since the w^is are *exact* solutions to the same equation but with different initial data, we need to obtain an estimate on the size of errors of propagation due to small changes in initial data. The following theorem addresses this and comes from a theoretic study of ODEs:

Theorem 3.1 (Stability of ODEs)**.** *Consider*

$$u_i' = g(u_i, t) \; i = 1, 2 \quad , \quad u_1(0) = \alpha_1 \quad , \quad u_2(0) = \alpha_2$$

Assume g is Lipschitz in u, i.e.

$$|g(z_1, t) - g(z_2, t)| \leq L |z_1 - z_2| \quad \forall z, t$$

Then the following estimate holds:

$$|u_1(t) - u_2(t)| \leq \exp(Lt) |u_1(0) - u_2(0)|$$

This is useful in our current analysis since we can write

$$\begin{aligned}
|u^n(T) - u(T)| &\leq |u^n(T) - w^n(T)| + \left|w^n(T) - w^{n-1}(T)\right| + \\
&\quad + \cdots + \left|w^2(T) - w^1(T)\right| \\
&\leq (\Delta t)^{1+\alpha} + c\exp\left(L(2\Delta t)\right)(\Delta t)^{1+\alpha}\left|u^{n-1}\right| + \\
&\quad + \cdots + c\exp\left(L(n\Delta t)\right)(\Delta t)^{1+\alpha}u_0 \\
&\leq (\Delta t)^{1+\alpha} c\exp\left(LT\right)\frac{T}{\Delta t}M \\
&\sim (\Delta t)^{\alpha}
\end{aligned}$$

4 Initial Value Problems for PDEs

4.1 Heat Equation

Consider the initial value problem

$$\begin{aligned}
u_t - \Delta u = 0 \quad &\text{in } \mathbb{R}^d \times \mathbb{R}_+ \\
u(\cdot, 0) = v \quad &\text{in } \mathbb{R}^d
\end{aligned} \tag{13}$$

The solution to (13) is given by

$$u(x,t) = (4\pi t)^{-\frac{d}{2}} \int_{\mathbb{R}^d} v(y) \exp\left(-\frac{|x-y|^2}{4t}\right) dy \tag{14}$$

and the so-called *fundamental solution* is given by

$$u(x,t) = (4\pi t)^{-\frac{d}{2}} \exp\left(-\frac{|x|^2}{4t}\right) \tag{15}$$

Theorem 4.1. *If v is a bounded continuous function on $\mathbb{R}^d$ then the function $u(x,t)$ defined by Equation (15) is a solution of the Heat equation (13) for all $t > 0$ and $u \to v$ as $t \to 0$.*

In a general setting, we consider a system like

$$\begin{aligned}
u_t - \Delta u = f \quad &\text{in } \Omega \times I := \Omega \times (0,T) \\
u = g \quad &\text{on } \Gamma \times I := \partial\Omega \times I \\
u(\cdot, 0) = v \quad &\text{in } \Omega
\end{aligned} \tag{16}$$

We also define the *parabolic boundary*

$$\Gamma_p = \left(\Gamma \times \overline{I}\right) \cup \left(\Omega \times \{t = 0\}\right)$$

which is the boundary of the parabolic cylinder, minus the lid, so to speak.

Theorem 4.2. *Let u be a smooth function and assume that $u_t - \Delta u \leq 0$ in $\Omega \times I$. Then u attains its maximum on the parabolic boundary.*

Theorem 4.3. *The solution of (16) satisfies*

$$\sup_{(x,t)\in\overline{\Omega}\times\overline{I}} |u(x,t)| \le \max\left\{ \max_{x\in\Gamma\times I} |g|, \max_{x\in\Omega} |v| \right\} + \frac{r^2}{2d} \max_{(x,t)\in\overline{\Omega}\times\overline{I}} |f|$$

where r is the radius of the smallest ball that contains Ω.

Theorem 4.4. *The initial value problem (13) has at* most *one solution that is bounded in $\mathbb{R}^d \times [0,T]$, where T is arbitrary.*

4.2 FDMs for Parabolic Problems

Consider the PDE

$$\begin{aligned} u_t &= u_{xx} \quad \text{in } \mathbb{R}\times\mathbb{R}_+ \\ u(\cdot,0) &= v \quad \text{in } \mathbb{R} \end{aligned} \tag{17}$$

for some given function v. Assuming that v is *smooth* and *bounded*, we have the following properties:

1. The problem (17) has a unique solution.
2. $\sup_x |u(x,t)| \le \sup_x |v(x)|$ for every $t \ge 0$.
3. The solution is given by

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left(-\frac{|x-y|^2}{4t}\right) v(y)\, dy$$

   which is just a special case of Equation (14) with $d = 1$.
4. $|u(\cdot,t)|_{\mathcal{C}^4} \le |v|_{\mathcal{C}^4}$, where $|v|_{\mathcal{C}^4} := \max_{|\alpha|\le 4} |D^\alpha v|$.

To apply the Finite Difference Method, we need to discretize in time and space. We introduce a grid of mesh points: $(x,t) = (x_j, t_n)$ where $x_j = jh$ with $j \in \mathbb{Z}$ and $t_n = n\Delta t$ with $n \in \mathbb{N}$.

The simplest FD scheme is given by

$$u_t(x_j,t_n) \approx \frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} \approx u_{xx}(x_j,t_n)$$

where $u_j^{n+1} \approx u(x_j, t_{n+1})$ and $u_j^n \approx u(x_j,t_n)$ for $j \in \mathbb{Z}$ and $n \in \mathbb{N}$. This is an (explicit) Forward Euler method, so we can write this as the system

$$\begin{aligned} u_j^{n+1} &= \left(E_{\Delta t} u^n\right)_j := \lambda u_{j-1}^n + (1-2\lambda) u_j^n + \lambda u_{j+1}^n \quad \text{for } n \in \mathbb{N}, j \in \mathbb{Z} \\ u_j^0 &= v_j \end{aligned} \tag{18}$$

where $\lambda := \frac{\Delta t}{h^2}$. This is a recursive formula, so we can write

$$u_j^n = \left(E_{\Delta t} u^{n-1}\right)_j = \left(E_{\Delta t} E_{\Delta t} u^{n-2}\right)_j = \cdots = \left(E_{\Delta t}^n u_0\right)_j$$

Assume that $0 \leq \lambda \leq \frac{1}{2}$ so that all of the coefficients above in the operator $E_{\Delta t}$ are ≥ 0. This allows us to write

$$\begin{aligned}
|u_j^{n+1}| = \left|\left(E_{\Delta t} u^n\right)_j\right| &\leq \lambda |u_{j-1}^n| + (1-2\lambda)|u_j^n| + \lambda |u_{j+1}^n| \\
&\leq \lambda \sup_{j\in\mathbb{Z}} |u_j^n| + (1-2\lambda) \sup_{j\in\mathbb{Z}} |u_j^n| + \lambda \sup_{j\in\mathbb{Z}} |u_j^n| \\
&= \sup_{j\in\mathbb{Z}} |u_j^n|
\end{aligned}$$

Iterating this procedure yields the bound $\sup_j |u_j^0| = \sup_j |v_j|$, and so

$$\sup_{j\in\mathbb{Z}} |u_j^n| \leq \sup_{j\in\mathbb{Z}} |v_j|$$

This is the *stability estimate* for the FD scheme.

Example 4.5. Let's consider what happens for an *unstable* scheme. Suppose $\lambda > \frac{1}{2}$ and $\Delta t > \frac{h^2}{2}$. Let $v_j = (-1)^j \varepsilon$ for $\varepsilon > 0$ small. Then $\sup_j |v_j| = \varepsilon$. Notice that

$$u_j^1 = \left(\lambda(-1)^{j-1} + (1-2\lambda)(-1)^j + \lambda(-1)^{j+1}\right)\varepsilon = (1-4\lambda)(-1)^j \varepsilon$$

Iterating, we obtain

$$u_j^n = (1-4\lambda)^n (-1)^j \varepsilon \;\Rightarrow\; \sup_{j\in\mathbb{Z}} |u_j^n| = (4\lambda - 1)^n \varepsilon \xrightarrow[n\to\infty]{} \infty$$

4.2.1 Error Analysis

Consider $\lambda \leq \frac{1}{2}$. We want to characterize $\max_j \left|u_j^n - \hat{u}_j^n\right|$, where $\hat{u}_j^n = u(x_j, t_n)$ the true solution. We write $\bar{x}_j \in (x_{j-1}, x_{j+1})$ and $\bar{t}_n \in (t_n, t_{n+1})$. Also, notice that $u_{tt} = (u_t)_{xx} = u_{xxxx}$. We start to define the local truncation error by using $\hat{u}$ in Equation (4.2)

$$\begin{aligned}
\tau_j^n := &\frac{\hat{u}_j^{n+1} - \hat{u}_j^n}{\Delta t} - \frac{\hat{u}_{j+1}^n - 2\hat{u}_j^n + \hat{u}_{j-1}^n}{h^2} - \underbrace{(u_t - u_{xx})}_{=0} \\
= &\left(\frac{\hat{u}_j^{n+1} - \hat{u}_j^n}{\Delta t} - \hat{u}_t(x_j, t_n)\right) - \left(\frac{\hat{u}_{j+1}^n - 2\hat{u}_j^n + \hat{u}_{j-1}^n}{h^2} - \hat{u}_{xx}(x_j, t_n)\right)
\end{aligned}$$

and then using the Taylor expansion around t_n

$$\begin{aligned}
\hat{u}_j^{n+1} = u(x_j, t_{n+1}) &= u(x_j, t_n) + u_t(x_j, t_n)\Delta t + \frac{\Delta t^2}{2} u_{tt}(x_j, \bar{t}_n) \\
&= \hat{u}_j^n + \Delta t u_t(x_j, t_n) + \frac{\Delta t^2}{2} u_{tt}(x_j, \bar{t}_n)
\end{aligned}$$

This allows us to rewrite the first term in the difference above as

$$\frac{\hat{u}_j^{n+1} - \hat{u}_j^n}{\Delta t} - u_t(x_j, t_n) = \frac{\Delta t}{2} u_{tt}(x_j, \bar{t}_n)$$

