Department of Mathematics – University of York

Numerical Methods for PDEs

Lecture Notes [1]

Spring 2017

Module coordinators:
Gustav Delius, Rooms G/110 and RCH/328, gustav.delius@york.ac.uk
Richard Southwell, Room RCH/333, richard.southwell@york.ac.uk

Contents

1 Introduction **2**

2 Parabolic partial differential equations **5**
2.1 Explicit finite-difference method for the heat equation 5
2.2 Stability 9
2.3 Implicit methods 12
2.3.1 Backward-difference method 12
2.3.2 Double-sweep algorithm 13
2.3.3 Richardson method 14
2.3.4 Crank-Nicolson method 16
2.4 Consistency, stability, and convergence 18
2.5 Non-homogeneous heat equation with non-homogeneous boundary conditions 21
2.6 Boundary conditions of other types 22
2.7 Variable coefficients 25
2.8 Nonlinear heat equation 29
2.9 Two-dimensional heat equation. 32

3 Elliptic partial differential equations **38**
3.1 Poisson equation 38
3.2 Equations with variable coefficients 44
3.3 Arbitrary (not rectangular) domains 45

4 Hyperbolic partial differential equations **47**
4.1 Wave equation 47
4.2 Hyperbolic systems of first-order partial differential equations 50

5 Appendix A **54**

[1] These lecture notes were written by Kostia Ilin and edited by Gustav Delius and Richard Southwell.

1 Introduction

The overall aim of this module is to show how computers can be used to solve various mathematical problems for partial differential equations (PDEs). This involves both theoretical and practical components. The theoretical part is an introduction to the most commonly used numerical methods for solving PDEs. Here we will discuss how and why these methods work. In the practical part, we will use R and C++ to demonstrate how the numerical methods discussed in the theoretical part can be implemented in practice.

We will discuss numerical methods for parabolic (the heat equation), elliptic (the Laplace and Poisson equations) and hyperbolic (the wave equation) PDEs.

There are of course many good textbooks on the subject. Three that we will refer to from time to time are:

1. RL Burden & JD Faires, *Numerical Analysis* (6th ed.), Brooks/Cole Publishing Company, 1997;
2. WF Ames, *Numerical Methods for Partial Differential Equations*, Academic Press, 1977;
3. WH Press, *Numerical Recipes: the Art of Scientific Computing*, CUP, 2007.

Below are some facts from Calculus that are used throughout this course.

Theorem 1.1 (Taylor's theorem for functions of one variable)**.** *Let $f \in C^{n+1}$ in the neighbourhood of the point x_0 (i.e. f is continuous and has continuous derivatives of all orders up to the $(n+1)$th order). Then, for all x in this neighbourhood,*

$$f(x) = T_n + R_n$$

where T_n in the nth Taylor polynomial

$$T_n = f(x_0) + (x - x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \cdots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0)$$

and R_n is the remainder term:

$$R_n = \frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi)$$

for some point ξ between x_0 and x

Example 1.1. Let us obtain the Taylor series expansion of $f(x) = \sin x$ about the point $x_0 = 0$. We have

$$f(0) = 0, \quad f'(0) = \cos x|_{x=0} = 1, \quad f''(0) = -\sin x|_{x=0} = 0,$$
$$f'''(0) = -\cos x|_{x=0} = -1, \quad \text{etc.}$$

Hence,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{n=1}^{\infty}(-1)^{n-1}\frac{x^{2n-1}}{(2n-1)!}.$$

If we restrict our attention to the nth Taylor polynomial for $\sin x$, then the remainder term R_n can be estimated using the fact that $f^{(n+1)}(\xi)$ is equal to either $\pm\sin x$ or $\pm\cos x$ depending on n. In both cases $|f^{(n+1)}(\xi)| \leq 1$. Hence, we obtain

$$|R_n| \leq \frac{|x-x_0|^{n+1}}{(n+1)!}.$$

Definition 1.2. Let $\lim_{x\to 0} g(x) = 0$ and $\lim_{x\to 0} f(x) = f_0$. If there exists a positive constant K such that

$$|f(x) - f_0| \le K|g(x)|,$$

at least when x is sufficiently close to zero, we write

$$f(x) = f_0 + O(g(x))$$

as $x \to 0$.

Note that in the above definition it is important that we specify the "as $x \to 0$". We are interested in the behaviour as x gets smaller, and we say something is $O(g(x))$ if it goes to zero at least as fast as $g(x)$. One could also use the Big Oh notation for other limits, in particular $x \to \infty$.

Example 1.2. The function $f(x) = \sin(x)/x$ converges to 1 as fast as x^2 converges to zero (as $x \to 0$). To show this, it suffices to consider the second Taylor polynomial for $\sin(x)$:

$$\sin(x) = x - \frac{x^3}{3!}\cos(\xi)$$

where ξ is some number between 0 and x. We have

$$\left|\frac{\sin x}{x} - 1\right| = \frac{|x|^2}{3!}|\cos(\xi)| \le \frac{x^2}{3!} = \frac{x^2}{6} \quad \Rightarrow \quad \frac{\sin x}{x} = 1 + O(x^2).$$

Here we used the fact that $|\cos(\xi)| \le 1$ for all ξ.

Lemma 1.3 (Properties of $O(x^n)$ as $x \to 0$)**.** *We have*

1. $O(x^n) + O(x^m) = O(x^l)$ *for* $n, m \ge 0$ *and* $l = \min\{n, m\}$.
2. $O(x^n)O(x^m) = O(x^{n+m})$ *for* $n, m \ge 0$.
3. $x^m O(x^n) = O(x^{n+m})$ *for* $n \ge 0$ *and* $n + m \ge 0$.

For example,

$$O(x^2) + O(x^3) = O(x^2), \quad O(x^2)O(x^3) = O(x^5), \quad x^{-2}O(x^3) = O(x).$$

Note that the first property holds for $x \to 0$ but if we were instead considering $x \to \infty$ the min would change to a max.

Theorem 1.4 (Taylor's theorem for functions of two variables)**.** *Suppose that $f(x, y)$ and all its partial derivatives of order less than or equal to $(n+1)$ are continuous in $D = \{(x, y) \mid a < x < b,\ c < y < d\}$, and let $(x_0, y_0) \in D$. For every $(x, y) \in D$, there exist ξ between x and x_0 and μ between y and y_0 such that*

$$f(x, y) = T_n(x, y) + R_n(x, y)$$

where

$$\begin{aligned} T_n(x,y) &= f(x_0,y_0) + \left[(x-x_0)\frac{\partial f}{\partial x}(x_0,y_0) + (y-y_0)\frac{\partial f}{\partial y}(x_0,y_0)\right] \\ &\quad + \left[\frac{(x-x_0)^2}{2}\frac{\partial^2 f}{\partial x^2}(x_0,y_0) + (x-x_0)(y-y_0)\frac{\partial^2 f}{\partial x \partial y}(x_0,y_0)\right. \\ &\quad \left. + \frac{(y-y_0)^2}{2}\frac{\partial^2 f}{\partial y^2}(x_0,y_0)\right] + ... \\ &\quad + \left[\frac{1}{n!}\sum_{j=0}^{n}\binom{n}{j}(x-x_0)^{n-j}(y-y_0)^j\frac{\partial^n f}{\partial x^{n-j}\partial y^j}(x_0,y_0)\right] \end{aligned} \tag{1.1}$$

and

$$R_n(x,y) = \frac{1}{(n+1)!}\sum_{j=0}^{n+1}\left(\begin{array}{c} n+1 \\ j \end{array}\right)(x-x_0)^{n+1-j}(y-y_0)^j\frac{\partial^{n+1}f}{\partial x^{n+1-j}\partial y^j}(\xi,\mu).$$

Here

$$\left(\begin{array}{c} n \\ j \end{array}\right) = \frac{n!}{j!(n-j)!}$$

are binomial coefficients. $T_n(x,y)$ is called the n-th Taylor polynomial in two variables and $R_n(x,y)$ is the remainder term.

2 Parabolic partial differential equations

2.1 Explicit finite-difference method for the heat equation

We will illustrate the finite-difference methods for parabolic PDEs with the heat, or diffusion, equation

$$\frac{\partial u}{\partial t}(x,t) = K\frac{\partial^2 u}{\partial x^2}(x,t), \quad 0 < x < L, \quad 0 < t < T, \tag{2.1}$$

subject to the Neumann boundary conditions

$$u(0,t) = u(L,t) = 0 \quad \text{for} \quad t \in (0,T), \tag{2.2}$$

and the initial condition

$$u(x,0) = u_0(x), \tag{2.3}$$

where $u_0(x)$ is a given function. In Eq. (2.1), K is a positive constant. In what follows, we assume that (i) the initial condition (2.3) is consistent with the boundary conditions (2.2) (i.e. $u_0(0) = u_0(L) = 0$), (ii) $u_0(x)$ is twice differentiable in x on $[0, L]$ and (iii) a unique solution of the initial boundary-value problem (2.1)–(2.3) exists.

First we choose integers N and M and define h and τ as

$$\tau = \frac{T}{M}, \quad h = \frac{L}{N}.$$

Then we define the grid points (or mesh points) (x_k, t_j), where $x_k = hk$ for $k = 0, 1, \ldots, N$ and $t_j = \tau j$ for $j = 0, 1, 2, \ldots, M$. The problem is to find numbers w_{kj} (for $k = 0, 1, \ldots, N$ and $j = 0, 1, 2, \ldots$) such that w_{kj} approximates the value of the exact solution $u(x,t)$ at the grid point (x_k, t_j).

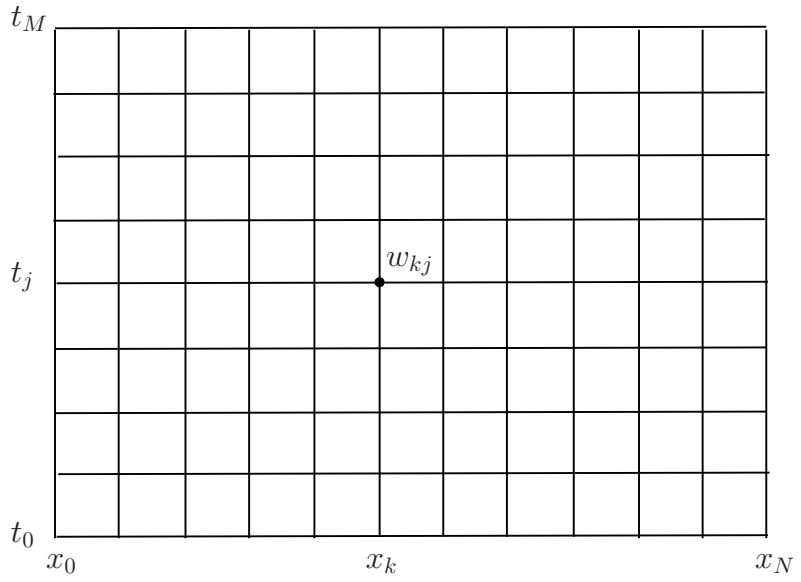

Figure 2.1:

To obtain a finite-difference method, we need to approximate the partial derivatives with respect to t and x at the grid points. By definition, the partial derivative of the function $u(x,t)$ with respect to t at point (x_k, t_j) is

$$\frac{\partial u}{\partial t}(x_k, t_j) = \lim_{\tau \to 0} \frac{u(x_k, t_j + \tau) - u(x_k, t_j)}{\tau}.$$

It is natural to expect that

$$\frac{\partial u}{\partial t}(x_k, t_j) \approx \frac{u(x_k, t_j + \tau) - u(x_k, t_j)}{\tau} \tag{2.4}$$

for sufficiently small τ. What is the error of this formula? To find this, we assume that u is sufficiently smooth (so that its first and second derivatives with respect to t are continuous in the interval $(0, T)$ for some $T > 0$) and write the first Taylor polynomial for $u(x_k, t_j + \tau)$:

$$u(x_k, t_j + \tau) = u(x_k, t_j) + \tau \frac{\partial u}{\partial t}(x_k, t_j) + \frac{\tau^2}{2} \frac{\partial^2 u}{\partial t^2}(x_k, \xi)$$

where ξ is between t_j and $t_j + \tau$. It follows that the error, which is called the **truncation error** and denoted by τ_{kj}, is given by

$$\tau_{kj} = \frac{\partial u}{\partial t}(x_k, t_j) - \frac{u(x_k, t_j + \tau) - u(x_k, t_j)}{\tau} = -\frac{\tau}{2} \frac{\partial^2 u}{\partial t^2}(x_k, \xi).$$

Thus, if $\frac{\partial^2 u}{\partial t^2}(x, t)$ is bounded for all x and t, then

$$\tau_{kj} = O(\tau).$$

Formula (2.4) with $\tau > 0$ is called the **forward-difference formula** for the first derivative. If in (2.4) we replace τ by $-\tau$, we obtain the formula

$$\frac{\partial u}{\partial t}(x_k, t_j) \approx \frac{u(x_k, t_j) - u(x_k, t_j - \tau)}{\tau}, \tag{2.5}$$

which is called the **backward-difference formula** (for the first derivative).

