CHAPTER 5

Finite difference methods for the heat equation

1. One space dimension

1.1. Semi-discretized system. We start our analysis of numerical methods for PDE's with finite difference methods for the heat equation. The heat equation defined on the entire real line only requires initial conditions. Typically we solve the heat equation on a finite domain with boundary conditions that can be of Dirichlet, Neumann, or mixed type. We shall choose a model problem with Dirichlet boundary conditions

$$\left\{\begin{array}{c} q_t = q_{xx} \\ q(x, t=0) = f(x) \\ q(x=0, t) = g_0(t),\ q(x=2\pi, t) = g_1(t) \end{array}\right. . \tag{1.1}$$

Physical units have been chosen to obtain the simple form shown above. A time honored scientific method is to reduce a complicated unknown problem to one whose solution we already know. We therefore ask whether it is possible to reduce (1.1) to a system of ODE's. The idea is to carry out the discretization of just one of the differential operators. We choose to do this for the spatial derivatives and we shall approximate ∂_x^2 by a finite difference expression.

First we define a computational grid $x_j = jh$, $h = 2\pi/(M+1)$, $t^n = nk$ with step size h,k in space and time. Define $Q_j(t)$ to be the restriction of $q(x,t)$ to $x = x_j$

$$Q_j(t) = q(x_j, t),\ j = 0, 1, \ldots, M+1 . \tag{1.2}$$

The indices $j = 1, 2, \ldots, M$ correspond to points in the interior of the computation domain. The indices $j = 0$ and $j = M+1$ correspond to points on the boundary of the computation domain where we can apply the Dirichlet boundary conditions to get

$$Q_0(t) = g_0(t),\ Q_{M+1}(t) = g_1(t) . \tag{1.3}$$

There are many possible finite difference approximations of ∂_{xx}. To show the general issues involved consider a simple centered approximation which is second order accurate

$$q_{xx}(x_j, t) \cong \frac{\delta_x^2}{h^2} Q_j(t) = \frac{Q_{j+1}(t) - 2Q_j(t) + Q_{j-1}(t)}{h^2} . \tag{1.4}$$

Using this in the heat equation at $x = x_j$ gives an ODE

$$\frac{dQ_j(t)}{dt} = \frac{Q_{j+1}(t) - 2Q_j(t) + Q_{j-1}(t)}{h^2} . \tag{1.5}$$

If we consider all interior points we obtain a system of ODE's which can be written in matrix form as

$$\frac{d}{dt}\mathbf{Q} = \frac{1}{h^2}A\mathbf{Q} + \mathbf{b} \tag{1.6}$$

with

$$\mathbf{Q} = \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_M \end{bmatrix}^T, \ \mathbf{b} = \begin{bmatrix} Q_0 & 0 & \cdots & Q_{M+1} \end{bmatrix}^T . \tag{1.7}$$

$$A = \begin{matrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{matrix} \tag{1.8}$$

We have carried our program of reducing the initial PDE to a system of ODE's. This method is also known as the *method of lines*, since we're solving the initial PDE on the $x = x_j$ lines. Of course there is an approximation involved. We have used a finite spatial step size h to obtain a finite sized ODE system whereas the initial PDE is equivalent to an infinite system of ODE's. We can now apply what we know from ODE's to solve (1.6). The boundary condition term does not play an essential role in the analysis of algorithms so we'll consider the special case $g_0(t) = g_1(t) = 0$ which leads to $\mathbf{b} = 0$.

1.1.1. *Forward Euler method.* Using the forward Euler method leads to

$$\mathbf{Q}^{n+1} = \mathbf{Q}^n + \frac{k}{h^2}A\mathbf{Q}^n = \left(I + \frac{k}{h^2}A\right)\mathbf{Q}^n . \tag{1.9}$$

A typical component from this equation reads

$$Q_j^{n+1} = Q_j^n + \sigma\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) \tag{1.10}$$

with $\sigma = k/h^2$. The same procedure could have been obtained by a Forward in Time, Centered in Space finite difference approximation of both derivatives appearing in the heat equation hence the name FTCS scheme. It is of $O(k, h^2)$, i.e. first order in time, second order in space. We need to choose a time step size k that will ensure stability of the method. We're looking at finite step sizes so the appropriate concept to use is absolute stability. From Euler's method for systems we know that the region of stability is

$$|1 + z| \leq 1 \tag{1.11}$$

where $z = k\lambda$ and λ is an eigenvalue of $\left(1/h^2\right)A$. The eigenvalues of a general matrix are difficult to determine in general. For the particular matrix here however we can use the fruitful analogy between discrete and continuum operators. The matrix A is an approximation of the second derivative operator ∂_x^2. The eigenvalues and eigenfunctions of the continuum operator are given by

$$\partial_x^2 e^{i\xi x} = -\xi^2 e^{i\xi x} . \tag{1.12}$$

We suspect therefore that the eigenvectors of A are discretizations of the continuum eigenfunction $e^{i\xi x}$. We therefore try a discretization of the eigenfunction $e^{i\xi x}$

$$\mathbf{W}_p = \begin{bmatrix} e^{iph} & e^{ip2h} & \cdots & e^{ipjh} & \cdots & e^{ipMh} \end{bmatrix} \tag{1.13}$$