Likewise, we can use the Taylor expansion around x_j to write

$$\hat{u}_{j+1}^n = u(x_{j+1}, t_n) = \hat{u}_j^n + hu_x(x_j, t_n) + \frac{h^2}{2}u_{xx}(x_j, t_n) + \frac{h^3}{3!}u_{xxx}(x_j, t_n) + \frac{h^4}{4!}u_{xxxx}(\bar{x}_j, t_n)$$

and

$$\hat{u}_{j-1}^n = u(x_{j-1}, t_n) = \hat{u}_j^n + hu_x(x_j, t_n) + \frac{h^2}{2}u_{xx}(x_j, t_n) + \frac{h^3}{3!}u_{xxx}(x_j, t_n) + \frac{h^4}{4!}u_{xxxx}(\bar{x}_{j-1}, t_n)$$

Combining these in the second term in the difference equation above yields

$$\frac{\hat{u}_{j+1}^n - 2\hat{u}_j^n + \hat{u}_{j-1}^n}{h^2} - \hat{u}_{xx}(x_j, t_n) = \frac{h^2}{24}\left(u_{xxxx}(\bar{x}_j, t_n) + u_{xxxx}(\bar{x}_{j-1}, t_n)\right)$$

Finally, this gives us the local truncation error

$$\tau_j^n = \frac{1}{2}\Delta t u_{tt}(x_j, \bar{t}_n) - \frac{h^2}{24}\left(u_{xxxx}(\bar{x}_j, t_n) + u_{xxxx}(\bar{x}_{j-1}, t_n)\right) \tag{19}$$

Then,

$$\max_j \left|\tau_j^n\right| \leq \frac{\Delta t}{2}\max_{j,n}|u_{tt}| + \frac{h^2}{12}\left|u(\cdot, t_n)\right|_{C^4}$$

where

$$\left|u(\cdot, t_n)\right|_{C^4} = \max_{|\alpha|\leq 4}|D_x^\alpha u(\cdot, t_n)|$$

Recall that $|u(\cdot, t_n)|_{C^4} \leq |v|_{C^4}$ and $u_{tt} = u_{xxxx}$, so then

$$\max_j |\tau_j^n| \leq \frac{h^2}{3}|v|_{C^4}$$

Theorem 4.6. *Let $u^n \approx \hat{u}(x_j, t_n)$ for $j \in \mathbb{Z}$ for $\hat{u}$ a solution to the heat equation. Let $\lambda = \frac{\Delta t}{h^2} \leq \frac{1}{2}$. Then $\exists C$ such that*

$$\max_{j\in\mathbb{Z}}|u^n - \hat{u}^n| \leq Ct_n h^2|v|_{C^4}$$

where C is independent of Δ and h.

4.3 Mixed IVP

Consider the mixed initial value problem

$$\begin{aligned} u_t = u_{xx} \quad &\text{in } \Omega = (0,1),\ t > 0 \\ u(0,t) = u(1,t) = 0 \quad &\text{for } t > 0 \\ u(\cdot, 0) = v \end{aligned} \tag{20}$$

We write $(x_j, t_n) = (jh, n\Delta t)$ for $j = 0, \ldots, M$ and $n = 0, \ldots,$ and $U_j^n \approx u(x_j, t_n)$ with $U_0^n = U_M^n = 0$ for $n > 0$ and $U_j^0 = V_j = v(x_j)$ for $j = 0, \ldots, M$. Let $\lambda = \frac{\Delta t}{h^2}$. Consider the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}$$

for $j = 1, \ldots, M-1$ with $U_0^{n+1} = U_M^{n+1} = 0$. This can be rewritten as

$$\begin{aligned} U_j^{n+1} &= \lambda\left(U_{j-1}^n + U_{j+1}^n\right) + (1-2\lambda)U_j^n \quad \text{for } j = 1, \ldots, M-1 \\ U_0^{n+1} &= U_M^{n+1} = 0 \end{aligned} \tag{21}$$

We write the vector of unknowns $U^n = (U_0^n, U_1^n, \ldots, U_{M-1}^n, U_M^n)$ and follow the procedure outlined below:

1. For $n = 0$, $U^0 = (0, V_1, \ldots, V_{M-1}, 0)$.
2. Solve the system in (21) with $n = 0$

$$\begin{aligned} U_j^1 &= \lambda\left(U_{j-1}^0 + U_{j+1}^0\right) + (1-2\lambda)U_j^0 \quad \text{for } j = 1, \ldots, M-1 \\ U_0^1 &= U_M^1 = 0 \end{aligned}$$

   to get U^1.
3. Replace U^0 with U^1 and repeat step 1. Iterate.

For $\lambda \leq \frac{1}{2}$, we have the stability estimate

$$\max_{0 \leq j \leq M} \left|U_j^{n+1}\right| \leq \max_{0 \leq j \leq M} \left|U_j^n\right| \leq \max_{0 \leq j \leq M} |V_j|$$

For $\lambda > \frac{1}{2}$, consider $U_j^0 = (-1)^j \sin(\pi j h)$ for $j = 0, \ldots, M$. Then

$$U_j^n = (1 - 2\lambda - 2\lambda\cos(\pi h))^n U_j^0 \quad \text{for } j = 0, \ldots, M$$

Now, if h is sufficiently small, $2\lambda\cos(\pi h) \approx 2\lambda$ so

$$|1 - 2\lambda - 2\lambda\cos(\pi h)| \approx |4\lambda - 1| \geq \gamma > 1$$

and thus

$$\max_{0 \leq j \leq M} \left|U_j^n\right| \geq \gamma^n \max_{0 \leq j \leq M} \left|U_j^0\right| \xrightarrow[n \to \infty]{} \infty$$

which shows that the scheme is *unstable*.

Theorem 4.7. *Let U^n and u be the solutions to a 1D parabolic problem. Then*

$$\max_{0 \leq j \leq m} \left|U_j^n - u_j^n\right| \leq C t_n h^2 \max_{t \leq t_n} |u(\cdot, t)|_{C^4}$$

for some $t_n \geq 0$.

4.3.1 Implicit Scheme

Consider the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{h^2}$$

for $j = 1, \ldots, M-1$ and $n \geq 0$, and $U_0^{n+1} = U_M^{n+1} = 0$ for $n \geq 0$. This can be rewritten as

$$\begin{gathered}(1+2\lambda)U_j^{n+1} - \lambda\left(U_{j-1}^{n+1} + U_{j+1}^{n+1}\right) = U_j^n \quad \text{for } j = 1, \ldots, M-1 \\ U_0^{n+1} = U_M^{n+1} = 0\end{gathered} \tag{22}$$

Write the vector of unknowns $\bar{U}^{n+1} = (U_1^{n+1}, \ldots, U_{M-1}^{n+1})$, so that the system (22) can be written as

$$B\bar{U}^{n+1} = \bar{U}^n \quad \text{where } B = \begin{bmatrix} 1+2\lambda & -\lambda & & \\ -\lambda & 1+2\lambda & -\lambda & \\ & & \ddots & \\ & & -\lambda & 1+2\lambda \end{bmatrix}$$

Notice that B is diagonally dominant and symmetric, two desirable properties for solving linear systems.

This scheme is stable without *any* restrictions on Δt or h. Notice that

$$U_j^{n+1} = \frac{1}{1+2\lambda}U_j^n + \frac{\lambda}{1+2\lambda}\left(U_{j-1}^{n+1} + U_{j+1}^{n+1}\right)$$

and so

$$\max_{0\leq j\leq M}\left|U_j^{n+1}\right| \leq \frac{2\lambda}{1+2\lambda}\max_{0\leq j\leq M}\left|U_j^{n+1}\right| + \frac{1}{1+2\lambda}\max_{0\leq j\leq M}\left|U_j^n\right|$$

Rearranging tells us

$$\max_{0\leq j\leq M}\left|U_j^{n+1}\right| \leq \max_{0\leq j\leq M}\left|U_j^n\right| \leq \max_{0\leq j\leq M}|V_j|$$

for *any* $\lambda > 0$.

Consider the (local) truncation error, given by

$$\tau_j^n = \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} = O\left(\Delta t + h^2\right)$$

as $\Delta t, h \to 0$.

Theorem 4.8. *Let U^n and u^n be solutions of the heat equation. Then*

$$\max_j\left|U_j^n - u_j^n\right| \leq Ct_n\left(h^2 + \Delta t\right)\max_{t\leq t_n}|u(\cdot,t)|_{C^4} \quad \textit{for } t_n \geq 0$$

4.3.2 Crank-Nicholson Scheme

The Crank-Nicholson scheme, outlined below, is second-order accurate with respect to time. We solve

$$\begin{aligned}\frac{U_j^{n+1}-U_j^n}{\Delta t} &= \frac{U_{j+1}^n-2U_j^n+U_{j-1}^n}{2h^2}+\frac{U_{j+1}^{n+1}-2U_j^{n+1}+U_{j-1}^{n+1}}{2h^2}\ \forall j=1,\dots,M\\ U_0^n &= U_M^n=0\ \forall n\\ U_j^0 &= V_j=v(jh)\ \forall j=1,\dots,M\end{aligned} \tag{23}$$

Let $\lambda=\frac{\Delta t}{h^2}$. Then we can rewrite the system as

$$(1+\lambda)U_j^{n+1}-\frac{\lambda}{2}\left(U_{j-1}^{n+1}+U_{j+1}^{n+1}\right)=(1-\lambda)U_j^n+\frac{\lambda}{2}\left(U_{j-1}^n+U_{j+1}^n\right)$$

We write the vector of unknowns $\bar{U}^n=(U_1,U_2,\dots,U_{M-1})$, and so the system is $B\bar{U}^{n+1}=A\bar{U}^n$, where

$$B=\begin{bmatrix}1+\lambda & -\frac{\lambda}{2} & & \\ -\frac{\lambda}{2} & 1+\lambda & -\frac{\lambda}{2} & \\ & & \ddots & \\ & & -\frac{\lambda}{2} & 1+\lambda\end{bmatrix}\quad\text{and } A=\begin{bmatrix}1-\lambda & \frac{\lambda}{2} & & \\ \frac{\lambda}{2} & 1-\lambda & \frac{\lambda}{2} & \\ & & \ddots & \\ & & \frac{\lambda}{2} & 1-\lambda\end{bmatrix}$$

The stability result is

$$\max_j\left|U_j^{n+1}\right|\le\max_j\left|U_j^n\right|\quad\text{if }\lambda\le 1$$

For $\lambda>1$, we get

$$\max_j\left|U_j^{n+1}\right|\le\underbrace{(2\lambda-1)}_{>1}\max_j\left|U_j^n\right|$$

and iterating this inequality will yield a coefficient $\to\infty$, which is inconclusive (i.e. not *necessarily* unstable).