To derive a finite difference formula for $\partial^2 u(x_k, t_j)/\partial x^2$, we first write the Taylor series expansions of $u(x_k + h, t_j)$ and $u(x_k - h, t_j)$ at the point (x_k, t_j):

$$\begin{aligned}
u(x_k + h, t_j) = u(x_k, t_j) &+ h \frac{\partial u}{\partial x}(x_k, t_j) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_k, t_j) \\
&+ \frac{h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_k, t_j) + \frac{h^4}{24} \frac{\partial^4 u}{\partial x^4}(\xi_1, t_j), \\
u(x_k - h, t_j) = u(x_k, t_j) &- h \frac{\partial u}{\partial x}(x_k, t_j) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_k, t_j) \\
&- \frac{h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_k, t_j) + \frac{h^4}{24} \frac{\partial^4 u}{\partial x^4}(\xi_2, t_j)
\end{aligned}$$

for some ξ_1 between x_k and $x_k + h$ and some ξ_2 between $x_k - h$ and x_k. The sum of these equations gives us the formula

$$\begin{aligned}
u(x_k + h, t_j) + u(x_k - h, t_j) = 2u(x_k, t_j) + h^2 \frac{\partial^2 u}{\partial x^2}(x_k, t_j) \\
+ \frac{h^4}{24} \left(\frac{\partial^4 u}{\partial x^4}(\xi_1, t_j) + \frac{\partial^4 u}{\partial x^4}(\xi_2, t_j) \right).
\end{aligned} \tag{2.6}$$

If $\partial^4 u(x, t)/\partial x^4$ is continuous, we can write this in a more compact form. Indeed, the number

$$\frac{1}{2} \left(\frac{\partial^4 u}{\partial x^4}(\xi_1, t_j) + \frac{\partial^4 u}{\partial x^4}(\xi_2, t_j) \right)$$

is between the numbers $\frac{\partial^4 u}{\partial x^4}(\xi_1, t_j)$ and $\frac{\partial^4 u}{\partial x^4}(\xi_2, t_j)$. Therefore, by the intermediate value theorem, there is a number ξ between ξ_1 and ξ_2 such that $\frac{1}{2} \left(\frac{\partial^4 u}{\partial x^4}(\xi_1, t_j) + \frac{\partial^4 u}{\partial x^4}(\xi_2, t_j) \right) = \frac{\partial^4 u}{\partial x^4}(\xi, t_j)$. Therefore, we obtain

$$\frac{\partial^2 u}{\partial x^2}(x_k, t_j) = \frac{u(x_{k+1}, t_j) - 2u(x_k, t_j) + u(x_{k-1}, t_j)}{h^2} - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi, t_j), \tag{2.7}$$

where ξ is a number between $x_k - h$ and $x_k + h$. This is called the **central difference formula** for $u_x x$. When u_{xxxx} is bounded, then the truncation error in the central difference approximation is $O(h^2)$.

Now we are ready to obtain the finite-difference equations that approximate the heat equation (2.1). First, we introduce numbers w_{kj} for $k = 0, 1, \ldots, N$ and $j = 0, 1, 2, \ldots$ that approximate the exact solution $u(x, t)$ of Eq. (2.1) at the grid points (x_k, t_j):

$$w_{kj} \approx u(x_k, t_j).$$

Then, we use Eqs. (2.4) and (2.7) to approximate the corresponding derivatives in the heat equation. As a result we obtain the following difference equations

$$\frac{w_{k,j+1} - w_{kj}}{\tau} - K \frac{w_{k+1,j} - 2w_{kj} + w_{k-1,j}}{h^2} = 0, \tag{2.8}$$

for each $k = 1, 2, \ldots, N-1$ and $j = 0, 1, \ldots, M-1$. Equation (2.8) which approximates our PDE at point (x_k, t_j) uses approximations to the solution not only at this point but also at three neighbouring points (x_{k+1}, t_j), (x_{k-1}, t_j) and (x_k, t_{j+1}).

In approximating the heat equation by the difference equation (2.8) we introduced a truncation error of from approximating the t derivative of $O(\tau)$ and another truncation error from approximating the x derivative of $O(h^2)$. Thus it is natural to expect that the truncation error in the difference equation (2.8) is of order $O(\tau) + O(h^2)$. However we will meet more complicated finite-difference methods where the errors do not simply add, so we need a proper definition of what we mean by the truncation error of a finite-difference approximation.

Definition 2.1. Let us represent a PDE as $D\,u = 0$, where D is a differential operator, and the corresponding difference equation as $D_{f.d.}\,w = 0$, where $D_{f.d.}$ is the finite-difference operator that approximates D. Then the **local truncation error** τ_{kj} of the finite-difference approximation at grid point (x_k, t_j) is

$$\tau_{kj} = (D_{f.d.}\,u)_{kj},$$

i.e., it is equal to the value of the left hand side of the difference equation evaluated on the exact solution of the differential equation.

The local truncation error of the difference equation (2.8) is given by

$$\begin{aligned}
\tau_{kj} &= (D_{f.d.}\,u)_{kj} = \frac{u_{k,j+1} - u_{kj}}{\tau} - K \frac{u_{k+1,j} - 2u_{kj} - u_{k-1,j}}{h^2} \\
&= \frac{\partial u}{\partial t}(x_k, t_j) - K \frac{\partial^2 u}{\partial x^2}(x_k, t_j) + \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2}(x_k, \xi_1) - K \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi, t_j) \\
&= \frac{\tau}{2} \frac{\partial^2 u}{\partial t^2}(x_k, \xi_1) - K \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\xi, t_j) \\
&= O(\tau) + O(h^2) = O(\tau + h^2).
\end{aligned}$$

Here we used the notation $u_{kj} = u(x_k, t_j)$. So in this case the local truncation error is just the sum of the truncation error from the finite-difference approximation for the time derivative and the truncation error from the finite-difference approximation for the space derivative, as we expected.

Equation (2.8) can be written as

$$w_{k,j+1} = (1 - 2\gamma)\, w_{kj} + \gamma \left(w_{k+1,j} + w_{k-1,j} \right), \tag{2.9}$$

for $k = 1, \ldots, N-1$ and $j = 0, 1, \ldots, M-1$. In Eq. (2.9), $\gamma \equiv K\tau/h^2$. Since the initial condition $u(x, 0) = u_0(x)$ implies that $w_{k,0} = u_0(x_k)$ for each $k = 0, 1, \ldots, N$, these values can be used in Eq. (2.9) to find the value of $w_{k,1}$ for each $k = 1, 2, \ldots, N-1$. The boundary conditions $u(0, t) = u(l, t) = 0$

imply that $w_{0,1} = w_{N,1} = 0$. Now we know $w_{k,1}$ for each $k = 0, 1, \ldots, N$. Then the same procedure is applied to find $w_{k,2}$, $w_{k,3}$, etc.

The method described above is called the **forward-difference method**. It can also be written in the matrix form

$$\mathbf{w}^{(j)} = A\mathbf{w}^{(j-1)} \quad \text{for} \quad j = 1, 2, \ldots, \tag{2.10}$$

where

$$A = \begin{bmatrix} 1-2\gamma & \gamma & 0 & \ldots & \ldots & 0 \\ \gamma & 1-2\gamma & \gamma & \ddots & & \vdots \\ 0 & \gamma & 1-2\gamma & \gamma & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \gamma \\ 0 & \ldots & \ldots & 0 & \gamma & 1-2\gamma \end{bmatrix}, \quad \mathbf{w}^{(j)} = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ w_{N-1,j} \end{bmatrix}. \tag{2.11}$$

The forward-difference method for the heat equation is an example of an **explicit finite-difference method**. It is explicit because we do not need to solve any equations, we just use the explicit formula (2.10).

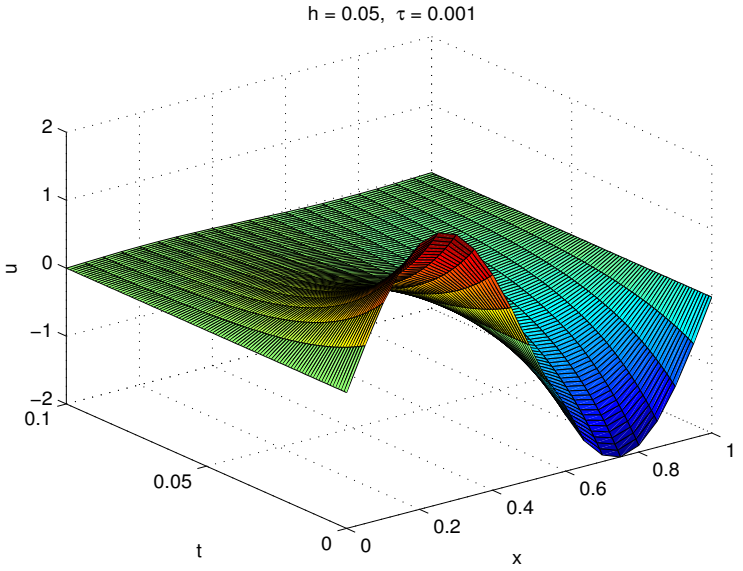

Figure 2.2: Solution of the forward-difference approximation of the heat equation (2.1) with boundary condition (2.2) and initial condition $u(x, 0) = 2\sin(2\pi x)$ for $K = 1$ and $t \in [0, 0.1]$. Step sizes are $h = 0.05$ and $\tau = 0.001$.

A surface plot of the solution $u(x, t)$ of problem (2.1)–(2.3) for $L = 1$, $T = 0.1$ and $u_0(x) = 2\sin(2\pi x)$, obtained with the help of the forward-difference method, is shown in Figure 2.2. So, the method works!

Will it always work? It turns out that it doesn't work if the time step τ is not small enough. For example, if we solve the same problem on a grid with a bigger τ, we get what is shown in Figure 2.3. So, in this case, the forward-difference method doesn't work properly. The reason for this is that the finite-difference scheme becomes **unstable** when τ is not sufficiently small.

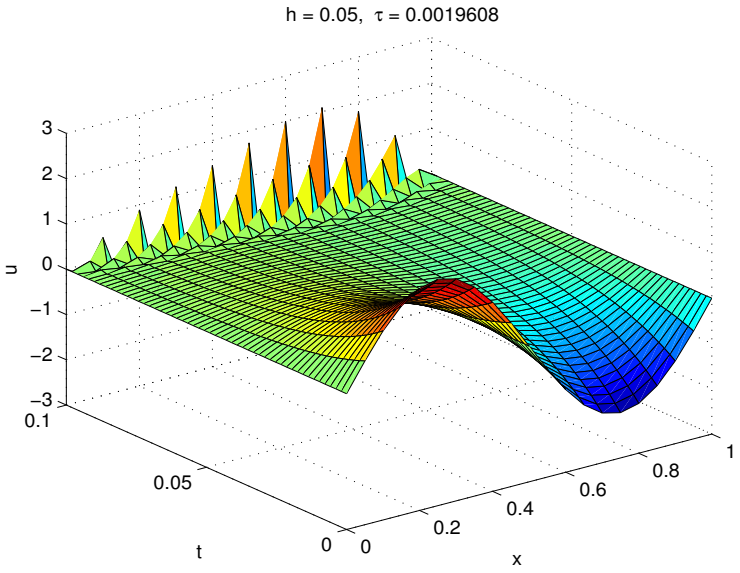

Figure 2.3: Solution of the forward-difference approximation for the same equation as in Figure 2.2 but with step sizes are $h = 0.05$ and $\tau = 0.1/51$.

2.2 Stability

If there are errors $z_{10}, z_{20}, \ldots, z_{N-1,0}$ in the initial data $w_{10}, w_{20}, \ldots, w_{N-1,0}$ (or at any particular step, the choice of the initial step is simply for convenience), the errors propagates to $w_{11}, w_{21}, \ldots, w_{N-1,1}$, then to $w_{12}, w_{22}, \ldots, w_{N-1,2}$, etc. If the errors grow with each time step, then the difference method in unstable. If they do not grow, it is stable. How to find out whether a finite-difference method is stable?