Here p is a discrete wavenumber (an integer in this case because we've chosen the definition domain as $[0, 2\pi]$). We compute $A\mathbf{W}_p$ hoping to obtain an eigenrelationship. The j^{th} component of the result is

$$(A\mathbf{W}_p)_j = e^{ip(j-1)h} - 2e^{ipjh} + e^{ip(j+1)h} = 2\left[\cos ph - 1\right]\ e^{ipjh} \tag{1.14}$$

and we see that we obtain an unique scalar for an arbitrary component j. Our initial guess was correct and the eigenvalues of $(1/h^2)A$ are

$$\lambda_p = -\left(\frac{2}{h}\sin\frac{ph}{2}\right)^2 \ . \tag{1.15}$$

The stability region is given by

$$-1 \leq 1 - k\left(\frac{2}{h}\sin\frac{ph}{2}\right)^2 \leq 1 \tag{1.16}$$

which reduces to

$$k \leq \frac{h^2}{2}\frac{1}{\sin^2(ph/2)} \tag{1.17}$$

The most restrictive condition arise for high values of p, the fast Fourier modes. A condition that includes all possbile p values is

$$k \leq \frac{h^2}{2} \tag{1.18}$$

This condition is quite restrictive in practice since a halving of the spatial step size (to get better accuracy) imposes a time step four times smaller.

Let us derive stability bounds using Von Neumann analysis. We replace the grid function values in (1.10) with their Fourier representation to obtain

(1.19)
$$\int_{-\pi/h}^{\pi/h} \hat{Q}^{n+1}(\xi)e^{ijh\xi}d\xi = \int_{-\pi/h}^{\pi/h} \hat{Q}^n(\xi)e^{ijh\xi}d\xi + \sigma\left(\int_{-\pi/h}^{\pi/h} \hat{Q}^n(\xi)e^{i(j+1)h\xi}d\xi - \right.$$

$$\left. 2\int_{-\pi/h}^{\pi/h} \hat{Q}^n(\xi)e^{ijh\xi}d\xi + \int_{-\pi/h}^{\pi/h} \hat{Q}^n(\xi)e^{i(j-1)h\xi}d\xi\right) \tag{1.20}$$

Grouping together terms this can be rewritten as

$$\int_{-\pi/h}^{\pi/h} \left[G(\xi) - 1 - 2\sigma\left(\cos h\xi - 1\right)\right]\hat{Q}^n(\xi)e^{ijh\xi}d\xi = 0\ . \tag{1.21}$$

Since the relation has to be true for all ξ the coefficient of the exponential functions (which form a basis) must be zero. The only interesting possibility is

$$G = 1 + 4\frac{k}{h^2}\sin^2\frac{h\xi}{2} \tag{1.22}$$

Imposing the condition $|G| \leq 1$ leads to the same stability criterion as above, $k \leq h^2/2$. This application of von Neumann analysis was carried out in full detail. Typically we know that we shall arrive at the stage where a relation between Fourier coefficients is obtained and a number of intermediate steps can be short-circuited.

1.1.2. *Trapezoidal method.* Applying the trapezoidal method to (**??**) leads to following relation between componets

$$Q_j^{n+1} = Q_j^n + \frac{\sigma}{2}\left(Q_{j+1}^n - 2Q_j^n + Q_{j-1}^n\right) + \frac{\sigma}{2}\left(Q_{j+1}^{n+1} - 2Q_j^{n+1} + Q_{j-1}^{n+1}\right) \ . \tag{1.23}$$

The method is known as the *Crank-Nicolson scheme*. This is now an implicit relation which requires the solution of a tridiagonal system at each time step. This is more work than the explicit FTCS method but not prohibitively so. If the method allows large step sizes the overall work for attaining a specific precision could be less than for FTCS. The stability criterion for the trapezoidal method is

$$|1 - z| \geq 1 \tag{1.24}$$

which leads to

$$\left| 1 + k\left(\frac{2}{h}\sin\frac{ph}{2}\right)^2 \right| \geq 1 \ . \tag{1.25}$$

This relation is satisfied for all $k > 0$ so the Crank-Nicolson method is unconditionally stable. This means that the only restriction on step size comes from the accuracy that we wish to attain.

Von Neumann stability analysis leads to

$$G = 1 + \sigma\left(\cos\xi h - 1\right) + G\sigma\left(\cos\xi h - 1\right) \tag{1.26}$$

or

$$G = \frac{1 - 2\sin^2\frac{\xi h}{2}}{1 + 2\sin^2\frac{\xi h}{2}} \tag{1.27}$$

from which we see that $|G| \leq 1$ always so we again obtain that the Crank-Nicolson scheme is unconditionally stable.