Recall $V=(V_0,\dots,V_M)^T$. Define the inner product

$$(V,W):=h\sum_{j=0}^{M}V_jW_j$$

and use this to define the norm

$$\|V\|_{2,h}=(V,V)^{1/2}=\left(h\sum_{j=0}^{M}V_j^2\right)^{1/2}$$

One can use this to prove the stability result for the Crank-Nicholson scheme:

$$\|U^n\|_{2,h}\le\|V\|_{2,h}\quad\forall\lambda>0 \tag{24}$$

Remark 4.9. Note that (24) holds for Backwards Euler for any $\lambda > 0$, but only holds for Forward Euler for $\lambda \leq \frac{1}{2}$.

Theorem 4.10. *The following estimate holds* $\forall \lambda > 0$*:*

$$\|U^n - u^n\|_{2,h} \leq Ct_n \left(h^2 + (\Delta t)^2\right) \cdot \max_{t \leq t_n} |u(\cdot, t)|_{C^6} \quad \textit{for } t_n \geq 0$$

Remark 4.11. For $\lambda \leq 1$, we also have $\|U^n - u^n\| = O(h^2 + \Delta t^2)$ in maximum norm.

4.4 FDMs for Hyperbolic Equations

4.4.1 First Order Scalar Equation

Consider the problem

$$\begin{aligned} u_t - au_x &= 0 \quad \text{in } \mathbb{R} \times \mathbb{R}_+ \\ u(\cdot, 0) &= v \quad \text{in } \mathbb{R} \end{aligned} \tag{25}$$

Recall: If $v \in C^1$ then the problem (25) admits the unique classical solution given by

$$u(x,t) = (E(t)v) = v(x + at)$$

where $x + at = \text{const.}$ are the characteristic lines. We have the estimates

$$\max |E(t)v| = \max |v| \quad \text{and } \|E(t)v\|_{\mathcal{L}^2} = \|v\|_{\mathcal{L}^2} \ \forall t \geq 0$$

which imply that $E(t)$ is stable in $\mathcal{L}^\infty$ and $\mathcal{L}^2$.

Define the grid $(x_j, t_n) = (jh, n\Delta t)$ and the approximations $U_j^n \approx u(x_j, t_n)$ for $j \in \mathbb{Z}$ and $n \in \mathbb{N}$. Assume, for now, that $a > 0$. Consider the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = a \frac{U_{j+1}^n - U_j^n}{h}$$

Let $\lambda = \frac{\Delta t}{h}$. Then we can write

$$U_j^{n+1} = a\lambda U_{j+1}^n + (1 - a\lambda) U_j^n$$

Stability: If $a\lambda \leq 1$, then

$$\max_j |U_j^n| \leq \max_j |V_j|$$

Theorem 4.12 (Convergence result)**.** *The estimate*

$$\max_j |U_j^n - u_j^n| \leq Ct_n h |v|_{C^2}$$

holds for $t_n \geq 0$.

Now, suppose $a < 0$; we use the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = a\frac{U_j^n - U_{j-1}^n}{h}$$

which can be written as

$$U_j^{n+1} = (1 + a\lambda)U_j^n - a\lambda U_{j-1}^n$$

for $0 < -a\lambda \leq 1$ with $|a\lambda| \leq 1$. These are *upwind schemes.*

Use the central-difference approximation

$$u_x \approx \frac{u(x+h,t) - u(x-h,t)}{2h}$$

to write

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{a}{2h}\left(U_{j+1}^n - U_{j-1}^n\right)$$

which can be written as

$$U_j^{n+1} = U_j^n + \frac{a\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right)$$

It can be shown that this method is *unstable* if $\lambda = \frac{\Delta t}{h} = \text{const}$. Note: this method is not used in practice.

We can make a minor modification to get a new scheme

$$U_j^{n+1} = \frac{1}{2}\left(U_{j+1}^n + U_{j-1}^n\right) + \frac{a\lambda}{2}\left(U_{j+1}^n - U_{j-1}^n\right)$$

which is known as the **Lax-Friedrichs Method**; it is stable when $|a\lambda| \leq 1$.

Compare this to the **Lax-Wendroff Method**

$$U_j^{n+1} = U_j^n + \frac{a\Delta t}{2h}\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{a^2\Delta t^2}{8h^2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n\right)$$

which is second-order, and stable when $|a\lambda| \leq 1$. To show second-order accuracy, we use $u_{tt} = a^2 u_{xx}$ and write

$$\begin{aligned} u(x, t+\Delta t) &= u(x,t) + \Delta t u_t(x,t) + \frac{\Delta t^2}{2}u_{tt}(t,x) + \cdots \\ &= u + \Delta t a u_x + \frac{\Delta t^2}{2}a^2 u_{xx} + \cdots \end{aligned}$$

Recall the *upwind methods*

$$\begin{aligned} U_j^{n=1} &= U_j^n - \frac{a\Delta t}{h}(U_j^n - U_{j-1}^n) \quad \text{for } a > 0 \\ U_j^{n=1} &= U_j^n - \frac{a\Delta t}{h}(U_{j+1}^n - U_j^n) \quad \text{for } a < 0 \end{aligned}$$

with the stability constraint $0 \leq \left|\frac{a\Delta t}{h}\right| \leq 1$.

The **Beam-Warning method** is second-order accurate and based on a one-sided approximation of the spatial derivatives. For $a > 0$, we write

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{2h}(3U_j^n - 4U_{j-1}^n - U_{j-2}^n) + \frac{a^2\Delta t^2}{2h^2}(U_j^n - 2U_{j-1}^n + U_{j-2}^n)$$

and for $a < 0$ we write

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{2h}(-3U_j^n + 4U_{j-1}^n - U_{j-2}^n) + \frac{a^2\Delta t^2}{2h^2}(U_j^n - 2U_{j-1}^n + U_{j-2}^n)$$

This scheme is stable for $0 \leq \left|\frac{a\Delta t}{h}\right| \leq 2$.

4.4.2 Characteristic tracing and interpolation

Because the solution is constant along characteristics, we wonder when

$$u(x_j, t_{n+1}) = u(x_j - a\Delta t, t_n)$$

When $0 < \frac{a\Delta t}{h} < 1$, then the new spatial point $x_j - a\Delta t$ will be between x_{j-1} and x_j. When $\frac{a\Delta t}{h} = 1$, then the new spatial point is exactly x_{j-1}; in this case, we can set $U_j^{n+1} = U_{j-1}^n$, and the method is actually *exact*.

Our goal now is to approximate

$$U_j^{n+1} \approx u(x_j - a\Delta t, t_n)$$

We use a linear interpolation between U_{j-1}^n and U_j^n

$$P(x) = U_j^n + (x - x_j)\left(\frac{U_j^n - U_{j-1}^n}{x_j - x_{j-1}}\right) \quad , \quad P(x_{j-1}) = U_{j-1}^n \quad , \quad P(x_j) = U_j^n$$

and approximate

$$U_j^{n+1} = P(x_j - a\Delta t) = U_j^n - \frac{a\Delta t}{h}(U_j^n - U_{j-1}^n)$$

which is valid for $0 < \frac{a\Delta t}{h} < 1$. Notice that

$$U_j^{n+1} = \left(1 - \frac{a\Delta t}{h}\right)U_j^n + \frac{a\Delta t}{h}U_{j-1}^n$$

so that our approximation is a *convex combination* of U_{j-1}^n, U_j^n.

We can also write

$$P(x) = U_{j-1}^n \frac{(x - x_j)(x - x_{j+1})}{(x_{j-1} - x_j)(x_{j-1} - x_{j+1})} + U_j^n \frac{(x - x_{j-1})(x - x_{j+1})}{(x_j - x_{j-1})(x_j - x_{j+1})} + U_{j+1}^n \frac{(x - x_{j-1})(x - x_j)}{(x_{j+1} - x_{j-1})(x_{j+1} - x_j)}$$

Notice that for constant h the denominators are ch^2.

4.5 The CFL Condition

The Courant-Friedrichs-Levy condition deals with the hyperbolic PDE (25) with solution $\eta(x-at)$ for intitial data $\eta(x)$. A scheme satisfies

$$U(x_j, t+\Delta t) = U(x_j - a\Delta t, t)$$

A *necessary* condition (in general) for any method developed for the advection equation: If U_j^{n+1} is computed based on walues $U_{j+p}^n, U_{j+p+1}^n, \ldots, U_{j+q}^n$ with $p \le q$ (note p, q may be negative), then we must have $x_{j+p} \le x_j - a\Delta t \le x_{j+q}$. When $x_j = jh$ for a uniform mesh, we can rearrange to write the condition as $-q \le \frac{a\Delta t}{h} \le -p$. We define $\nu := \frac{a\Delta t}{h}$ to be the *Courant* number.

The *domain of dependence* of the point (X, T) is $X - aT$ since $u(X, T) = \eta(X - aT)$. We write

$$D(X, T) = \{X - aT\}$$

In general, we will see that the solution at (X, T) will depend on the initial data at several points or over a whole interval. Recall that for the PDE $u_t = u_{xx}$ and $u(x, 0) = g(x)$ the solution is

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-y)}{4t}\right) dy$$

which depends on the whole real line.

Definition 4.13. *The domain of dependence of a grid point (x_j, t_n) is the set of grid points x_i at the initial time $t = 0$ with the property that the data U_i^0 at x_i has an effect on the solution U_j^n.*

Example 4.14. The Lax-Wendroff method uses a three-point stencil, so that descending each level in time yields the dependencies $j \to \{j-1, j, j+1\}$. So the solution U_j^n depends on the initial data at $x_{j-n}, \ldots, x_{j+n}$. If we refine the grid but keep $\frac{\Delta t}{h} = r$ fixed, then the numerical domain of dependence of the point (X, T) will fill the interval $[X - \frac{T}{r}, X + \frac{T}{r}]$. The Courant number tells us that we also need

$$X - \frac{T}{r} \le X - aT \le X + \frac{T}{r}$$

and solving this inequality tells us $|a| \le \frac{1}{r}$, or equivalently

$$\left|\frac{a\Delta t}{h}\right| \le 1 \tag{26}$$

This is a *necessary* condition for stability and convergence of the scheme.