Let $\mathbf{z}^{(0)} = (z_1^{(0)}, z_2^{(0)}, \ldots, z_{N-1}^{(0)})^T$ be the initial error (or perturbation). Then it follows from (2.10) that

$$\tilde{\mathbf{w}}^{(1)} = A\tilde{\mathbf{w}}^{(0)} = A\left(\mathbf{w}^{(0)} + \mathbf{z}^{(0)}\right) = A\mathbf{w}^{(0)} + A\mathbf{z}^{(0)},$$

i.e.

$$\mathbf{z}^{(1)} = \tilde{\mathbf{w}}^{(1)} - \mathbf{w}^{(1)} = A\mathbf{z}^{(0)}.$$

At the n-th time step, the error in $\tilde{\mathbf{w}}^{(n)}$ due to $\mathbf{z}^{(0)}$ is $\mathbf{z}^{(n)} = A^n\mathbf{z}^{(0)}$. The method is stable if these errors do not grow as n increases, i.e. if and only if for any initial error $\mathbf{z}^{(0)}$ we have $\|A^n\mathbf{z}^{(0)}\| \le \|\mathbf{z}^{(0)}\|$ for all n or, equivalently, $\|A\mathbf{z}^{(0)}\| \le \|\mathbf{z}^{(0)}\|$ (here $\|\cdot\|$ is any vector norm). This, in turn, is equivalent to the condition that magnitudes of all eigenvalues of A are equal to or smaller than 1, i.e.

$$|\lambda_i| \le 1 \quad \text{for} \quad i = 1, 2, \ldots, N-1.$$

So, to solve the stability problem we need to calculate the eigenvalues of A.

Calculating the eigenvalues of the matrix in eq.(2.11) analytically is a bit hard. Instead, we will consider the problem with the boundary conditions (2.2) replaced by periodic boundary conditions

$$u(x, 0) = u(x, L). \tag{2.12}$$

In most cases the stability will not be affected much by this change in boundary condition. Physically the heat equation with this periodic boundary condition would model the heat of a circular rod, where heat flowing out of the right end flows back in at the left end.

The periodic boundary condition implies that at the boundary point $k = N$ we have the same value as at $k = 0$, but this value is no longer fixed. Instead we have to use equation (2.9) also for the boundary point $k = N$, with the convention that $k = N + 1$ is identified with $k = 1$. Therefore eq. (2.11) gets replaced by

$$A = \begin{bmatrix} 1-2\gamma & \gamma & 0 & \dots & 0 & \gamma \\ \gamma & 1-2\gamma & \gamma & \ddots & \vdots & 0 \\ 0 & \gamma & 1-2\gamma & \gamma & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & \ddots & \gamma \\ \gamma & 0 & \dots & 0 & \gamma & 1-2\gamma \end{bmatrix}, \quad \mathbf{w}^{(j)} = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ w_{N,j} \end{bmatrix}. \tag{2.13}$$

Luckily this matrix is a circulant matrix, where each row is equal to the row above, just circularly shifted to the right by one. The eigenvectors of such a matrix are of the form

$$\mathbf{v} = (1, z, z^2, \dots, z^{N-1})$$

where z is any of the N-th roots of unity, $z = \exp(2\pi i n)$ for $n = 1, \dots, N$. This knowledge of the eigenvectors of course makes it simple to calculate the eigenvalues of the matrix A simply by acting with A on each of the eigenvectors.

Because the matrix for periodic boundary conditions differs from the matrix for Neumann boundary conditions only in two places, it is plausible that the eigenvalues will also not differ much between the two cases.

We can also do the stability analysis with periodic boundary conditions without first writing down the matrix A but instead directly solving the difference equations. We refer to this approach as the **Fourier method**. Let w_{kj} be the exact solution of the difference equation (2.8)

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} - K\frac{w_{k+1,j} - 2w_{k,j} + w_{k-1,j}}{h^2} = 0 \tag{2.14}$$

(the forward-difference method). The initial data

$$z_{k,0} = \tilde{w}_{k,0} - w_{k,0} = \tilde{w}_{k,0} - u_0(x_k) \quad \text{for} \quad k = 0, 1, \dots, N,$$

propagates with each step in time resulting in another solution of $\tilde{w}_{kj}$. Let $z_{kj} = \tilde{w}_{kj} - w_{kj}$ be the error at the mesh point (x_k, t_j) for each $k = 0, 1, 2, \dots, N$ and $j = 0, 1, \dots, M$. z_{kj} satisfies the difference equation

$$\frac{z_{k,j+1} - z_{k,j}}{\tau} - K\frac{z_{k+1,j} - 2z_{k,j} + z_{k-1,j}}{h^2} = 0 \tag{2.15}$$

for $k = 1, 2, \dots, N$ and $j = 0, 1, \dots, M-1$. (Note that Eq. (2.15) coincides with (2.8). This is because Eq. (2.8) is linear and homogeneous.) In the Fourier method, we seek a particular solution of (2.15) in the form

$$z_{k,j} = \rho_q^j e^{iqx_k}, \quad q \in \mathbb{R}. \tag{2.16}$$

(Here $i = \sqrt{-1}$.) Then the finite-difference method (2.8) is stable, if all solutions having the form (2.16) are such that

$$|\rho_q| \leq 1$$

for all $q \in \mathbb{R}$.

Substitution of (2.16) into (2.15) yields

$$\frac{\rho_q^{j+1} e^{iqx_k} - \rho_q^j e^{iqx_k}}{\tau} - K\frac{\rho_q^j \left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right)}{h^2} = 0$$

or

$$\rho_q - 1 - \gamma\left(e^{iqh} - 2 + e^{-iqh}\right) = 0.$$

Since

$$e^{iqh} - 2 + e^{-iqh} = \left(e^{iqh/2} - e^{-iqh/2}\right)^2 = -4\sin^2\frac{qh}{2},$$

we obtain

$$\rho_q = 1 - 4\gamma\sin^2\frac{qh}{2}.$$

The method is stable if

$$\left|1 - 4\gamma\sin^2\frac{qh}{2}\right| \le 1$$

for each q, which is equivalent to

$$-1 \le 1 - 4\gamma\sin^2\frac{qh}{2} \le 1.$$

This double inequality is satisfied provided that

$$0 \le \gamma\sin^2\frac{qh}{2} \le \frac{1}{2}.$$

Evidently, the last inequality holds for all q if

$$0 \le \gamma \le \frac{1}{2} \quad \text{or equivalently} \quad 0 \le \tau \le \frac{h^2}{2K}. \tag{2.17}$$

Thus, we arrive at the conclusion that the forward-difference method is stable only if (2.17) is satisfied.

A method which is stable only if a certain condition holds is called **conditionally stable**. Thus, the forward-difference method for the heat equation is conditionally stable.

2.3 Implicit methods

2.3.1 Backward-difference method

To obtain a method which is **unconditionally stable**, we consider an implicit-difference method that results from using the backward-difference formula for $\partial u/\partial t(x,t)$:

$$\frac{\partial u}{\partial t}(x_k,t_j)=\frac{u(x_k,t_j)-u(x_k,t_{j-1})}{\tau}+\frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}(x_k,\mu_j),$$

where $\mu_j\in(t_{j-1},t_j)$. Using this instead of the forward-difference formula yields the following difference equation

$$\frac{w_{kj}-w_{k,j-1}}{\tau}-K\frac{w_{k+1,j}-2w_{kj}+w_{k-1,j}}{h^2}=0 \tag{2.18}$$

for $k=1,2,\dots,N-1$ and $j=1,2,\dots,M$. This equation is different from Eq. (2.8) and cannot be solved using a procedure as straightforward as that in the forward-difference method.

Equation (2.18) can be rewritten as

$$(1+2\gamma)w_{kj}-\gamma(w_{k+1,j}+w_{k-1,j})=w_{k,j-1} \tag{2.19}$$

for $k=1,2,\dots,N-1$ and $j=1,2,\dots,M$, or, equivalently,

$$A\mathbf{w}^{(j)}=\mathbf{w}^{(j-1)}\quad\text{for}\quad j=1,2,\dots, \tag{2.20}$$

where

$$A=\begin{bmatrix}1+2\gamma & -\gamma & 0 & \dots & \dots & 0\\ -\gamma & 1+2\gamma & -\gamma & \ddots & & \vdots\\ 0 & -\gamma & 1+2\gamma & -\gamma & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \ddots & 0\\ \vdots & & \ddots & \ddots & \ddots & -\gamma\\ 0 & \dots & \dots & 0 & -\gamma & 1+2\gamma\end{bmatrix},\quad \mathbf{w}^{(j)}=\begin{bmatrix}w_{1,j}\\ w_{2,j}\\ \vdots\\ \vdots\\ \vdots\\ w_{N-1,j}\end{bmatrix}. \tag{2.21}$$

If we know $\mathbf{w}^{(j-1)}$, Eq. (2.20) represents a system of linear equations for $\mathbf{w}^{(j)}$ which can be solved, e.g., by the double-sweep method that we will describe in the next subsection.

The finite-difference method described above is called the **backward-difference method**. It is an example of **implicit** finite-difference methods.

Now let us investigate the stability of the backward-difference scheme by the Fourier method. If we introduce the error $z_{k0}=w_{k0}-u_0(x_k)$ into the initial condition, it will propagate with each step in time. Let $z_{kj}=w_{kj}-u_0(x_k,t_j)$ be the error at the mesh point (x_k,t_j) for each $k=0,1,2,\dots,N$ and $j=0,1,\dots$. It follows from (2.18) that z_{kj} satisfies the difference equation

$$\frac{z_{kj}-z_{k,j-1}}{\tau}-K\frac{z_{k+1,j}-2z_{kj}+z_{k-1,j}}{h^2}=0 \tag{2.22}$$

for $k=1,2,\dots,N-1$ and $j=1,2,\dots,M$. We seek a particular solution of (2.22) in the form

$$z_{k,j}=\rho_q^j e^{iqx_k},\quad q\in\mathbb{R}. \tag{2.23}$$

The finite-difference method (2.18) is stable, if

$$|\rho_q|\le 1$$

for all q.

Substitution of (2.23) into (2.22) yields

$$e^{iqx_k}\left(\rho_q^j - \rho^{j-1}\right) - \gamma\rho_q^j\left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right) = 0$$

or

$$1 - \frac{1}{\rho_q} - \gamma\left(e^{iqh} - 2 + e^{-iqh}\right) = 0.$$

Since

$$e^{iqh} - 2 + e^{-iqh} = \left(e^{iqh/2} - e^{-iqh/2}\right)^2 = -4\sin^2\frac{qh}{2},$$

we obtain

$$\rho_q = \frac{1}{1 + 4\gamma\sin^2\frac{qh}{2}}.$$

Evidently, $|\rho_q| \le 1$ for all q, and therefore, the method is stable. Note that the method is **unconditionally stable** (independent of the choice of $\gamma = K\tau/h^2$). It can be shown that its local truncation error is $O(\tau + h^2)$ (the same as in the forward-difference method).

2.3.2 Double-sweep algorithm

The double-sweep method. How to solve system (2.19)? This system is tridiagonal. The most efficient algorithm of solving such systems is called **the double-sweep method**. Consider the following tridiagonal system:

$$A_i v_{i-1} - C_i v_i + B_i v_{i+1} = F_i \quad \text{for} \quad i = 1, \ldots, N-1; \quad v_0 = v_N = 0; \tag{2.24}$$

where the coefficients A_i, B_i and C_i satisfy the conditions

$$A_i, B_i, C_i > 0, \quad C_i \ge A_i + B_i. \tag{2.25}$$

(One can verify that system (2.19) satisfies the conditions (2.25).) To solve (2.24), we will seek α_i and β_i such that

$$v_{i-1} = \alpha_i v_i + \beta_i \quad \text{for} \quad i = 1, 2, \ldots, N. \tag{2.26}$$

Substitution of (2.26) into (2.24) yields

$$(\alpha_i A_i - C_i)v_i + B_i v_{i+1} + \beta_i A_i - F_i = 0 \quad \text{for} \quad i = 1, \ldots, N-1. \tag{2.27}$$

From (2.26), we also have

$$v_i = \alpha_{i+1} v_{i+1} + \beta_{i+1} \quad \text{for} \quad i = 0, 1, \ldots, N-1.$$

Substituting this into (2.27), we find that

$$[(\alpha_i A_i - C_i)\alpha_{i+1} + B_i]v_{i+1} + [(\alpha_i A_i - C_i)\beta_{i+1} + \beta_i A_i - F_i] = 0$$

for $i = 1, \ldots, N-1$. This equation is satisfied if the two expressions in the square brackets are both zero. This leads to the following recursive formulas:

$$\alpha_{i+1} = \frac{B_i}{C_i - \alpha_i A_i}, \quad \beta_{i+1} = \frac{\beta_i A_i - F_i}{C_i - \alpha_i A_i}, \quad \text{for} \quad i = 1, \ldots, N-1. \tag{2.28}$$

Now if α_1 and β_1 are known, then α_i and β_i for $i = 2, 3, \ldots, N$ can be computed from Eqs. (2.28). α_1 and β_1 can be determined from Eq. (2.26) and the fact that $v_0 = 0$. Indeed

$$v_0 = \alpha_1 v_1 + \beta_1 \quad \text{and} \quad v_0 = 0 \quad \Rightarrow \quad \alpha_1 v_1 + \beta_1 = 0.$$

To satisfy the last equation, we choose $\alpha_1 = 0$ and $\beta_1 = 0$. Once we know all α_i and β_i, we compute $v_{N-1}, v_{N-2}, \ldots, v_1$ using formula (2.26).