The CFL condition: A numerical method can be convergent only if its numerical domain of dependence contains the true domain of dependence of the PDE, at least in the limit as $\Delta t, h \to 0$.

4.6 Modified Equations

Upwind scheme: $u_t + au_x = 0$ for $a > 0$.

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{h}\left(U_j^n - U_{j-1}^n\right)$$

Consider v

$$v(x, t+\Delta t) = v(x,t) - \frac{a\Delta t}{h}\left(v(x,t) - v(x-h,t)\right)$$

We do a Taylor expansion around (x,t):

$$v(x, t+\Delta t) = v(x,t) + \Delta t v_t + \frac{\Delta t^2}{2}v_{tt} + \frac{\Delta t^3}{3!}v_{ttt} + \cdots$$

and

$$v(x-h,t) = v(x,t) - hv_x + \frac{h^2}{2}v_{xx} - \frac{h^3}{3!}v_{xxx} + \cdots$$

Combining with the line above, we have

$$\Delta t v_t + \frac{\Delta t^2}{2}v_{tt} + \frac{\Delta t^3}{6}v_{ttt} + \cdots = -\frac{a\Delta t}{h}\left(hv_x + \frac{h^2}{2}v_{xx} - \frac{h^3}{6}v_{xxx} + \cdots\right)$$

Then

$$v_t + av_x = \frac{1}{2}\left(ahv_{xx} - \Delta t v_{tt}\right) + \frac{1}{6}\left(ah^2 v_{xxx} - \Delta t^2 v_{ttt}\right) + \cdots$$

for $\frac{\Delta t}{h}$ fixed. Drop all $O(\Delta t), O(\Delta t^2)$ terms to get the advection equation. Or, keep $O(\Delta t)$ terms but drop $O(\Delta t^2)$ terms to get

$$v_t + av_x = \frac{1}{2}\left(ahv_{xx} - \Delta t v_{tt}\right)$$

Differentiate this with respect to x and t to get

$$v_{tt} = -av_{xt} + \frac{1}{2}\left(ahv_{xxt} - \Delta t v_{ttt}\right)$$

and

$$v_{tx} = -av_{xx} + \frac{1}{2}\left(ahv_{xxx} - \Delta t v_{ttx}\right)$$

Then

$$v_t + av_x = \frac{1}{2}\left(ahv_{xx} - a^2\Delta t v_{xx}\right) + O(\Delta t^2)$$

Dropping the $O(\Delta t^2)$ terms yields

$$v_t + av_x = \frac{1}{2}ah\left(1 - \frac{a\Delta t}{h}\right)v_{xx}$$

an advection/diffusion equation. Examine the coefficient $\frac{1}{2}(ah - a^2\Delta t)$.

- If $a\Delta t = h$ then the upwind method is exact for the advection equation.
- If $0 < \frac{a\Delta t}{h} < 1$ then the diffusion coefficient is > 0.
- Otherwise, the problem is ill-posed.

Lax-Wendroff scheme: Follow a similar procedure to the one above to get

$$v_t + av_x = \frac{1}{6}ah^2\left(1 - \left(\frac{a\Delta t^2}{h}\right)\right)v_{xxx} = 0$$

The v_{xxx} is called the "dispersive" term. This shows the L-W method is 3rd-order accurate. In fact, L-W leads to dispersive behavior of the numerical solution, which implies an oscillation and shift in the location of the max. (See picture on page 219 in textbook.)

4.6.1 Higher Order Methods

Given data U_j for $j = 1, 2, \ldots, m$, we write $W_j \approx U_x(x_j)$ and define either

$$W_j = \left(\frac{U_j - U_{j-1}}{h}\right) = U_x + O(h) \tag{27}$$

or

$$W_j = \left(\frac{U_{j+1} - U_{j-1}}{2h}\right) = U_x + O(h^2) \tag{28}$$

Another way to derive (27) and (28) is to use interpolating polynomials $p_j(x)$ with $W_j = p_j'(x)$. Recall that interpolating polynomials satisfy $p_j(x_i) = U_i$. To regain (27) we construct a linear interpolating polynomial, where $p_1(x_j) = U_j$ and $p_1(x_{j-1}) = U_{j-1}$. Then we have

$$p_1(x) = \frac{x - x_j}{x_{j-1} - x_j}U_{j-1} + \frac{x - x_{j-1}}{x_j - x_{j-1}}U_j$$

and so for a uniform mesh

$$W_j = p_1'(x) = \frac{U_j}{h} - \frac{U_{j-1}}{h} = \frac{U_j - U_{j-1}}{h}$$

as above. For (28) we construct a quadratic interpolating polynomial, where $p_2(x_{j-1}) = U_{j-1}$, $p_2(x_j) = U_j$, $p_2(x_{j+1}) = U_{j+1}$, so we have

$$p_2(x) = \frac{(x - x_{j+1})(x - x_j)}{(x_{j+1} - x_j)(x_{j+1} - x_{j-1})}U_{j+1} + \frac{(x - x_j)(x - x_{j+1})}{(x_{j-1} - x_j)(x_{j-1} - x_{j+1})}U_{j-1} + \frac{(x - x_{j-1})(x - x_{j+1})}{(x_j - x_{j-1})(x_j - x_{j+1})}U_j$$

For example, interpolating with a polynomial of degree k will use the points $U_{j-2}, U_{j-1}, U_j, U_{j+1}, U_{j+2}$ and satisfy $p_4(x_i) = U_i$ for $i = j-2, \ldots, j+2$ and $W_j = p_4'(x)$. Then

$$U_x(x_j) \approx W_j = \frac{4}{3}\left(\frac{U_{j+1} - U_{j-1}}{2h}\right) - \frac{1}{3}\left(\frac{U_{j+2} - U_{j-2}}{4h}\right)$$

and this will be 4th order accurate. A 6th order accurate formula is given by

$$W_j = \frac{3}{2}\left(\frac{U_{j+1} - U_{j-1}}{2h}\right) - \frac{3}{5}\left(\frac{U_{j+2} - U_{j-2}}{4h}\right) + \frac{1}{10}\left(\frac{U_{j+3} - U_{j-3}}{6h}\right)$$

Differentiating this to get p_6'' will give an approximation to $U_x x(x_j)$, but it will only be 5th order accurate, for instance.

4.6.2 Mixed Equations and Fractional Step Methods

We look at these methods for advection-reaction equations

$$\begin{aligned} u_t + a u_x &= -\lambda u \\ u(x,0) &= \eta(x) \end{aligned} \tag{29}$$

This models, for example, transport of a radioactive material in a fluid, flowing at a constant speed a down a pipe. The exact solution to (29) is given by

$$u(x,t) = \exp(-\lambda t)\eta(x - at)$$

Unsplit methods: Extend the Upwind Methods for $a > 0$

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{h}\left(U_j^n - U_{j-1}^n\right) - \Delta t \lambda U_j^n$$

which is 1st order accurate for $0 < \frac{a\Delta t}{h} < 1$.

Lax-Wendroff method: Write

$$u(x, t+\Delta t) \approx u(x,t) + \Delta t u_t(x,t) + \frac{1}{2}(\Delta t)^2 u_{tt}(x,t)$$

and

$$u_{tt} = -a u_{xt} - \lambda u_t \quad , \quad u_{tx} = -a u_{xx} - \lambda u_x \quad , \quad u_{tt} = a^2 u_{xx} + 2\lambda a u_x + \lambda^2 u$$

and then plug these into the line above, using $u_{tx} = u_{xt}$ and $u_t = -a u_x - \lambda u$. From this, we derive

$$U_j^{n+1} = \left(1 - \lambda\Delta t + \frac{\lambda^2 \Delta t^2}{2}\right) U_j^n - \frac{a\Delta t}{2h}\left(1 - \frac{\lambda \Delta t}{2}\right)\left(U_{j+1}^n - U_{j-1}^n\right) + \frac{a^2 \Delta t^2}{2h^2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n\right)$$

Fractional Step Method. Consider Problem A $u_t^\star + a u_x^\star = 0$, and Problem B $u_t^{\star\star} = -\lambda u^{\star\star}$. Our goal is to find $U_i^{n+1} \approx u$, the solution to (29).

Step A: Define

$$u^\star \approx U_i^\star := U_i^n - \frac{a\Delta t}{\Delta x}\left(U_i^n - U_{i-1}^n\right)$$

Step B: Define

$$U_i^{n+1} = U_i^\star - \lambda \Delta t U_i^\star$$

Using Step A, this allows us to write

$$\begin{aligned} U_i^{n+1} &= (1-\lambda\Delta t)\, U_i^\star \\ &= (1-\lambda\Delta t)\left(U_i^n - \frac{a\Delta t}{\Delta x}\left(U_i^n - U_{i-1}^n\right)\right) \\ &= \underbrace{U_i^n - \lambda\Delta t U_i^n - \frac{a\Delta t}{\Delta x}(U_i^n - U_{i-1}^n)}_{\text{upwind scheme for } u_t + a u_x = -\lambda u} + \underbrace{\frac{\lambda\Delta t^2}{\Delta x}(U_i^n - U_{i-1}^n)}_{O(\Delta t^2)} \end{aligned}$$

where the right-hand term is $O(\Delta t^2)$ since $\frac{U_i^n - U_{i-1}^n}{\Delta x} \approx u_x = O(1)$. Another approach would be to take Step A to be the Lax-Wendroff method and Step B to be a second-order Runge-Kutta scheme.

In general, we have an equation $u_t = A(u) + B(u)$ and follow the procedure:

$$\begin{aligned} \textbf{Step A} &: U^\star = N_A(U^n, \Delta t) \\ \textbf{Step B} &: U^{n+1} = N_B(U^\star, \Delta t) \end{aligned}$$

where $N_A(U^n, \Delta t)$ is a one step method that solves $u_t = A(u)$ starting with initial data U^n, and $N_B(U^\star, \Delta t)$ is a one step method that solves $u_t = B(u)$ with initial data $U^\star$.

The advantages of this general procedure are

1. we can use a very different numerical approximation for Steps A and B, and
2. the decomposition of the problem into smaller problems is beneficial.

In general, the approximation will be only first-order accurate.