Now let's show that conditions (2.25) are sufficient for the double-sweep method to work. Numbers α_i and β_i are well-defined by Eqs. (2.28) provided that $C_i - \alpha_i A_i \neq 0$ for $i = 1, \ldots, N$. First we show that $0 \leq \alpha_i < 1$ for $i = 1, \ldots, N$. We will show this by induction. Since $\alpha_1 = 0$, it satisfies the required inequality. Now we need to show that if $0 \leq \alpha_i < 1$ for some $i \geq 1$, then $0 \leq \alpha_{i+1} < 1$. it follows from (2.25) that

$$C_i - \alpha_i A_i > A_i + B_i - \alpha_i A_i > B_i \quad \Rightarrow \quad \alpha_{i+1} = \frac{B_i}{C_i - \alpha_i A_i} < 1.$$

So, all α_i satisfy $0 \leq \alpha_i < 1$. Therefore,

$$C_i - \alpha_i A_i > A_i + B_i - \alpha_i A_i > B_i > 0 \quad \text{for} \quad i = 1, \ldots, N,$$

as required.

Is it possible to generalise the double-sweep method to the case when we have non-zero boundary conditions? The answer is 'Yes', and it is not difficult.

Suppose that we need to solve the same equations (2.24), but with non-zero boundary conditions:

$$v_0 = A, \quad v_N = B.$$

The only thing that we need to modify is the choice of α_1 and β_1. It follows from Eq. (2.26) and the boundary condition that

$$\alpha_1 v_1 + \beta_1 = A.$$

To satisfy this we simply choose

$$\alpha_1 = 0, \quad \beta_1 = A.$$

Boundary condition $v_N = B$ is taken into account automatically when v_{N-1} is computed using Eq. (2.26).

The double-sweep method can also be generalised for the following boundary conditions:

$$v_0 = \lambda v_1 + B \quad \text{and} \quad v_N = \mu v_{N-1} + B,$$

where λ, μ, A and B are constants and $|\lambda| \leq 1$, $|\mu| \leq 1$. (Prove this!)

2.3.3 Richardson method

How to improve the order of approximation of the heat equation by finite-difference schemes? The easiest way to obtain a finite-difference method whose truncation error is $O(\tau^2 + h^2)$ is to replace the forward (backward) difference formulas for u_t with the central difference formula:

$$\frac{\partial u}{\partial t}(x_k, t_j) = \frac{u(x_k, t_j + \tau) - u(x_k, t_j - \tau)}{2\tau} - \frac{\tau^2}{6}\frac{\partial^3 u}{\partial t^3}(x_k, \xi) \tag{2.29}$$

[It is an exercise for you to derive this formula.] This leads to the Richardson method, given by

$$\frac{w_{k,j+1} - w_{k,j-1}}{2\tau} - K\frac{w_{k+1,j} - 2w_{kj} + w_{k-1,j}}{h^2} = 0. \tag{2.30}$$

(for $k = 1, 2, \ldots, N-1$ and $j = 1, 2, \ldots, M-1$). [2]

[2] This was the first published finite-difference methods for PDEs, introduced by Lewis Fry Richardson already in 1911 in "The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations, with an Application to the Stresses in a Masonry Dam", Philosophical Transactions of the Royal Society of London A 210, pp.307 – 357.

Problem 2.1. Show that the Richardson method has the local truncation error $O(\tau^2 + h^2)$.

Solution. The local truncation error for Richardson's method is

$$\tau_{kj} = \frac{u_{k,j+1} - u_{k,j-1}}{2\tau} - K\frac{u_{k+1,j} - 2u_{kj} + u_{k-1,j}}{h^2}. \tag{2.31}$$

Here $u_{kj} = u(x_k, t_j)$. Expanding $u_{k,j\pm1} = u(x_k, t_{j\pm1}) = u(x_k, t_j \pm \tau)$ in Taylor's series at point (x_k, t_j), we obtain

$$u_{k,j+1} = u_{kj} + \tau\frac{\partial u}{\partial t}(x_k, t_j) + \frac{\tau^2}{2}\frac{\partial^2 u}{\partial t^2}(x_k, t_j) + O(\tau^3),$$
$$u_{k,j-1} = u_{kj} - \tau\frac{\partial u}{\partial t}(x_k, t_j) + \frac{\tau^2}{2}\frac{\partial^2 u}{\partial t^2}(x_k, t_j) + O(\tau^3).$$

It follows that

$$\frac{u_{k,j+1} - u_{k,j-1}}{2\tau} = \frac{\partial u}{\partial t}(x_k, t_j) + O(\tau^2). \tag{2.32}$$

Expanding $u_{k\pm1,j} = u(x_{k\pm1}, t_j) = u(x_k \pm h, t_j)$ in Taylor's series at point (x_k, t_j), we find that

$$u_{k+1,j} = u_{kj} + h\frac{\partial u}{\partial x}(x_k, t_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_k, t_j) + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_k, t_j) + O(h^4),$$
$$u_{k-1,j} = u_{kj} - h\frac{\partial u}{\partial x}(x_k, t_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_k, t_j) - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_k, t_j) + O(h^4).$$

Hence,

$$\frac{u_{k+1,j} - 2u_{kj} + u_{k-1,j}}{h^2} = \frac{\partial^2 u}{\partial x^2}(x_k, t_j) + O(h^2). \tag{2.33}$$

Substitution of (2.32) and (2.33) in Eq. (2.31) yields

$$\tau_{kj} = \frac{\partial u}{\partial t}(x_k, t_j) - K\frac{\partial^2 u}{\partial x^2}(x_k, t_j) + O(\tau^2) + O(h^2).$$

Taking into account the fact that $u(x, t)$ satisfies the heat equation, we obtain

$$\tau_{kj} = O(\tau^2 + h^2).$$

Problem 2.2. Investigate the stability of Richardson's method, given by (2.30), by the Fourier method.

Solution. If w_{kj} and $\tilde{w}_{kj}$ are two solutions of the difference equation (2.30) that correspond to different initial conditions, then the perturbation (error) $z_{kj} = \tilde{w}_{kj} - w_{kj}$ satisfies the homogeneous difference equation

$$\frac{z_{k,j+1} - z_{k,j-1}}{2\tau} - K\frac{z_{k+1,j} - 2z_{kj} + z_{k-1,j}}{h^2} = 0. \tag{2.34}$$

We seek a particular solution of (2.34) in the form

$$z_{k,j} = \rho_q^j e^{iqx_k}.$$

Substituting this into (2.34), we obtain

$$e^{iqx_k}\left(\rho_q^{j+1} - \rho_q^{j-1}\right) - 2\gamma\rho_q^j\left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right) = 0$$

or

$$\rho_q^2 - 1 - 2\gamma\rho_q\left(e^{iqh} - 2 + e^{-iqh}\right) = 0.$$

Since

$$e^{iqh} - 2 + e^{-iqh} = \left(e^{iqh/2} - e^{-iqh/2}\right)^2 = -4\sin^2\frac{qh}{2},$$

we obtain

$$\rho_q^2 + \left(8\gamma\sin^2\frac{qh}{2}\right)\rho_q - 1 = 0.$$

One of the roots of this quadratic equation, namely,

$$\rho_q = -4\gamma\sin^2\frac{qh}{2} - \sqrt{16\gamma^2\sin^4\frac{qh}{2} + 1},$$

is greater than 1 in magnitude for any q such that $\sin(qh/2) \neq 0$. Thus, Richardson's method is (unconditionally) unstable.

2.3.4 Crank-Nicolson method

Let δ_x^2 be a finite-difference operator defined by

$$\delta_x^2 w_{kj} = w_{k+1,j} - 2w_{kj} + w_{k-1,j}. \tag{2.35}$$

Consider the initial boundary value problem for the heat equation

$$\frac{\partial u}{\partial t}(x,t) = K\frac{\partial^2 u}{\partial x^2}(x,t), \quad 0 < x < L, \quad 0 < 0 < T, \tag{2.36}$$

$$u(0,t) = u(L,t) = 0 \quad \text{for} \quad t > 0, \tag{2.37}$$

$$u(x,0) = u_0(x). \tag{2.38}$$

With the help of the above notation the (explicit) forward-difference method for Eq. (2.36) can be written as

$$\frac{w_{k,j+1} - w_{kj}}{\tau} - K\frac{\delta_x^2 w_{kj}}{h^2} = 0, \tag{2.39}$$

the (implicit) backward-difference method can be written as

$$\frac{w_{k,j+1} - w_{kj}}{\tau} - K\frac{\delta_x^2 w_{k,j+1}}{h^2} = 0. \tag{2.40}$$

A better method is obtained by averaging the forward-difference method (2.39) and the backward-difference method (2.40). Adding Eq. (2.39) to Eq. (2.40) with weights 1/2, we obtain the difference equation

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} - \frac{K}{2h^2}\left(\delta_x^2 w_{kj} + \delta_x^2 w_{k,j+1}\right) = 0 \tag{2.41}$$

or, equivalently,

$$w_{k,j+1} - w_{k,j} - \frac{\gamma}{2}\left(w_{k+1,j} - 2w_{kj} + w_{k-1,j} + w_{k+1,j+1} - 2w_{k,j+1} + w_{k-1,j+1}\right) = 0 \tag{2.42}$$

where $\gamma = K\tau/h^2$. This method is called the **Crank-Nicolson method**. It can be shown that the local truncation error of the Crank-Nicolson method is $O(\tau^2 + h^2)$ (see problem sheet).

In matrix form, Eq. (2.42) can be written as

$$A\mathbf{w}^{(j+1)} = B\mathbf{w}^{(j)} \quad \text{for} \quad j = 0, 1, 2, \ldots, \tag{2.43}$$

where

$$A = \begin{bmatrix} 1+\gamma & -\gamma/2 & 0 & \dots & \dots & 0 \\ -\gamma/2 & 1+\gamma & -\gamma/2 & \ddots & & \vdots \\ 0 & -\gamma/2 & 1+\gamma & -\gamma/2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -\gamma/2 \\ 0 & \dots & \dots & 0 & -\gamma/2 & 1+\gamma \end{bmatrix},$$

$$B = \begin{bmatrix} 1-\gamma & \gamma/2 & 0 & \dots & \dots & 0 \\ \gamma/2 & 1-\gamma & \gamma/2 & \ddots & & \vdots \\ 0 & \gamma/2 & 1-\gamma & \gamma/2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \gamma/2 \\ 0 & \dots & \dots & 0 & \gamma/2 & 1-\gamma \end{bmatrix}.$$

If we know $\mathbf{w}^{(j)}$, we can find $\mathbf{w}^{(j+1)}$ by solving the tridiagonal linear system (2.43) by the double-sweep method.

To investigate the stability of the Crank-Nicolson method, we employ the Fourier method. If $z_{kj} = w_{kj} - \tilde{w}_{kj}$, then

$$z_{k,j+1} - z_{k,j} - \frac{\gamma}{2}\left(z_{k+1,j} - 2z_{kj} - z_{k-1,j} + z_{k+1,j+1} - 2z_{k,j+1} + z_{k-1,j+1}\right) = 0 \tag{2.44}$$

for $k = 1, 2, \dots, N-1$ and $j = 1, 2, \dots$. We seek a particular solution of (2.44) in the form

$$z_{k,j} = \rho_q^j e^{iqx_k}, \quad q \in \mathbb{R}. \tag{2.45}$$

Then, the finite-difference method (2.42) is stable, if

$$|\rho_q| \le 1$$

for all q.