Example 4.15. Consider $u_t = Au + Bu$, and splitting means we have to solve $u_t = Au$ and $u_t = Bu$. Consider

$$\begin{aligned} \textbf{Step A} &: U^\star = N_A(U^n, \Delta t) = \mathrm{e}^{A\Delta t} U \\ \textbf{Step B} &: U^{n+1} = N_B(U^\star, \Delta t) = \mathrm{e}^{B\Delta t} U^\star \end{aligned}$$

so then

$$U^{n+1} = \mathrm{e}^{B\Delta t} U^\star = \mathrm{e}^{B\Delta t}\mathrm{e}^{A\Delta t} U^n$$

But the exact solution satisfies

$$u(t_{n+1}) = \mathrm{e}^{(A+B)\Delta t} u) t_n)$$

and for $\Delta t = t_{n+1} - t_n$, we have

$$\begin{aligned} \mathrm{e}^{(A+B)\Delta t} &= I + \Delta t(A+B) + \frac{1}{2}\Delta t^2 (A+B)^2 + \cdots \\ \mathrm{e}^{B\Delta t}\mathrm{e}^{A\Delta t} &= \left(I + \Delta t A + \frac{1}{2}\Delta t^2 A^2 + \cdots\right) \cdot \left(I + \Delta t B + \frac{1}{2}\Delta t^2 B^2 + \cdots\right) \\ &= I + \Delta t(A+B) + \frac{1}{2}\Delta t^2 (A^2 + 2AB + B^2) + \cdots \end{aligned}$$

so if A and B commute (e.g. are scalars) then the splitting idea is *exact*. If A and B do *not* commute then the scheme is $O(\Delta t)$.

Strang Splitting: The idea is similar, except we use

$$\begin{aligned}
&\textbf{Step A}: U^\star = N_A\left(U^n, \frac{\Delta t}{2}\right)\\
&\textbf{Step B}: U^{\star\star} = N_B(U^\star, \Delta t)\\
&\textbf{Step C}: U^{n+1} = N_A\left(U^{\star\star}, \frac{\Delta t}{2}\right)
\end{aligned}$$

where $u_t = A(u) \approx N(A)$ and $u_t = B(U) \approx N_B$. It can be shown that this is $O(\Delta t^2)$.

4.7 More Mixed Problems

1. Advection-reaction: $u_t + au_x = R(u)$
2. Reaction-diffusion: $u_t = ku_{xx} + R(u)$
3. Advection-diffusion (convection-diffusion): $u_t + au_x = ku_{xx}$
4. Advection-diffusion-reaction: $u_t + f(u)_x = ku_{xx} + R(u)$ for some nonlinear $f(\cdot)$

Let's focus on the Convection-Diffusion equation:

$$u_t + au_x = bu_{xx} \tag{30}$$

Let $y = x - at$ and set $w(t, y) = u(t, y + at)$. Then

$$w_t = u_t + au_x = bu_{xx} \text{ and } w_y = u_x \text{ and } w_{yy} = u_{xx}$$

so $w_t = bw_{yy}$. Thus, the solution to (30) travels with a speed a (convection) and is dissipated with strength b (diffusion).

Example 4.16 (Fokker-Planck equation). Consider a discrete process with states $X_i = \eta i$ where $i \in \mathbb{Z}$ and $\eta \in \mathbb{R}^+$ (like discretization in space). Transitions occur only between neighboring states at the discrete times $t_n = \tau n$ for $n = 0, 1, 2, \ldots$. Let p_i^n be the probability that $i \to i+1$ occurs in one time unit starting at time t_n. Let q_i^n be the corresponding probability for $i \to i - 1$. So the probability of staying at X_i is $1 - p_i^n - q_i^n$.

Let u_i^n be the probability density function at time t_n (i.e. u_i^n is the probability that the object is at X_i at time t_n). Then

$$u_i^{n+1} = p_{i-1}^n \cdot u_{i-1}^n + q_{i+1}^n \cdot u_{i+1}^n + (1 - p_i^n - q_i^n) \cdot u_i^n \tag{31}$$

This is known as the Chapman-Kolmogorov Equation. To derive a continuous Fokker-Planck (F-P) equation from this, we write $t_{n+1} = t_n + \tau$ and let $\tau, \eta \to 0$.

We rewrite (31) as

$$u_i^{n+1} - u_i^n = \frac{1}{2}\left(p_{i-1}^n - q_{i-1}^n\right) u_{i-1}^n - \frac{1}{2}\left(p_{i+1}^n - q_{i+1}^n\right) u_{i+1}^n$$
$$+ \frac{1}{2}\left[\left(p_{i-1}^n + q_{i-1}^n\right) u_{i-1}^n - 2\left(p_i^n + q_i^n\right) u_i^n + \left(p_{i+1}^n + q_{i+1}^n\right) u_{i+1}^n\right] \quad (32)$$

Thus,

$$\frac{u_i^{n+1} - u_i^n}{\tau} = \frac{1}{\eta}\left[\frac{1}{2}\eta \cdot \frac{p_{i-1}^n - q_{i-1}^n}{\tau} \cdot u_{i-1}^n - \frac{1}{2}\eta \cdot \frac{p_{i+1}^n - q_{i+1}^n}{\tau} \cdot u_{i+1}^n\right]$$
$$+ \frac{1}{2\eta^2}\left[\frac{p_{i-1}^n + q_{i-1}^n}{\tau} \cdot \eta^2 u_{i-1}^n - 2\frac{p_i^n + q_i^n}{\tau} \cdot \eta^2 u_i^n + \frac{p_{i+1}^n + q_{i+1}^n}{\tau} \cdot \eta^2 u_{i+1}^n\right] \quad (33)$$

Assume

$$\frac{p_i^n - q_i^n}{2\tau} \cdot \eta \xrightarrow[\eta,\tau\to 0]{} C(t_n, X_i) \text{ and } \frac{p_i^n + q_i^n}{2\tau} \cdot \eta^2 \xrightarrow[\eta,\tau\to 0]{} d(t_n, X_i)$$

Take the limit of (33) as $\tau, \eta \to 0$ to get the F-K equation

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial X}\left(C(t,X) \cdot u\right) + \frac{\partial^2}{\partial X^2}\left(d(t,X) \cdot u\right) \quad (34)$$

Since u is a probability density, it must satisfy $\int_{\mathbb{R}} u(t,x)\,dx = 1$ and $u \geq 0$ a.e.

To handle the Convection-Diffusion equation (30), we write

$$\frac{U_m^{n+1} - U_m^n}{\Delta t} + a\frac{U_{m+1}^n - U_{m-1}^n}{2h} = b\frac{U_{m+1}^n - 2U_m^n + U_{m-1}^n}{h^2}$$

which is a forward in time, central scheme. One can show that the stability requirement is

$$\frac{b\Delta t}{h^2} \leq \frac{1}{2}$$

Let $\mu = \frac{\Delta t}{h^2}$ and $\alpha = \frac{ha}{2b}$, so that $b\alpha\mu = \frac{a\Delta t}{2h}$. Then we can rewrite the scheme above as

$$U_m^{n+1} = (1 - 2b\mu)\,U_m^n + b\mu\,(1-\alpha)\,U_{m+1}^n + b\mu\,(1+\alpha)\,U_{m-1}^n$$

Recall the property of the solution to (30)

$$\sup_x |u(t,x)| \leq \sup_x |u(t',x)| \quad \text{for } t > t'$$

but notice the solutions to the discretized scheme above will have a similar property $\iff$ $\alpha \leq 1$. We already know $b > 0$ (diffusion coefficient), and assuming $a > 0$ then $\alpha = \frac{ha}{2b} > 0$ and $b\mu \leq \frac{1}{2}$. From these assumptions, we obtain the estimate

$$|U_m^{n+1}| \leq (1 - 2b\mu)\,|U_m^n| + b\mu(1-\alpha)|U_{m+1}^n| + b\mu(1+\alpha)|U_{m-1}^n|$$

so that

$$|U_m^{n+1}| \le (1-2b\mu)\max_m |U_m^n| + b\mu(1-\alpha)\max_m |U_m^n| + b\mu(1+\alpha)\max_m |U_m^n|$$

and collecting terms, we have

$$\max_m |U_m^{n+1}| \le \max_m |U_m^n|$$

for $\alpha \le 1 \sim h \le \frac{2b}{a}$. The quantity $\frac{a}{b}$ is known as the Reynolds number for fluid flow problems and the Péclet number for heat flow problems.

What happens if $h > \frac{2b}{a}$ or $\alpha < 1$? In general, the maximum estimate written above will not necessarily hold, as the following example shows.

Example 4.17. Let $U_m^0 = 1$ for $m \le 0$ and $U_m^0 = -1$ for $m > 0$. Putting these into the scheme above, we can find

$$U_0^1 = (1-2b\mu) - b\mu(1-\alpha) + b\mu(1+\alpha) = 1 + 2b\mu(\alpha - 1)$$

so that $U_0^1 > 1 = U_0^0 = \max$ for $\alpha > 1$, so the maximum principle does not hold. That is, the numerical solution will exhibit unphysical oscillations.

Compare this to an upwind scheme

$$\frac{U_m^{n+1} - U_m^n}{\Delta t} + a\frac{U_m^n - U_{m-1}^n}{h} = b\frac{U_{m+1}^n - 2U_m^n + U_{m-1}^n}{h^2}$$

which is 1st-order accurate. Defining α, μ as before, we have

$$U_m^{n+1} = (1 - 2b\mu(1+\alpha))U_m^n + b\mu U_{m+1}^n + b\mu(1+2\alpha)U_{m-1}^n$$

If $1 - 2b\mu(1+\alpha) \ge 0$, then $\max_m |U_m^{n+1}| \le \max_m |U_m^n|$. That is, we require $b\mu(1+\alpha) \le \frac{1}{2}$ to have a maximum principle, which is actually *less* restrictive than requiring $h \le \frac{2b}{a}$.

We rewrite the upwind scheme in the equivalent form

$$\frac{U_m^{n+1} - U_m^n}{\Delta t} + a\frac{U_{m+1}^n - U_{m-1}^n}{2h} = \left(b + \frac{ah}{2}\right)\frac{U_{m+1}^n - 2U_m^n + U_{m-1}^n}{h^2}$$

to recognize it as a central scheme with an extra term. That is, upwind scheme = central scheme + $\frac{ah}{2}u_{xx}$, where this extra term represents an "artificial viscosity".