Substitution of (2.45) into (2.44) yields

$$\begin{aligned} 0 = {} & e^{iqx_k}\left(\rho_q^{j+1} - \rho_q^j\right) \\ & - \frac{\gamma}{2}\left[\rho_q^j\left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right) + \rho_q^{j+1}\left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right)\right] \end{aligned} \tag{2.46}$$

or

$$0 = \rho_q - 1 - \frac{\gamma}{2}\left[e^{iqh} - 2 + e^{-iqh} + \rho_q\left(e^{iqh} - 2 + e^{-iqh}\right)\right]$$

$$\Rightarrow \quad 0 = \rho_q - 1 - \frac{\gamma}{2}(\rho_q + 1)\left(e^{iqh} - 2 + e^{-iqh}\right).$$

Since

$$e^{iqh} - 2 + e^{-iqh} = \left(e^{iqh/2} - e^{-iqh/2}\right)^2 = -4\sin^2\frac{qh}{2},$$

we obtain

$$\rho_q - 1 + (\rho_q + 1)2\gamma\sin^2\frac{qh}{2} = 0 \quad \Rightarrow \quad \rho_q = \frac{1 - 2\gamma\sin^2\frac{qh}{2}}{1 + 2\gamma\sin^2\frac{qh}{2}}.$$

To find out whether $|\rho_q| \leq 1$, we consider the function $f(\alpha) = \frac{1-\alpha}{1+\alpha}$ for $\alpha \in [0, \infty)$. Since $f'(\alpha) = -\frac{2}{(1+\alpha)^2} < 0$ for all $\alpha \in [0, \infty)$, it is a decreasing function. In addition to this, we have $f(0) = 1$ and $\lim_{\alpha \to \infty} f(\alpha) = -1$. We conclude that $-1 < f(\alpha) \leq 1$ for $\alpha \geq 0$. Therefore, $|\rho_q| \leq 1$ for all $q \in \mathbb{R}$, and the Crank-Nicolson method is unconditionally stable.

2.4 Consistency, stability, and convergence

By definition, a finite-difference approximation to a differential equation is **consistent** with this differential equation if local truncation errors tend to zero as the step size goes to zero, i.e.

$$\max_{k,j} |\tau_{ki}(h, \tau)| \to 0 \quad \text{as} \quad h, \tau \to 0.$$

Finite difference equation for PDEs arising in initial value problems may display a phenomenon which has no counterpart in ordinary differential equations. Successive reduction of step sizes in x and t may generate a finite difference solution which is stable, but which may converge to the solution of a different differential equation. For example, the Du Fort - Frankel method for solving the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \tag{2.47}$$

is given by

$$\frac{w_{k,j+1} - w_{k,j-1}}{2\tau} - \frac{w_{k+1,j} - w_{k,j-1} - w_{k,j+1} + w_{k-1,j}}{h^2} = 0.$$

It can be shown that this method is always stable and has the local truncation error

$$\tau_{kj} = O\left(\tau^2 + h^2 + \frac{\tau^2}{h^2}\right).$$

Thus, the Du Fort-Frankel formula is consistent with Eq. (2.47) only if τ goes to zero faster than h. If they go to zero at the same rate, so that $\tau/h = \beta$ is fixed, then this approximation is consistent not with the diffusion equation but with the hyperbolic equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} + \beta^2 \frac{\partial^2 u}{\partial t^2} = 0.$$

By definition, a finite-difference method is said to be **convergent** if the total error of the method

$$E = \max_{k,j} |u_{kj} - w_{kj}|$$

tends to zero as $h \to 0$ and $\tau \to 0$:

$$E \to 0 \quad \text{as} \quad h \to 0 \quad \text{and} \quad \tau \to 0.$$

We know that an approximation for the initial boundary value problem (2.36)–(2.38) can be obtained using the forward-difference method

$$w_{k,j+1} = (1 - 2\gamma)\, w_{kj} + \gamma\, (w_{k+1,j} + w_{k-1,j}) \quad \text{for} \quad k = 1, 2, \ldots, N-1, j = 0, 1, \ldots, M-1, \tag{2.48}$$

$$w_{0,j} = w_{N,j} = 0 \quad \text{for} \quad j = 0, 1, \ldots, M \quad \text{and} \tag{2.49}$$

$$w_{k,0} = u_0(x_k) \quad \text{for} \quad k = 1, 2, \ldots, N-1. \tag{2.50}$$

Here $\gamma \equiv K\tau/h^2$. Also, we know that the local truncation error for the method is

$$\tau_{kj} = \frac{u_{k,j+1} - u_{kj}}{\tau} - \frac{\gamma}{\tau}\left(u_{k+1,j} - 2u_{kj} + u_{k-1,j}\right) = O(\tau + h^2), \tag{2.51}$$

so that the method is consistent, and that the method is stable under the condition

$$\gamma \le \frac{1}{2}. \tag{2.52}$$

Theorem 2.2. *If the stability condition (2.52) is satisfied, the forward-difference method (2.48), (2.50) is convergent.*

Proof. It follows from (2.51) that

$$u_{k,j+1} = (1-2\gamma)u_{kj} + \gamma\left(u_{k+1,j} + u_{k-1,j}\right) + O(\tau^2 + h^2\tau). \tag{2.53}$$

Let

$$z_{kj} = u_{kj} - w_{kj}. \tag{2.54}$$

Then it follows from (2.53) and (2.48) that

$$z_{k,j+1} = (1-2\gamma)z_{kj} + \gamma\left(z_{k+1,j} + z_{k-1,j}\right) + O(\tau^2 + h^2\tau) \tag{2.55}$$

for $k = 1, 2, \ldots, N-1$ and $j = 0, 1, \ldots, M-1$. In view of (2.50),

$$z_{0,j} = z_{N,j} = 0, \quad j = 0, 1, \ldots, M \quad \text{and} \quad z_{k,0} = 0, \quad k = 1, 2, \ldots, N-1. \tag{2.56}$$

If the stability condition (2.52) is satisfied, then all coefficients on the right side of (2.55) are positive.

By definition of O, Eq. (2.55) implies that

$$|z_{k,j+1} - [(1-2\gamma)z_{kj} + \gamma\left(z_{k+1,j} + z_{k-1,j}\right)]| \le A_{k,j+1}\,|\tau^2 + h^2\tau|$$

for some positive $A_{k,j+1}$. This and the inequality

$$|a| = |b + (a-b)| \le |b| + |a-b|$$

(which is valid for any a and b) have a consequence that

$$|z_{k,j+1}| = |(1-2\gamma)z_{kj} + \gamma\left(z_{k+1,j} + z_{k-1,j}\right)| + A_{k,j+1}\,|\tau^2 + h^2\tau| \tag{2.57}$$

for $k = 1, 2, \ldots, N-1$ and $j = 0, 1, \ldots, M-1$.

Let $\mathbf{z}^{(j)} = (z_{1j}, z_{2j}, \ldots, z_{N-1,j})^T$ and let

$$\|\mathbf{z}^{(j)}\| = \|\mathbf{z}^{(j)}\|_\infty = \max_{1\le k\le N-1} |z_{kj}|.$$

It follows from Eq. (2.55) that

$$\begin{aligned}
\|\mathbf{z}^{(j+1)}\| &= \max_k |z_{k,j+1}| \\
&= \max_k \left\{|(1-2\gamma)z_{kj} + \gamma\left(z_{k+1,j} + z_{k-1,j}\right)| + A_{j+1}\,|\tau^2 + h^2\tau|\right\} \\
&\le (1-2\gamma)\max_k |z_{kj}| + \gamma \max_k |z_{k+1,j}| + \gamma \max_k |z_{k-1,j}| \\
&\qquad + A_{j+1}\,|\tau^2 + h^2\tau| \\
&\le \|\mathbf{z}^{(j)}\| + A_{k,j+1}\,|\tau^2 + h^2\tau|.
\end{aligned}$$

Here $A_{j+1} = \max\limits_k A_{k,j+1}$. This inequality is valid for all $j = 0, 1, \ldots, M-1$. It follows that

$$\begin{aligned} \|\mathbf{z}^{(j)}\| &\le \|\mathbf{z}^{(j-1)}\| + A_j\,|\tau^2 + h^2\tau| \\ &\le \|\mathbf{z}^{(j-2)}\| + (A_j + A_{j-1})\,|\tau^2 + h^2\tau| \\ &\quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ &\le \|\mathbf{z}^{(0)}\| + (A_j + \cdots + A_1)\,|\tau^2 + h^2\tau| \\ &\le \|\mathbf{z}^{(0)}\| + Aj\tau\,|\tau + h^2| = \|\mathbf{z}^{(0)}\| + At_j\,|\tau + h^2| \end{aligned}$$

where

$$A = \max_j A_j = \max_{k,j} A_{kj}.$$

Since $\mathbf{z}_0 = 0$, we obtain

$$\|\mathbf{z}^{(j)}\| \le At_j\,|\tau + h^2| \le AT|\tau + h^2|. \tag{2.58}$$

Since this inequality is valid for all $j = 0, 1, \ldots, M$, we obtain

$$E = \max_{k,j} |z_{kj}| = \max_j \|\mathbf{z}^{(j)}\| \le AT|\tau + h^2| \quad \Rightarrow \quad E \to 0 \quad \text{as} \quad \tau, h \to 0.$$

Thus, we have proved that the forward-difference method is convergent. □

Similar statements can be proved for other finite-difference methods.

Lax and Richtmyer[3] and others studied the relation between consistency, stability, and convergence of the approximations of linear initial value problems by finite difference equations. The major result of that study is

Theorem 2.3 (Lax equivalence theorem)**.** *Given a well-posed initial boundary value problem and a finite difference approximation to it that satisfies the consistency condition, then stability is the necessary and sufficient condition for convergence.*

[3]Lax, P.D., and Richtmyer, R.D. (1956). Survey of the stability of linear finite difference equations. Comm. Pure Appl. Math. 9, 267293.

2.5 Non-homogeneous heat equation with non-homogeneous boundary conditions

So far, we considered the homogeneous heat equation with homogeneous (zero) boundary conditions. Consider first the non-homogeneous heat equation with homogeneous boundary conditions, given by

$$\frac{\partial u}{\partial t}(x,t) - K\frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < L, \quad 0 < t < T, \tag{2.59}$$

$$u(0,t) = u(L,t) = 0 \quad \text{for} \quad t > 0, \tag{2.60}$$

$$u(x,0) = u_0(x). \tag{2.61}$$

The generalisation of the forward-difference and backward-difference methods is straightforward.

Forward-difference method:

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} - K\frac{w_{k+1,j} - 2w_{kj} + w_{k-1,j}}{h^2} = f(x_k, t_j). \tag{2.62}$$

Backward-difference method:

$$\frac{w_{k,j} - w_{k,j-1}}{\tau} - K\frac{w_{k+1,j} - 2w_{kj} + w_{k-1,j}}{h^2} = f(x_k, t_j). \tag{2.63}$$

The non-homogeneous version of the Crank-Nicolson method is given by

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} - K\frac{w_{k+1,j} - 2w_{kj} + w_{k-1,j} + w_{k+1,j+1} - 2w_{k,j+1} + w_{k-1,j+1}}{2h^2} = f(x_k, t_j + \tau/2). \tag{2.64}$$

Note that the right hand side of Eq. (2.64) can be replaced by $[f(x_k,t_j) + f(x_k,t_{j+1})]/2$ because

$$f(x_k, t_j + \tau/2) - \frac{1}{2}[f(x_k,t_j) + f(x_k,t_{j+1})] = O(\tau^2).$$

Do the above modifications affect the stability of the methods? The answer is NO, because in spite of the fact that the difference equations are non-homogeneous, the corresponding equations for perturbation z_{kj} remain the same homogeneous equations as before.

Non-homogeneous boundary conditions. Consider the initial boundary value problem

$$\frac{\partial u}{\partial t}(x,t) - K\frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < L, \quad 0 < t < T, \tag{2.65}$$

$$u(0,t) = \mu_1(t), \quad u(L,t) = \mu_2(t) \quad \text{for} \quad t > 0, \tag{2.66}$$

$$u(x,0) = u_0(x), \tag{2.67}$$

where $\mu_1(t)$ and $\mu_2(t)$ are some given functions. The non-homogeneous boundary conditions (2.66) can be dealt with in two ways. First, one can simply change the boundary conditions for w_{kj} by letting

$$w_{0,j} = \mu_1(t_j), \quad w_{N,j} = \mu_2(t_j). \tag{2.68}$$

Note that for implicit schemes where we employ the double-sweep method, the latter should be modified in order to allow non-zero boundary conditions.