4.8 Implicit-Explicit Methods

Here we explore some IMEX methods. Consider the problem $u_t = A(u) + B(u)$ where $A(u)$ is stiff and $B(u)$ is non-stiff. We have a first-order scheme

$$U^{n+1} = U^n + \Delta t\left(A(U^{n+1}) + B(U^n)\right)$$

and a second-order scheme

$$U^{n+1} = U^n + \frac{\Delta t}{2}\left(A(U^n) + A(U^{n+1}) + 3B(U^n) - B(U^{n-1})\right)$$

that is a multi-step method.

5 Analyses of Finite Difference Schemes

5.1 Fourier Analysis

Fourier Analysis is a helpful tool in the study of stability and well-posedness. For a function $u(x) \in \mathbb{R}$, its Fourier transform $\hat{u}(\omega)$ is defined by

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

for $\omega \in \mathbb{R}$. Note $\hat{u}(\omega) \in \mathbb{C}$. The Fourier Inversion Formula is given by

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

Example 5.1. Define

$$u(x) = \begin{cases} \mathrm{e}^{-x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

One can find that

$$\hat{u}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \exp(-i\omega x) \exp(-x)\, dx = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{1 + i\omega}$$

If v is a grid function defined on all integers m, then its Fourier transform is given by

$$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \exp(-im\xi) v_m$$

for $\xi \in [-\pi, \pi]$ with $\hat{v}(-\pi) = \hat{v}(\pi)$, and the inversion formula is given by

$$v_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \exp(im\xi) \hat{v}(\xi)\, d\xi$$

If the spacing between grid points is a fixed value h, then

$$\hat{v}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \exp(-imh\xi) v_m h$$

for $\xi \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right]$, and the inversion formula is

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

We have the following properties of the L^2 norm:

1. $\displaystyle\int_{-\infty}^{\infty} |u(x)|^2\, dx = \int_{-\infty}^{\infty} |\hat{u}(\omega)|\, d\omega$

2. $\|\hat{v}\|_h^2 = \displaystyle\int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 \, d\xi = \sum_{m=-\infty}^{\infty} |v_m|^2 h = \|v\|_h^2$

Property (2) is known as *Parseval's relation*, and we can prove it by observing that

$$\begin{aligned}
\|\hat{v}\|_h^2 &= \int_{-\pi/h}^{\pi/h} |\hat{v}(\xi)|^2 \, d\xi \\
&= \int_{-\pi/h}^{\pi/h} \overline{\hat{v}(\xi)} \cdot \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \exp(-imh\xi) v_m h \, d\xi \\
&= \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} \int_{-\pi/h}^{\pi/h} \exp(-imh\xi) \overline{\hat{v}(\xi)} \, d\xi \cdot v_m h \\
&= \sum_{m=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \exp(imh\xi) \hat{v}(\xi) \, d\xi \cdot v_m h \\
&= \sum_{m=-\infty}^{\infty} v_m \cdot v_m h = \|v\|_h^2
\end{aligned}$$

Example 5.2. Define

$$v_m = \begin{cases} 1 & \text{if } |x_m| < 1 \\ \frac{1}{2} & \text{if } |x_m| = 1 \\ 0 & \text{if } |x_m| > 1 \end{cases}$$

and let $h = \frac{1}{M}$. Then

$$\begin{aligned}
\hat{v}(\xi) &= \frac{h}{\sqrt{2\pi}} \Bigg(\underbrace{\frac{1}{2}}_{=v_m} \exp(\underbrace{iMh\xi}_{m=-M}) + \frac{1}{2} \exp(\underbrace{-iMh\xi}_{m=M}) \Bigg) \\
&\quad + \frac{h}{\sqrt{2\pi}} \sum_{m=-(M-1)}^{M-1} \exp(-imh\xi) \\
&= \frac{h}{\sqrt{2\pi}} \cos(\xi) + \frac{h}{\sqrt{2\pi}} \cdot \frac{\sin\left(M - \frac{1}{2}\right) h\xi}{\sin\left(\frac{h\xi}{2}\right)} \\
&= \frac{h}{\sqrt{2\pi}} \sin(\xi) \cot\left(\frac{h\xi}{2}\right)
\end{aligned}$$

We also notice that

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

and so we have the general property

$$\frac{\partial \hat{u}}{\partial x}(\omega) = i\omega \hat{u}(\omega)$$

Important fact: $u(x)$ has L^2-integrable derivatives of order up to $r \iff$

$$\int_{-\infty}^{\infty} (1+|\omega|^2)^r |\hat{u}(\omega)|^2 \, d\omega < \infty$$

Proving this amounts to showing

$$\int_{-\infty}^{\infty} \left| \frac{\partial^r u(x)}{\partial x^r} \right|^2 dx = \int_{-\infty}^{\infty} |\omega|^{2r} |\hat{u}(\omega)|^2 \, d\omega$$

Definition 5.3. *Define the space of functions H^r, for $r \geq 0$, to be set of all functions in $L^2(\mathbb{R})$ such that the norm*

$$\|u\|_{H^r} = \left(\int_{-\infty}^{\infty} (1+|\omega|^2)^r |\hat{u}(\omega)|^2 \, d\omega \right)^{1/2} < \infty$$

Remark 5.4. $H^0 \equiv L^2$. Also,

$$\|D^r u\|_0^2 = \int_{-\infty}^{\infty} \left| \frac{\partial^r u}{\partial x^r} \right|^2 dx = \int_{-\infty}^{\infty} |\omega|^{2r} |\hat{u}(\omega)|^2 \, d\omega$$

5.2 Von Neumann Analysis

This is based on Fourier Analysis, and gives necessary and sufficient conditions for the stability of FD schemes.

Consider the forward-time and backward-space scheme

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

Let $\lambda = \frac{\Delta t}{h}$ and write

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

By the Fourier inversion formula,

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

Plugging this into the scheme above,

$$v_m^{n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} \exp(imh\xi) \left((1 - a\lambda) + a\lambda \exp(-ih\xi) \right) \hat{v}^n(\xi) \, d\xi$$

Recall

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

From uniqueness, we get

$$\hat{v}^{n+1}(\xi) = \underbrace{\left((1-\lambda) + a\lambda \exp(-ih\xi)\right)}_{=:g(h\xi)} \hat{v}^n(\xi)$$

This function g is called the *amplification factor*, and

$$\hat{v}^{n+1}(\xi) = g(h\xi)\hat{v}^n(\xi) \Rightarrow \hat{v}^n(\xi) = (g(h\xi))^n \hat{v}^0(\xi)$$

Checking the L^2 norm, we have

$$\begin{aligned} h \sum_{m=-\infty}^{\infty} |v_m^n|^2 &= \int_{-\pi/h}^{\pi/h} |\hat{v}^n(\xi)|^2 \, d\xi \\ &= \int_{-\pi/h}^{\pi/h} |g(h\xi)|^{2n} |\hat{v}^0(\xi)|^2 \, d\xi \end{aligned}$$

By letting $\theta = h\xi$, we can write

$$g(\theta) = (1 - a\lambda) + a\lambda \exp(-i\theta) = (1 - a\lambda) + a\lambda(\cos\theta - i\sin\theta)$$

and then we want to use this to bound $|g(\theta)|^2$. We will use the identities

$$1 - \cos\theta = 2\sin^2\left(\frac{\theta}{2}\right) \quad \text{and} \quad \sin\theta = 2\sin\left(\frac{\theta}{2}\right) \cdot \cos\left(\frac{\theta}{2}\right)$$

to write

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

since $t = a + ib \Rightarrow |t|^2 = a^2 + b^2$. Thus,

$$|g(\theta)|^2 = 1 - 4a\lambda(1 - a\lambda)\sin^2\left(\frac{\theta}{2}\right) \quad \text{for } a > 0$$

and so $|g(\theta)| \leq 1$ if $0 \leq a\lambda \leq 1$. However,

$$|g(\theta)|^{2n} \xrightarrow[\Delta t \to 0]{n \to \infty} \infty \quad \text{if } a\lambda > 1 \Leftarrow |g(\theta)|^2 > 1$$

5.2.1 Stability Condition

Theorem 5.5. *A one-step finite difference scheme (first-order accurate with respect to time) with constant coefficients is stable in the stability region $L \iff \exists K$ (constant, independent of $\theta, \Delta t, h$) such that*

$$|g(\theta, \Delta t, h)| \leq 1 + \Delta t \cdot K$$

with $(\Delta t, h) \in L$.

If $g(\theta, \Delta t, h)$ is independent of $\Delta t, h$, then the stability condition can be replaced by $|g(\theta)| \leq 1$.

Example 5.6. Consider the scheme

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

so then

$$g(\theta) = 1 + a\lambda - a\lambda \mathrm{e}^{i\theta}$$

If $a > 0$ and $\lambda = \text{const.}$ then

$$|g|^2 = 1 + 4a\lambda(1 + a\lambda)\sin^2\left(\frac{\theta}{2}\right) > 1$$

and so the scheme is *unstable*.

If $a < 0$ then $|g|^2 \leq 1$ provided $-1 \leq a\lambda \leq 0$.

The stability theorem is due to Von Neumann, who first noticed the unstable schemes as in the above example, whence the term "Von Neumann analysis".

Example 5.7. Consider the scheme

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

Replace v_m^n by $g^n \mathrm{e}^{im\theta}$ to get

$$\frac{g^{n+1}\mathrm{e}^{im\theta} - g^n \mathrm{e}^{im\theta}}{\Delta t} + a\frac{g^n \mathrm{e}^{i(m+1)\theta} - g^n \mathrm{e}^{i(m-1)\theta}}{2h} = 0$$

which implies

$$g^n \mathrm{e}^{im\theta}\left(\frac{g-1}{\Delta t} + a\frac{\mathrm{e}^{i\theta} - \mathrm{e}^{-i\theta}}{2h}\right)$$

and solving yields

$$g = 1 - ia\lambda \sin\theta \;\Rightarrow\; |g|^2 = 1 + a^2\lambda^2 \sin^2\theta > 1$$

If $\theta \neq 0$ or $\theta \neq \pi$, $\frac{\Delta t}{h} = \pi \Rightarrow$ the scheme is unstable.

Example 5.8. Consider $u_t + au_x = 0$ and the Lax-Friedrichs scheme

$$\frac{U_m^{n+1} - \frac{1}{2}(U_{m+1}^n + U_{m-1}^n)}{\Delta t} + a\frac{U_{m+1}^n - U_{m-1}^n}{2h} - U_m^n = 0$$

Then

$$g(\theta, \Delta t, h) = \cos\theta - ia\lambda \sin\theta + \Delta t$$

so, supposing $|a\lambda| \leq 1$,

$$\begin{aligned} |g|^2 &= (\cos\theta + \Delta t)^2 + a^2\lambda^2 \sin^2\theta \\ &\leq 1 + 2\Delta t \cos\theta + \Delta t^2 \leq (1 + \Delta t)^2 \leq 1 + K\Delta t \end{aligned}$$

We will use to prove that we have stability when $|a\lambda| \leq 1$. (One can prove, with more difficulty, that $|a\lambda| > 1 \Rightarrow$ instability.) We know $|g(\theta, h, \Delta t)| \leq 1 + K\Delta t$ and

$$\|U^n\|_h^2 = \int_{-\pi/h}^{\pi/h} |g(h\xi, \Delta t, h)|^{2n} |\hat{U}^0(\xi)|^2 \, d\xi$$

and $|g(h\xi, \Delta t, h)| \leq 1 + K\Delta t$ for $(\Delta t, h) \in L$, the stability region, so then

$$\|U^n\|_h^2 \leq \int_{-\pi/h}^{\pi/h} (1 + K\Delta t)^{2n} |\hat{U}^0(\xi)|^2 \, d\xi = (1 + K\Delta t)^{2n} \|U^0\|_h^2$$

Now, $n \leq \frac{T}{\Delta t}$, so

$$(1 + K\Delta t)^n \leq (1 + K\Delta t)^{T/\Delta t} \leq \mathrm{e}^{KT} = \lim_{\Delta t \to 0} (1 + K\Delta t)^{T/\Delta t}$$

and thus

$$\|U^n\|_h^2 \leq \mathrm{e}^{2KT} \|U^0\|_h^2 = C\|U^0\|_h^2$$

which implies stability in the L^2 norm.

Next, suppose $|g(\theta, \Delta t, h)| \leq 1 + K\Delta t$ *cannot* be satisfied for $(\Delta t, h) \in L$ for any value of K. Then the scheme is unstable in L. This assumption says that for any $C > 0$, $\exists \theta \in [\theta_1, \theta_2]$ and $(\Delta t, h) \in L$ with $|g(\theta, \Delta t, h)| > 1 + C\Delta t$. We construct U_m^0 as

$$U_m^0(\xi) = \begin{cases} \sqrt{h(\theta_1 - \theta_2)^{-1}} & \text{if } \theta \in [\theta_1, \theta_2] \\ 0 & \text{otherwise} \end{cases}$$

and notice that U_m^0 has compact support and $|g(\theta, \Delta t, h)| > 1 + C\Delta t$ when $\theta \in [\theta_1, \theta_2]$. Also,

$$\|U^0\|_h^2 = \|\hat{U}^0\|_h^2 = \int_{-\pi/h}^{\pi/h} |\hat{U}^0(\xi)|^2 \, d\xi = \int_{\theta_1/h}^{\theta_2/h} \frac{h}{\theta_2 - \theta_1} \, d\xi = 1$$

and then

$$\begin{aligned}\|U^n\|_h^2 &= \int_{-\pi/h}^{\pi/h} |g(h\xi, \Delta t, h)|^{2n} |\hat{U}^0(\xi)|^2 \, d\xi \\ &= \int_{\theta_1/h}^{\theta_2/h} |g(h\xi, \Delta t, h)|^{2n} \frac{h}{\theta_2 - \theta_1} \, d\xi \\ &\geq (1 + C\Delta t)^{2n} \cdot 1 \geq \frac{1}{2} \mathrm{e}^{2TC}\end{aligned}$$

when n is close to $T/\Delta t$. Thus,

$$\|U^n\|_h^2 \geq \frac{1}{2} \mathrm{e}^{2TC} \|U^0\|_h^2 \;\Rightarrow\; \text{unstable}$$

Corollary 5.9. *If a scheme, as in Von Neumann Th, is modified so that the modifications result only in the addition of terms on the order $O(\Delta t)$ to the amplification factor (*uniformly *in ξ), then the modified scheme is stable $\iff$ the original scheme is stable.*

Proof. Since $|g| \leq 1 + K\Delta t$ for the original scheme, and (by assumption) $g' = g + O(\Delta t)$ for the modified scheme, then

$$|g'| \leq |g + O(\Delta t)| \leq |g| + |O(\Delta t)| \leq 1 + K\Delta t + C\Delta t$$

and so $|g'| \leq 1 + K'\Delta t|$. This works to prove both directions, actually. □

Theorem 5.10. *A consistent one-step scheme $u_t + au_x + bu = 0$ is stable $\iff$ it is stable for the equation with $b = 0$. Moreover, when $\Delta t = \lambda h$ and $\lambda = const.$, the stability condition on $g(h\xi, \Delta t, h)$ is $|g(\theta, 0, 0)| \leq 1$.*

Proof. Consider $u_t + au_x + bu = 0$. Since the scheme is consistent, the error of the approximation of bu will be proprotional to Δt. This implies that bu will contribute to g a term $\sim \Delta t$ (by the previous Lemma). It follows that setting $b = 0$ will not affect the stability of the scheme. We know

$$g(\theta, \Delta t, h) = g(\theta, 0, 0) + O9h) + O(\Delta t)$$

and $\Delta t = \lambda h$ or $h = \lambda^{-1} t$ so

$$g(\theta, \Delta t, h) = g(\theta, 0, 0) + O(\Delta t) \text{ for } \theta \in [-\pi, \pi]$$

which implies $O(\Delta t)$ is uniformly bounded with respect to θ. By the previous Corollary 5.9 we have $|g(\theta, \Delta t, h)| \leq 1 + K'\Delta t \sim |g(\theta, 0, 0)| \leq 1 + K\Delta t$ and we can let $\Delta t \to 0$ to remove the second term, since θ is independent of Δt. Thus, $|g(\theta, 0, 0)| \leq 1$. □

Example 5.11. Consider $u_t + au_x - u = 0$ with the Lax-Friedrichs scheme

$$\frac{U_m^{n+1} - \frac{1}{2}(U_{m+1}^n + U_{m-1}^n}{\Delta t} + a \frac{U_{m+1}^n - U_{m-1}^n}{2h} - U_m^n = 0$$

By Von Neumann Th,

$$\text{stability} \sim \frac{\overbrace{U_m^{n+1}}^{g_{n+1}\exp(im\theta)} - \frac{1}{2}(U_{m+1}^n + U_{m-1}^n)}{\Delta t} + a\frac{U_{m+1}^n - U_{m-1}^n}{2h} = 0$$

Since $g(\theta) = \cos\theta - ia\lambda\sin\theta$ then

$$|g|^2 = \cos^2\theta + \sin^2\theta + (a^2\lambda^2 - 1)\sin^2\theta$$

so $|g| \leq 1 \iff |a\lambda| \leq 1$. Thus, the Lax-Friedrichs scheme is stable $\iff$ $|a\lambda| \leq 1$.

Example 5.12. Consider $u_t + au_{xxx} = f$ and apply the Lax-Friedrichs scheme

$$\frac{U_m^{n+1} - \frac{1}{2}(U_{m+1}^n + U_{m-1}^n)}{\Delta t} + \frac{a}{2h^3}(U_{m+2}^n - 2U_{m+1}^n + 2U_{m-1}^n - U_{m-2}^n) = f_m^n$$

It can be shown that this scheme is consistent if $\frac{h^2}{\Delta t} \to 0$ as $h, \Delta t \to 0$. Then

$$g(\theta, \Delta t, h) = \cos\theta + \frac{4a\Delta t}{h^3} i\sin\theta\sin^2\left(\frac{\theta}{2}\right)$$

It can then be shown that $\frac{4|a|\Delta t}{h^3}$ has to be bounded to get $|g| \leq 1 + K\Delta t$ and achieve stability. This amounts to showing

$$\frac{4|a|}{h} \cdot \frac{h^2}{\Delta t} \to \infty \quad \text{if } \frac{h^2}{\Delta t} \to 0$$

Therefore, the scheme is not convergent for this problem. In general, don't expect one scheme to work for all problems just because it works with a particular one!

5.3 Stability conditions for variable coefficients

Consider $u_t + a(t,x)u_x = 0$ with the Lax-Friedrichs scheme $U_m^n \approx u(t_n, x_m)$ with $\lambda = \frac{\Delta t}{h}$:

$$U_m^{n+1} = \frac{1}{2}(U_{m+1}^n + U_{m-1}^n) - \frac{\lambda}{2}a(t_n, x_m)(U_{m+1}^n - U_{m-1}^n)$$

Here, the stability condition is $|a(t_n, x_m)|\lambda \leq 1$ for any point (t_n, x_m) in the domain of computation.

General procedure: "frozen coefficient"
Consider the frozen coefficient problem arising from the scheme by selecting (t, x). If each frozen coefficient problem is stable, then the variable coefficient problem is also stable.

Note: depending on $a(x,t)$ we could take our frozen coefficient to be, for example,

$$\frac{a(t_{n+1}, x_m) + a(t_n, x_m)}{2}$$

5.3.1 Stability of Lax-Wendroff and Crank-Nicholson

Consider $u_t + au_x = 0$ with the scheme

$$U_m^{n+1} = U_m^n - \frac{a\lambda}{2}(U_{m+1}^n - U_{m-1}^n) + \frac{a^2\lambda^2}{2}(U_{m+1}^n - 2U_m^n + U_{m-1}^n)$$

Write $U_m^n \to g^n \cdot \mathrm{e}^{im\theta}$, so

$$\begin{aligned} g^{n+1}\exp(im\theta) = g^n\exp(im\theta) &- \frac{a\lambda}{2}\left(\exp(i(m+1)\theta) - \exp(i(m-1)\theta)\right) \\ &+ \frac{a^2\lambda^2}{2}\left(\exp(i(m+1)\theta) - 2\exp(im\theta) + \exp(i(m-1)\theta)\right) \end{aligned}$$

and then we have the amplification factor

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

Accordingly,

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

Thus, $|g(\theta)| \le 1 \iff (1-a^2\lambda^2) \ge 0 \iff |a\lambda| \le 1$.