Second, one can reduce problem (2.65)–(2.67) to an initial boundary problem with homogeneous boundary conditions. To do this, we choose any function $g(x,t)$ satisfying the boundary conditions

$$g(0,t) = \mu_1(t), \quad g(L,t) = \mu_2(t). \tag{2.69}$$

Now if $u(x,t) = v(x,t) + g(x,t)$ and $u(x,t)$ is the solution of (2.65)–(2.67), then $v(x,t)$ must satisfy the initial boundary value problem

$$\frac{\partial v}{\partial t}(x,t) - K\frac{\partial^2 v}{\partial x^2}(x,t) = \tilde{f}(x,t), \quad 0 < x < L, \quad 0 < t < T, \tag{2.70}$$

$$v(0,t) = 0, \quad v(L,t) = 0 \quad \text{for} \quad t > 0, \tag{2.71}$$

$$v(x,0) = v_0(x), \tag{2.72}$$

where

$$\tilde{f}(x,t) = f(x,t) - \frac{\partial g}{\partial t}(x,t) + K\frac{\partial^2 g}{\partial x^2}(x,t), \quad v_0(x) = u_0(x) - g(x,0).$$

For example, we can choose the function

$$g(x,t) = \mu_1(t) + [\mu_2(t) - \mu_1(t)]\frac{x}{L}.$$

2.6 Boundary conditions of other types

.

In all the problems we discussed so far, the boundary conditions did not require approximations of any kind. A different situation arises when we have the boundary conditions for the derivative of u:

$$\frac{\partial u}{\partial x}(0,t) = \mu_1(t), \quad \frac{\partial u}{\partial x}(L,t) = \mu_2(t). \tag{2.73}$$

These are known as Dirichlet boundary conditions. Evidently, we need to approximate these conditions.

Consider the non-homogeneous heat equation

$$\frac{\partial u}{\partial t} - K\frac{\partial^2 u}{\partial x^2} = f(x,t), \quad 0 < x < L, \quad 0 < 0 < T, \tag{2.74}$$

subject to the boundary conditions (2.73) and the initial condition

$$u(x,0) = u_0(x). \tag{2.75}$$

As in the case of non-homogeneous boundary conditions for u, the problem (2.73)–(2.75) with non-homogeneous boundary conditions for u_x can be reduced to a problem with homogeneous boundary condition. To do this, we write $u(x,t) = v(x,t) + g(x,t)$ with any fixed function $g(x,t)$ satisfying conditions (2.73). Then for $v(x,t)$, we obtain the problem

$$\frac{\partial v}{\partial t} - K\frac{\partial^2 v}{\partial x^2} = \tilde{f}(x,t), \quad 0 < x < L, \quad 0 < 0 < T,$$

$$\frac{\partial v}{\partial x}(0,t) = 0, \quad \frac{\partial v}{\partial x}(L,t) = 0, \quad v(x,0) = v_0(x),$$

where $\tilde{f}(x,t) = f(x,t) - g_t(x,t) + Kg_{xx}(x,t)$ and $v_0(x) = u_0(x) - g(x,0)$. [An example of function g: $g(x,t) = \mu_1(t)x + [\mu_2(t) - \mu_1(t)]x^2/(2L)$.] Therefore, we will discuss here only the homogeneous conditions

$$\frac{\partial u}{\partial x}(0,t) = 0, \quad \frac{\partial u}{\partial x}(L,t) = 0. \tag{2.76}$$

In what follows we restrict our analysis to the boundary condition at $x = 0$ (the other boundary condition can be treated similarly). If we use the two-point forward-difference formula for the derivative at $(x = 0, t = t_j)$, then

$$\frac{w_{1,j} - w_{0,j}}{h} = 0.$$

However, this formula for $u_x(0,t)$ has the truncation error $O(h)$, while in all the methods that we have considered the truncation error of the relevant difference equations was at least $O(h^2)$. So, the use of this formula would increase the local truncation error of these methods.

Suppose that we use the forward-difference method to approximate the heat equation (2.74) at the interior grid points. How to approximate the boundary conditions (2.76) with truncation error $O(h^2)$ for this method? We will consider two ways of doing this. In the first one, we add a 'false' boundary at $x = x_{-1} = x_0 - h$ and assume that the forward-difference scheme approximates the heat equation at points (x_0, t_j) $(j = 1, 2, \dots)$. Then, we have

$$\frac{w_{0,j+1} - w_{0,j}}{\tau} - K\frac{w_{1,j} - 2w_{0j} + w_{-1,j}}{h^2} = f(0, t_j). \tag{2.77}$$

Approximating $u_x(0,t_j)$ by the central difference formula (whose truncation error is $O(h^2)$), we find that

$$\frac{w_{1,j} - w_{-1,j}}{2h} = 0. \tag{2.78}$$

Eliminating $w_{-1,j}$ from Eqs. (2.77) and (2.78), we obtain

$$w_{0,j+1} = (1 - 2\gamma)w_{0,j} + 2\gamma w_{1,j} + \tau f(0, t_j). \tag{2.79}$$

This is an explicit formula that relates the boundary values at the time levels t_j and t_{j+1}.

Similarly, for the boundary condition at $x = L$, one can obtain the formula

$$w_{N,j+1} = (1 - 2\gamma)w_{N,j} + 2\gamma w_{N-1,j} + \tau f(L, t_j). \tag{2.80}$$

In another (more general) technique which leads to Eqs. (2.79) and (2.80), we expand $u(x_1, t_j)$ in Taylor's series at point (x_0, t_j):

$$u(x_1, t_j) = u(x_0, t_j) + hu_x(x_0, t_j) + \frac{h^2}{2}u_{xx}(x_0, t_j) + O(h^3).$$

Taking account of the boundary condition and the fact that $u(x,t)$ is the solution of the heat equation, we find that

$$\begin{aligned} u(x_1, t_j) = u(x_0, t_j) &+ \frac{h^2}{2K}\left(\frac{\partial u}{\partial t}(x_0, t_j) - f(x_0, t_j)\right) + O(h^3) \\ = u(x_0, t_j) &+ \frac{h^2}{2K}\left(\frac{u(x_0, t_{j+1}) - u(x_0, t_j)}{\tau} - f(x_0, t_j)\right) \\ &+ O(\tau h^2) + O(h^3). \end{aligned}$$

The last equation leads to the difference equation (2.79).

What about the stability of this method? Evidently, the Fourier method would produce the same answer as before, because boundary conditions are ignored. The stability can be investigated with the help of the matrix stability analysis[4]. The forward-difference method described above can be written in the matrix form

$$\mathbf{w}^{(j)} = A\mathbf{w}^{(j-1)} + \tau\mathbf{F}^{(j-1)} \quad \text{for} \quad j = 1, 2, \dots, \tag{2.81}$$

[4]See W. F. Ames, Numerical methods for partial differential equation, Academic Press, 1977.

where

$$A = \begin{bmatrix} 1-2\gamma & 2\gamma & 0 & \dots & \dots & \dots & 0 \\ \gamma & 1-2\gamma & \gamma & \ddots & & & \vdots \\ 0 & \gamma & 1-2\gamma & \gamma & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \gamma & 0 \\ \vdots & & & \ddots & \gamma & 1-2\gamma & \gamma \\ 0 & \dots & \dots & \dots & 0 & 2\gamma & 1-2\gamma \end{bmatrix}, \tag{2.82}$$

$$\mathbf{w}^{(j)} = \begin{bmatrix} w_{0,j} \\ w_{1,j} \\ w_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ w_{N-1,j} \\ w_{N,j} \end{bmatrix}, \qquad \mathbf{F}^{(j)} = \begin{bmatrix} f_{0,j} \\ f_{1,j} \\ f_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ f_{N-1,j} \\ f_{N,j} \end{bmatrix}. \tag{2.83}$$

The method will be stable if for all eigenvalues λ of A,

$$|\lambda| \le 1,$$

and it can be shown (see W. F. Ames, Numerical methods for partial differential equations) that this results in the same stability condition $\gamma \le 1/2$.

The implicit backward-difference method for problem (2.73)–(2.75) can be written in the matrix form as follows

$$B\mathbf{w}^{(j)} = \mathbf{w}^{(j-1)} + \tau\mathbf{F}^{(j)} \quad \text{for} \quad j = 1, 2, \dots, \tag{2.84}$$

where

$$B = \begin{bmatrix} 1+2\gamma & -2\gamma & 0 & \dots & \dots & \dots & 0 \\ -\gamma & 1+2\gamma & -\gamma & \ddots & & & \vdots \\ 0 & -\gamma & 1+2\gamma & -\gamma & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & -\gamma & 0 \\ \vdots & & & \ddots & -\gamma & 1+2\gamma & -\gamma \\ 0 & \dots & \dots & \dots & 0 & -2\gamma & 1+2\gamma \end{bmatrix}, \tag{2.85}$$

$$\mathbf{w}^{(j)} = \begin{bmatrix} w_{0,j} \\ w_{1,j} \\ w_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ w_{N-1,j} \\ w_{N,j} \end{bmatrix}, \quad \mathbf{F}^{(j)} = \begin{bmatrix} f_{0,j} \\ f_{1,j} \\ f_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ f_{N-1,j} \\ f_{N,j} \end{bmatrix}. \tag{2.86}$$

It can be shown that it remains unconditionally stable.

More general boundary conditions (Robin boundary conditions)

$$\frac{\partial u}{\partial x}(0,t) + c_1(t)u(0,t) = \mu_1(t), \quad \frac{\partial u}{\partial x}(L,t) + c_2(t)u(L,t) = \mu_2(t), \tag{2.87}$$

where $c_1(t)$ and $c_2(t)$ are some given functions, can be treated similarly.

2.7 Variable coefficients

.

Consider the parabolic equation

$$\frac{\partial u}{\partial t} = a(x,t)\frac{\partial^2 u}{\partial x^2} + b(x,t)\frac{\partial u}{\partial x} + c(x,t)u + d(x,t), \quad 0 < x < L, \quad 0 < t < T, \tag{2.88}$$

subject to the initial and boundary conditions

$$u(x,0) = u_0(x), \tag{2.89}$$

$$u(0,t) = 0, \quad u(L,t) = 0. \tag{2.90}$$

Here we assume that

$$a(x,t) > 0 \quad \text{for} \quad 0 \le x \le L, \quad 0 < t < T. \tag{2.91}$$

Most of the previously discussed methods can be generalized to the case of the initial boundary value problem (2.88)–(2.90). How to do this?

Employing the two-point forward difference formula for u_t and the central difference formulae for u_x and u_{xx}, we obtain the following finite-difference approximation for Eq. (2.88):

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} = a_{kj}\frac{w_{k+1,j} - 2w_{k,j} + w_{k-1,j}}{h^2} + b_{kj}\frac{w_{k+1,j} - w_{k-1,j}}{2h} + c_{kj}w_{k,j} + d_{kj}, \tag{2.92}$$

for $k = 1, \ldots, N-1$ and $j = 0, 1, \ldots, M-1$. In Eq. (2.92), $a_{kj} = a(x_k, t_j)$, $b_{kj} = b(x_k, t_j)$, etc. Since the forward difference formula for $u_t t$ has the truncation error $O(\tau)$, and the errors of the central difference formulae for u_x and u_{xx} are $O(h^2)$, the local truncation error of the difference equation (2.92) is

$$\tau_{kj} = O(\tau + h^2).$$

This, together with obvious boundary conditions, gives us the explicit method for solving (2.88)–(2.90) which is similar to the explicit forward-difference method for the heat equation that we discussed earlier. Similarly, if we use the backward difference formula for u_t, we obtain the difference equation

$$\frac{w_{k,j} - w_{k,j-1}}{\tau} = a_{kj}\frac{w_{k+1,j} - 2w_{k,j} + w_{k-1,j}}{h^2} + b_{kj}\frac{w_{k+1,j} - w_{k-1,j}}{2h} + c_{kj}w_{k,j} + d_{kj}, \tag{2.93}$$

whose local truncation error is $O(\tau + h^2)$ and which is similar to the implicit backward-difference method for the heat equation.