For Crank-Nicholson, we have

$$\frac{U_m^{n+1} - U_m^n}{\Delta t} + a\frac{\overbrace{U_{m+1}^{n+1} - U_{m-1}^{n+1}}^{\approx u_x(t_{n+1},x_m)} + \overbrace{U_{m+1}^n - U_{m-1}^n}^{\approx u_x(t_n,x_m)}}{4h} = 0$$

Then writing $U_m^n = g^n\exp(im\theta)$ yields

$$0 \quad = \quad \frac{1}{\Delta t} \quad *$$

and so $g - 1 + a\lambda\frac{g+1}{2}i\sin\theta = 0$, which implies

$$g(\theta) = \frac{1 - \frac{i}{2}a\lambda\sin\theta}{1 + \frac{i}{2}a\lambda\sin\theta} = \frac{z}{\bar{z}} \;\Rightarrow\; |g(\theta)| = 1$$

Therefore, the Crank-Nicholson scheme is stable for *any* $\lambda = \frac{\Delta t}{h}$, i.e. it is *unconditionally stable.*

6 Solution of Linear Systems

We have discussed

1. Direct numerical methods: Gaussian elimination
2. Basic iterative methods: Jacobi, Gauss-Seidel, Successive Over-Relaxation

To this list, we will add the method of steepest descent and the conjugate gradient (CG) method; both are iterative methods.

In general, we want to solve $Ax = b$ $(\star)$ for some symmetric, positive-definite (spd) matrix A. Define

$$F(y) := \frac{1}{2}(y - x, A(y - x))$$

wher we assume that x is the solution to $(\star)$ and $(\cdot, \dot{c})$ is the standard inner product on $\mathbb{R}^k$ (with A being $k \times k$). Since $F(y) \geq 0$ then $y = x$ is the unique solution to $(\star)$ and it is a minimum for $F(\cdot)$. Define

$$\begin{aligned} E(y) := F(y) - F(0) &= \frac{1}{2}(y - x, A(y - x)) - \frac{1}{2}(x, Ax) \\ &= \frac{1}{2}(y, Ay) - \frac{1}{2}(x, Ay) - \frac{1}{2}(y, Ax) \\ &= \frac{1}{2}(y, Ay) - (y, b) \end{aligned}$$

since $Ax = b$ and $A^T = A$. Now, $y = x$ is the unique minimizer for $E(y)$; that is, we have recasted the original formulation $(\star)$ as the minimization problem

$$\min_y E(y) = \min_y \frac{1}{2}(y, Ay) - (y, b)$$

We think of $E(y)$ as the "energy" of the system. We write $\nabla E(y) = Ay - b =: -r$, and r is called the *residual*. Since y is the unique minimizer for $E(y)$, then $\nabla E(y)$ points in the direction of the *steepest ascent*.

6.1 Method of steepest descent

1. Start from x^0
2. Let $x^{k+1} = x^k + \alpha^k r^k$, where $r^k = b - Ax^k$ and α^k is some parameter

We want to choose α^k such that $E(x^{k+1})$ is minimal. We write

$$\begin{aligned} E(x^k + \alpha^k r^k) = \frac{1}{2}(x^k, Ax^k) + \alpha^k(r^k, Ax^k) + \frac{1}{2}(\alpha^k)^2(r^k, Ar^k) \\ - (x^k, b) - \alpha^k(r^k, b) \\ = E(x^k) - \alpha^k \underbrace{(r^k, r^k)}_{=\|r^k\|^2} + \frac{1}{2}(\alpha^k)^2(r^k, Ar^k) \end{aligned}$$

and since we want $\frac{\partial E}{\partial \alpha^k} = 0$, then we must have

$$\alpha^k = \frac{(r^k, r^k)}{(r^k, Ar^k)} = \frac{\|r^k\|^2}{\|r^k\|_A^2}$$

So now we have

$$\begin{aligned} r^{k+1} &= b - Ax^{k+1} = b - A(x^k + \alpha^k r^k) = b - Ax^k - \alpha^k Ar^k \\ &= r^k - \alpha^k Ar^k \end{aligned}$$

Notice that $(r^{k+1}, r^k) = 0$. Thus,

$$E(x^{k+1}) = E(x^k) - \frac{1}{2}\frac{\|r^k\|^4}{\|r^k\|_A^2}$$

and so $E(x^k)$ will decrease as k increases until the residual is 0.

Recalling the definition of $E(y)$ above, we find

$$E(x^k) = \frac{1}{2}(x^k - x, A(x^k - x)) - F(0)$$

where $x^k = A^{-1}(b - r^k)$, so

$$E(x^k) = \frac{1}{2}(A^{-1}r^k, r^k) - F(0)$$

It follows that

$$\left(A^{-1}r^{k+1}, r^{k+1}\right) = \left(A^{-1}r^k, r^k\right) - \frac{\|r^k\|^4}{\|r^k\|_A^2}$$

Theorem 6.1. *If A is a positive-definite matrix for which $A^T A^{-1}$ is also positive-definite, then the steepest descent algorithm converges to the unique solution.*

Proof. First, recall that A p-d $\Rightarrow$ A^{-1} p-d; suppose $A^T A^{-1}$, as well. It can be shown that an inequality of the form

$$c_0(x, A^{-1}x) \leq (x, A^T A^{-1}x)$$

holds for some $c_0 > 0$ (use eigenvalues ***). Similarly, $c_1(x, Ax) \leq (x, x)$ for some $c_1 > 0$. Consider $(r^{k+1}, A^{-1}r^{k+1})$, with $r^0 = b - Ax^0$ and $r^{k+1} = r^k - \alpha_k Ar^k$. We apply the equation stated above before the theorem to write

$$\left(r^{k+1}, A^{-1}r^{k+1}\right) = \left(r^k, A^{-1}r^k\right) - \alpha_k(r^k, r^k) - \alpha_k(Ar^k, A^{-1}r^k) + \alpha_k^2(r^k, Ar^k)$$

Then,

$$\left(r^{k+1}, A^{-1}r^{k+1}\right) = \left(r^k, A^{-1}r^k\right) - \alpha_k\left(r^k, A^T A^{-1}r^k\right)$$

and so

$$c_1 \leq \frac{\|r^k\|^2}{(r^k, Ar^k)} = \alpha_k \;\Rightarrow\; \alpha_k \geq c_1$$

Likewise,

$$c_0\left(r^k, A^{-1}r^k\right) \leq \left(r^k, A^T A^{-1} r^k\right) \Rightarrow \left(r^{k+1}, A^{-1}r^{k+1}\right) \leq \left(r^k, A^{-1}r^k\right)(1 - c_0 c_1)$$

Since A^{-1} is p-d, then all terms above are positive, so $1 > 1 - c_0 c_1 \geq 0$. This allows us to conclude

$$\left(r^k, A^{-1}r^k\right) \leq \left(r^0, A^{-1}r^0\right)(1 - c_0 c_1)^k \xrightarrow[k\to\infty]{} 0$$

so the scheme converges. Since A^{-1} is p-d then $r^k \to 0$, and since $r^k = b - Ax^k$, then $x^k = A^{-1}(b - r^k)$ and so

$$\lim_{k\to\infty} A^{-1}b = \lim_{k\to\infty} x^k = x$$

□

Corollary 6.2. *A s-p-d ⇒ steepest descent converges.*

To speed up the convergence, use the conjugate gradient method.

6.2 Conjugate Gradient Method

The method can be written as

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k(r^k + \gamma_k(x^k - x^{k-1})) \\ p^k &= r^k + \gamma_k(x^k - x^{k-1}) = r^k + \gamma_k \alpha_{k-1} p^{k-1} \\ &= r^k + \beta_{k-1} p^{k-1} \quad \text{where } \beta_{k-1} = \gamma_k \alpha_{k-1} \end{aligned} \tag{35}$$

so $x^{k+1} = x^k + \alpha_k p^k$. It can be shown that $r^{k+1} = r^k - \alpha_k A p^k$ and $p^{k+1} = r^{k+1} + \beta_k p^k$.

Given p^k, the search direction is known, but $\alpha_k, \beta_k =$? We want to minimize $E(x^{k+1})$

$$E(x^{k+1}) = E(x^k) - \alpha_k(p^k, r^k) + \frac{\alpha_k^2}{2}(p^k, Ap^k)$$

so we set

$$\frac{\partial E}{\partial \alpha_k} = 0 \;\Rightarrow\; \alpha_k = \frac{(p^k, r^k)}{(p^k, Ap^k)} \quad k \geq 0$$

Then

$$E(x^{k+1}) = E(x^k) - \frac{1}{2}\frac{(p^k, r^k)^2}{(p^k, Ap^k)}$$

and setting $p^0 := r^0 \Rightarrow E(x^1) < E(x^0)$. Also, note $(p^k, r^{k+1}) = 0$ and

$$(p^{k+1}, r^{k+1}) = |r^{k+1}|^2 \quad \text{for } k \geq 0$$

and since $(p^0, r^0) = |r^0|^2$ we can say $(p^k, r^k) = |r^k|^2$ for *every* $k \geq 0$. Therefore,

$$\alpha_k = \frac{|r^k|^2}{(p^k, Ap^k)} \;\Rightarrow\; E(x^{k+1}) = E(x^k) - \frac{1}{2}\frac{|r^k|^4}{(p^k, Ap^k)}$$

We want to minimize the functional (p^k, Ap^k) where

$$(p^k, Ap^k) = (r^k, Ar^k) + 2\beta_{k-1}(r^k, Ap^{k-1}) + \beta_{k-1}^2(p^{k-1}, Ap^{k-1})$$

so we differentiate w.r.t. β and set $= 0$. We get

$$\beta_k = \frac{|r^{k+1}|^2}{|r^k|^2}$$

Finally, we can summarize the Conjugate Gradient method:

1. For $k = 0$, set $p^0 := r^0 = b - Ax^0$.
2. For $k \geq 0$, write $x^{k+1} = x^k + \alpha_k p^k$ where

$$\alpha_k = \frac{|r^k|^2}{(p^k, Ap^k)} \ , \ r^{k+1} = r^k - \alpha_k Ap^k \ , \ p^{k+1} = r^{k+1} + \beta_k p^k \ , \ \beta_k = \frac{|r^{k+1}|^2}{|r^k|^2}$$