It can be shown that Crank-Nicolson's technique applied to (2.88) yields:

$$\begin{aligned}\frac{w_{k,j+1} - w_{k,j}}{\tau} &= \frac{a_{k,j+1/2}}{2h^2}\,\delta_x^2\,(w_{k,j} + w_{k,j+1}) + \frac{b_{k,j+1/2}}{4h}\,\delta_x\,(w_{k,j} + w_{k,j+1}) \\ &\quad + \frac{c_{k,j+1/2}}{2}\,(w_{k,j} + w_{k,j+1}) + d_{k,j+1/2},\end{aligned}$$

where

$$a_{k,j+1/2} = \frac{a(x_k,t_j) + a(x_k,t_{j+1})}{2}, \quad b_{k,j+1/2} = \frac{b(x_k,t_j) + b(x_k,t_{j+1})}{2}, \quad \text{etc.}$$

What about the stability of these methods? If the coefficients a, b, c and d do not involve t, the stability can be investigated [5]. In particular, it can be shown that the explicit method (2.92) is stable provided that

$$\frac{\tau}{h^2} < \frac{1}{2a(x,t)} \quad \text{for} \quad 0 \le x \le L, \quad t > 0.$$

Example 2.3. Consider the two-dimensional heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad \text{for} \quad 0 < t < T \tag{2.94}$$

in the circular domain $\mathcal{D}$:

$$\mathcal{D} = \{(x,y) | \sqrt{x^2+y^2} < R\},$$

subject to the boundary condition

$$u(x,y,t) = \mu(x,y,t) \quad \text{for} \quad \sqrt{x^2+y^2} = R, \tag{2.95}$$

and the initial condition

$$u(x,y,0) = u_0(x,y) \quad \text{for} \quad (x,y) \in \mathcal{D}. \tag{2.96}$$

In polar coordinates (r,θ) [such that $x = r\cos\theta$, $y = r\sin\theta$], the above initial boundary value problem takes the form

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \quad \text{in} \quad \mathcal{D}, \\ &u(r,\theta,t)|_{r=R} = \mu(\theta,t), \quad u(r,\theta,0) = u_0(r,\theta). \end{aligned} \tag{2.97}$$

If the initial and boundary conditions do not involve θ, i.e. $u_0 = u_0(r)$ and $\mu = \mu(t)$, then the solutions of problem (2.97) are rotationally symmetric, i.e. independent of θ, for all $t > 0$. In this case, (2.97) simplifies to

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} \quad \text{for} \quad 0 < r < R, \\ &u(r,t)|_{r=R} = \mu(t), \quad u(r,0) = u_0(r). \end{aligned} \tag{2.98}$$

There are two apparent difficulties with problem (2.98): (i) we have only one boundary condition and (ii) the term $(1/r)(\partial u/\partial r)$ is singular at $r = 0$. The first difficulty can be eliminated by using the following condition

$$u_r(0,t) = 0. \tag{2.99}$$

This condition holds for any sufficiently smooth solution of (2.98). To show this, we assume that u is twice continuously differentiable with respect to both t and r in D and integrate Eq. (2.98) in r from 0 to $\epsilon > 0$ with weight r:

$$\int_0^\epsilon u_t(r,t) r\, dr = \int_0^\epsilon \frac{1}{r}\frac{\partial}{\partial r}\Big(r\, u_r(r,t)\Big) r\, dr = r\, u_r(r,t)\,\Big|_0^\epsilon = \epsilon\, u_r(\epsilon,t).$$

Hence,

$$u_r(\epsilon,t) = \frac{1}{\epsilon}\int_0^\epsilon u_t(r,t) r\, dr. \tag{2.100}$$

[5] See W. F. Ames, Numerical methods for partial differential equation, Academic Press, 1977.

For small ϵ,

$$\int_0^\epsilon u_t(r,t)r\,dr = \int_0^\epsilon (u_t(0,t)+O(r))r\,dr = u_t(0,t)\int_0^\epsilon (r+O(r^2))\,dr$$
$$= u_t(0,t)\,\frac{\epsilon^2}{2} + O(\epsilon^3)$$

Passing to the limit as $\epsilon \to 0$ in (2.100) leads to condition (2.99):

$$u_r(0,t) = \lim_{\epsilon\to 0} u_r(\epsilon,t) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_0^\epsilon u_t(r,t)r\,dr = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(u_t(0,t)\,\frac{\epsilon^2}{2}+O(\epsilon^3)\right) = 0.$$

If we try to approximate the boundary condition (2.99) by introducing a false boundary at $r = r_{-1} = -h$ in the same way as before, we will need to use a finite-difference scheme at points $(r_0, t_j) = (r_0, t_j)$ and therefore to approximate the term $r^{-1}u_r$ at $r = 0$. However, we do not know $\lim_{r\to 0} r^{-1}u_r$ and, therefore, cannot approximate it. Nevertheless we can avoid the singularity at $r = 0$, if we choose the grid points in such a way that $r_0 = h/2$ and $r_k = r_0 + kh$ for $k = 1,\dots,N$ with $h = 2L/(2N+1)$. Then we introduce the false boundary at $r_{-1} = -h/2$ and approximate (2.99) by the central difference formula

$$\frac{w_{0,j} - w_{-1,j}}{h} = 0 \tag{2.101}$$

whose truncation error is $O(h^2)$.

When the Crank-Nicolson concept is applied to Eq. (2.98), we obtain the difference equation for the interior grid points (r_k, t_j):

$$\frac{w_{k,j+1}-w_{k,j}}{\tau} - \frac{1}{2h^2}\left(\delta_r^2 + \frac{h}{2r_k}\delta_r\right)(w_{k,j+1}+w_{kj}) = 0 \tag{2.102}$$

for $k = 1,\dots,N-1)$ and $j = 0,\dots,M-1$. Here δ_r^2 and δ_r are finite-difference operators, defined by

$$\delta_r^2 w_{kj} = w_{k+1,j} - 2w_{kj} + w_{k-1,j} \quad \text{and} \quad \delta_r w_{kj} = w_{k+1,j} - w_{k-1,j}. \tag{2.103}$$

For grid points (r_0, t_j), we use the same difference equations with $w_{-1,j}$ eliminated using (2.101), i.e.

$$\frac{w_{0,j+1}-w_{0,j}}{\tau} - \frac{1}{2h^2}\left(1+\frac{h}{2r_0}\right)(w_{1,j+1} - w_{0,j+1} + w_{1,j} - w_{0,j}) = 0 \tag{2.104}$$

for $j = 0, 1, \dots, M-1$.

The local truncation error of the approximation (2.102) can be shown to be $O(\tau^2 + h^2)$.

Example 2.4. One of the most well-known equation in mathematical finance is the the Black-Scholes equation:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0. \tag{2.105}$$

Here V is the value of the option, σ and r are constants. V depends on the current value S of the underlying asset and on time t.

When this equation describes European call option, it is solved subject to the following conditions:

$$V(0,t) = 0 \quad \text{for} \quad t < T, \tag{2.106}$$
$$V(S,t) \to S - Ee^{-r(T-t)} \text{ as } S \to \infty\,, \tag{2.107}$$
$$V(S,T) = \max\{S-E, 0\}. \tag{2.108}$$

In Eq. (2.108), E is a given positive constant. The problem is to solve Eq. (2.105) backwards in time, i.e. for $t < T$. In principle, Eq. (2.105) can be transformed to the heat equation on the whole lines and then solved exactly. Our task is to obtain a numerical solution.

It is convenient to introduce a new independent variable $\tilde{t} = T - t$ and a new dependent variable $\tilde{V}(S, \tilde{t}) = V(S, t) - S + Ee^{-r(T-t)}(1 - e^{-S})$. Then problem (2.105)–(2.108) takes the form

$$\frac{\partial \tilde{V}}{\partial \tilde{t}} = \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 \tilde{V}}{\partial S^2} + rS\frac{\partial \tilde{V}}{\partial S} - r\tilde{V} + f(S, \tilde{t}), \tag{2.109}$$

$$\tilde{V}(0, \tilde{t}) = 0 \quad \text{for} \quad t < T, \tag{2.110}$$

$$\tilde{V}(S, \tilde{t}) \to 0 \text{ as } S \to \infty\,, \tag{2.111}$$

$$\tilde{V}(S, 0) = \max\{0, E - S\} - Ee^{-S}. \tag{2.112}$$

Here $f(S, \tilde{t}) = E\left(\frac{1}{2}\sigma^2 S^2 - rS\right) e^{-r\tilde{t}-S}$. Problem (2.109)–(2.112) is an initial boundary value problem for a parabolic equation with variable coefficients.

To solve the problem, we choose sufficiently large S_0 and define the grid points:

$$(S_k, \tilde{t}_j) = (hk, \tau j) \quad \text{for} \quad k = 0, \ldots, N \quad \text{and} \quad j = 0, 1, \ldots, M$$

where $h = S_0/N$ and $\tau = T/M$ is the step length in time $\tilde{t}$.

To make sure that out scheme is stable, we employ the backward-difference formula for $\partial\tilde{V}/\partial\tilde{t}$ and central difference formulae for $\partial^2\tilde{V}/\partial S^2$ and $\partial\tilde{V}/\partial S$. As a result, we have

$$\frac{w_{k,j} - w_{k,j-1}}{\tau} = \frac{1}{2}\sigma^2 S_k^2 \frac{w_{k+1,j} - 2w_{k,j} + w_{k-1,j}}{h^2} + rS_k \frac{w_{k+1,j} - w_{k-1,j}}{2h} - rw_{k,j} + f_{k,j}, \tag{2.113}$$

$$w_{0,j} = 0, \quad w_{N,j} = 0, \tag{2.114}$$

$$w_{k,0} = \max\{0, E - S_k\} - Ee^{-S_k}. \tag{2.115}$$

These equations can be solved using the standard double-sweep method.

2.8 Nonlinear heat equation

Consider the nonlinear heat equation in the form

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(K(x,t,u)\frac{\partial u}{\partial x}\right) + f(x,t,u), \quad 0 < x < L, \quad t > 0, \tag{2.116}$$

subject to the initial and boundary conditions

$$u(x,0) = u_0(x), \tag{2.117}$$

$$u(0,t) = 0, \quad u(L,t) = 0. \tag{2.118}$$

To develop a finite-difference approximation for (2.116), we need a finite-difference formula for

$$\frac{d}{dx}\left(Q(x)\frac{dg}{\partial x}\right).$$

Let

$$x_{k\pm\frac{1}{2}} = x_k \pm \frac{h}{2} \quad \text{and} \quad G(x) = Q(x)\frac{dg}{\partial x}.$$

Then, on using central difference formula for derivative, we obtain

$$\begin{aligned}\frac{dG}{dx}(x_k) &= \frac{G(x_{k+\frac{1}{2}}) - G(x_{k-\frac{1}{2}})}{h} + O(h^2)\\ &= \frac{1}{h}\left(Q(x_{k+\frac{1}{2}})\frac{dg}{dx}(x_{k+\frac{1}{2}}) - Q(x_{k-\frac{1}{2}})\frac{dg}{dx}(x_{k-\frac{1}{2}})\right) + O(h^2).\end{aligned}$$

Applying central difference formulae $g'(x_{k\pm\frac{1}{2}})$, we find that

$$\frac{dg}{dx}(x_{k+\frac{1}{2}}) = \frac{g(x_{k+1}) - g(x_k)}{h} + O(h^2), \quad \frac{dg}{dx}(x_{k-\frac{1}{2}}) = \frac{g(x_k) - g(x_{k-1})}{h} + O(h^2).$$

It follows that

$$\begin{aligned}\frac{d}{dx}\left(Q(x)\frac{dg}{\partial x}\right)\Big|_{x=x_k} =& \frac{1}{h^2}\Big(Q(x_{k+\frac{1}{2}})\left[g(x_{k+1}) - g(x_k)\right]\\ &- Q(x_{k-\frac{1}{2}})\left[g(x_k) - g(x_{k-1})\right]\Big) + O(h).\end{aligned} \tag{2.119}$$

In fact, the error of formula (2.119) is $O(h^2)$ rather than $O(h)$. This can be verified by expanding all functions in (2.119) in Taylor's series at point x_k.

Note that within the error of $O(h^2)$ the quantities $Q(x_{k\pm\frac{1}{2}})$ in Eq. (2.119) can be replaced by

$$\frac{1}{2}\left[Q(x_k) + Q(x_{k\pm1})\right].$$

With the help of the forward difference formula for u_t and Eq. (2.119), we construct the following finite-difference approximation for Eq. (2.116):

$$\begin{aligned}\frac{w_{k,j+1} - w_{kj}}{\tau} - \frac{1}{h^2}\Big(\varkappa_{k+\frac{1}{2},j}\left[w_{k+1,j} - w_{k,j}\right] - \varkappa_{k-\frac{1}{2},j}\left[w_{k,j} - w_{k-1,j}\right]\Big)&\\ = f(x_k, t_j, w_{kj}),&\end{aligned} \tag{2.120}$$

where

$$\varkappa_{k\pm\frac{1}{2},j} \equiv \frac{1}{2}\left[K(x_k,t_j,w_{kj}) + K(x_{k\pm1},t_j,w_{k\pm1,j})\right].$$

The local truncation error of the difference equation (2.120) is $O(\tau + h^2)$. The difference method (2.120) is explicit. As we already know, explicit methods for linear problems are only conditionally stable. In nonlinear problem, *stability depends not only on the form of the finite difference equations but also generally upon the solution being obtained*, i.e. equation may be stable for some values of t and not for others. This fact leads to a strong restriction on the step size in time which, in turn, makes the method inefficient.

As was previously observed, implicit methods for linear problems have certain stability advantages. It is therefore natural to turn to implicit methods in seeking to avoid the restrictions on the time step. We will discuss only the method which is obtained from (2.120) by replacing the forward-difference formula for u_t with the backward-difference formula. This yields the implicit formula:

$$\frac{w_{k,j} - w_{k,j-1}}{\tau} - \frac{1}{h^2}\left(\varkappa_{k+\frac{1}{2},j}\left[w_{k+1,j} - w_{k,j}\right] - \varkappa_{k-\frac{1}{2},j}\left[w_{k,j} - w_{k-1,j}\right]\right) = f(x_k, t_j, w_{kj}), \quad (2.121)$$

whose local truncation error is $O(\tau + h^2)$. We can write it in the vector form as

$$A\left(\mathbf{w}_j\right)\mathbf{w}_j = \mathbf{w}_{j-1} + \tau\mathbf{F}_j \quad \text{for} \quad j = 1, 2, \ldots, \quad (2.122)$$

where

$$\mathbf{w}_j = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ w_{N-1,j} \end{bmatrix}, \quad \mathbf{F}_j = \begin{bmatrix} f(x_1, t_j, w_{1,j}) \\ f(x_2, t_j, w_{2,j}) \\ \vdots \\ \vdots \\ \vdots \\ f(x_{N-1}, t_j, w_{N-1,j}) \end{bmatrix}$$

and

$$A = \begin{bmatrix} a_1 & b_1 & 0 & \ldots & \ldots & 0 \\ b_1 & a_2 & b_2 & \ddots & & \vdots \\ 0 & b_2 & a_3 & b_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & b_{N-2} \\ 0 & \ldots & \ldots & 0 & b_{N-2} & a_{N-1} \end{bmatrix}$$

$$a_k = 1 + \frac{\tau}{h^2}(\varkappa_{k+\frac{1}{2},j} + \varkappa_{k-\frac{1}{2},j}), \quad b_k = -\frac{\tau}{h^2}\varkappa_{k+\frac{1}{2},j}.$$

Equations (2.121) and (2.122) represent the system of nonlinear algebraic equations which is difficult to solve. We can simplify Eq. (2.122) by replacing $\varkappa_{k\pm\frac{1}{2},j}$ with $\varkappa_{k\pm\frac{1}{2},j-1}$ and $f(x_k, t_j, w_{kj})$ with $f(x_k, t_{j-1}, w_{k,j-1})$. This results in the formula

$$A\left(\mathbf{w}_{j-1}\right)\mathbf{w}_j = \mathbf{w}_{j-1} + \tau\mathbf{F}_{j-1} \quad \text{for} \quad j = 1, 2, \ldots,$$

or, equivalently,

$$\frac{w_{k,j} - w_{k,j-1}}{\tau} - \frac{1}{h^2}\left(\varkappa_{k+\frac{1}{2},j-1}\left[w_{k+1,j} - w_{k,j}\right] - \varkappa_{k-\frac{1}{2},j-1}\left[w_{k,j} - w_{k-1,j}\right]\right) = f(x_k, t_{j-1}, w_{k,j-1}). \quad (2.123)$$

Equations (2.123) are linear in $w_{k-1,j}$, $w_{k,j}$ and $w_{k+1,j}$ and can be solved by the double-sweep method. The local truncation error of Eq. (2.123) is $O(\tau + h^2)$. However, practical computations show that the real accuracy of numerical solutions obtained using Eq. (2.123) is considerably lower that the accuracy achieved by using the iterative methods of solving the nonlinear equations (2.121). We will describe two iterative methods.

The first is called the **method of successive approximations**. In this method, we compute a sequence of numbers $w_{kj}^{(s)}$ $(s = 0, 1, \dots)$ at each time step. As an initial approximation, we take the solution at the previous time step:

$$w_{kj}^{(0)} = w_{k,j-1}.$$

Then, successive approximations are computed using the formula

$$w_{k,j}^{(s)} - \frac{\tau}{h^2}\left(\varkappa_{k+\frac{1}{2},j}^{(s-1)}\left[w_{k+1,j}^{(s)} - w_{k,j}^{(s)}\right] - \varkappa_{k-\frac{1}{2},j}^{(s-1)}\left[w_{k,j}^{(s)} - w_{k-1,j}^{(s)}\right]\right) = w_{k,j-1} + \tau f(x_k, t_j, w_{kj}^{(s-1)}), \quad (2.124)$$

or, in vector form,

$$\mathbf{w}_j^{(0)} = \mathbf{w}_{j-1}, \quad A\left(\mathbf{w}_j^{(s-1)}\right)\mathbf{w}_j^{(s)} = \mathbf{w}_{j-1} + \tau \mathbf{F}_j^{(s-1)}$$

for $s = 1, 2, \dots$ If we perform only one iteration, this method is equivalent to the 'linear' method (2.123). If successive approximations converge to the solution $w_{k,j}$ of the nonlinear system (2.121) (note that it may diverge), then the convergence is linear, i.e.

$$\|\mathbf{w}_j^{(s)} - \mathbf{w}_j\| = O\left(\|\mathbf{w}_j^{(s-1)} - \mathbf{w}_j\|\right) \quad \text{as} \quad s \to \infty.$$

Here $\mathbf{w}_j^{(s)} = (w_{1,j}^{(s)}, w_{2,j}^{(s)}, \dots, w_{N-1,j}^{(s)})$ and $\mathbf{w}_j = (w_{1,j}, w_{2,j}, \dots, w_{N-1,j})$ are $N-1$-dimensional vectors, $\|\cdot\|$ is any vector norm and $\mathbf{w}_j$ represents the exact solution of the nonlinear system (2.121).

The other method is the **Newton method**. Suppose that we have a system of nonlinear equations

$$\Phi_i(x_1, x_2, \dots, x_n) = 0, \quad i = 1, 2, \dots, n, \quad (2.125)$$

for n unknowns $x_1, x_2, \dots, x_n$. In the Newton method, we compute a sequence of approximations $\mathbf{x}^{(s)}$ $(s = 0, 1, 2, \dots)$ to the solution $\mathbf{x}^{(s)} = (x_1, x_2, \dots, x_n)$ of equations (2.125) using the formula

$$\mathbf{x}^{(s)} = \mathbf{x}^{(s-1)} + \mathbf{r}^{(s)}, \quad (2.126)$$

where $\mathbf{r}^{(s)}$ is the solution of the linear system

$$J\left(\mathbf{x}^{(s-1)}\right)\mathbf{r}^{(s)} = -\mathbf{\Phi}\left(\mathbf{x}^{(s-1)}\right) \quad (2.127)$$

with

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial \Phi_1}{\partial x_1} & \frac{\partial \Phi_1}{\partial x_2} & \cdots & \frac{\partial \Phi_1}{\partial x_n} \\ \frac{\partial \Phi_2}{\partial x_1} & \frac{\partial \Phi_2}{\partial x_2} & \cdots & \frac{\partial \Phi_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \Phi_n}{\partial x_1} & \frac{\partial \Phi_n}{\partial x_2} & \cdots & \frac{\partial \Phi_n}{\partial x_n} \end{pmatrix}. \quad (2.128)$$

The sequence of approximations generated by the Newton method converges if the initial approximation $\mathbf{x}^{(0)}$ is sufficiently close to the solution, and the convergence is quadratic

$$\|\mathbf{x}^{(s)} - \mathbf{x}\| = O\left(\|\mathbf{x}^{(s-1)} - \mathbf{x}\|^2\right) \quad \text{as} \quad s \to \infty.$$

Applying the Newton method to the nonlinear equations (2.121), we obtain the equations

$$w_{kj}^{(s)} = w_{kj}^{(s-1)} + r_{kj}^{(s)}, \tag{2.129}$$

where $r_{kj}^{(s)}$ is the solution of the linear system

$$\begin{aligned}
&r_{k+1,j}^{(s)} \left[\varkappa_{k+\frac{1}{2},j}^{(s-1)} + \frac{\partial \varkappa_{k+\frac{1}{2},j}^{(s-1)}}{\partial w_{k+1,j}^{(s-1)}} \left(w_{k+1,j}^{(s-1)} - w_{k,j}^{(s-1)} \right) \right] \\
&-r_{k,j}^{(s)} \left[\frac{h^2}{\tau} + \varkappa_{k+\frac{1}{2},j}^{(s-1)} + \varkappa_{k-\frac{1}{2},j}^{(s-1)} - \frac{\partial \varkappa_{k+\frac{1}{2},j}^{(s-1)}}{\partial w_{k,j}^{(s-1)}} \left(w_{k+1,j}^{(s-1)} - w_{k,j}^{(s-1)} \right) + \right. \\
&\left. + \frac{\partial \varkappa_{k-\frac{1}{2},j}^{(s-1)}}{\partial w_{k,j}^{(s-1)}} \left(w_{k,j}^{(s-1)} - w_{k-1,j}^{(s-1)} \right) - h^2 \frac{\partial f(x_k, t_j, w_{kj}^{(s-1)})}{\partial w_{k,j}^{(s-1)}} \right] + \\
&+r_{k-1,j}^{(s)} \left[\varkappa_{k-\frac{1}{2},j}^{(s-1)} + \frac{\partial \varkappa_{k-\frac{1}{2},j}^{(s-1)}}{\partial w_{k-1,j}^{(s-1)}} \left(w_{k,j}^{(s-1)} - w_{k-1,j}^{(s-1)} \right) \right] = \\
&= \frac{h^2}{\tau} \left(w_{k,j}^{(s-1)} - w_{k,j-1} \right) - \varkappa_{k+\frac{1}{2},j}^{(s-1)} \left[w_{k+1,j}^{(s-1)} - w_{k,j}^{(s-1)} \right] + \\
&\quad + \varkappa_{k-\frac{1}{2},j}^{(s-1)} \left[w_{k,j}^{(s-1)} - w_{k-1,j}^{(s-1)} \right] - h^2 f(x_k, t_j, w_{kj}^{(s-1)}).
\end{aligned} \tag{2.130}$$

Equations (2.130) look very complicated, but can be solved by the double sweep method. In general, the method based on equations (2.129)–(2.130) produces sequences that converge much faster than corresponding sequences for the method of successive approximations.

Note that if $K = const$ in Eq. (2.116), then Eq. (2.130) reduces to

$$\begin{aligned}
&-\gamma r_{k-1,j}^{(s)} - \gamma r_{k+1,j}^{(s)} + \left(1 + 2\gamma - \tau \frac{\partial f(x_k, t_j, w_{kj}^{(s-1)})}{\partial w_{kj}^{(s-1)}} \right) r_{k,j}^{(s)} = \\
&= -(1+2\gamma) w_{kj}^{(s-1)} + \gamma (w_{k+1,j}^{(s-1)} + w_{k-1,j}^{(s-1)}) + \tau f(x_k, t_j, w_{kj}^{(s-1)}) + w_{k,j-1}
\end{aligned} \tag{2.131}$$

where $\gamma = K\tau/h^2$.

2.9 Two-dimensional heat equation.

Consider the two-dimensional heat equation

$$\frac{\partial u}{\partial t} = K \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + f(x, y, t) \tag{2.132}$$

to be solved for $0 < t < T$ and in a connected region $\mathcal{D}$ of the (x, y) plane. Let S be the boundary of $\mathcal{D}$. We suppose that

$$u(x, y, t) = g(x, y, t) \quad \text{on} \quad S \tag{2.133}$$

and

$$u(x, y, 0) = u_0(x, y), \tag{2.134}$$

where $g(x, y, t)$ and $u_0(x, y)$ are given functions. In what follows, $\mathcal{D}$ is the rectangle (the general domain will be treated later):

$$\mathcal{D} = \{ (x, y) \,|\, 0 \le x \le L_1,\ 0 \le y \le L_2 \}.$$

We choose positive integers N_1, N_2 and M and define the grid points

$$(x_k, y_j, t_n) = (kh_1, jh_2, \tau n)$$

for $k = 0, 1, \ldots, N_1, \quad j = 0, 1, \ldots, N_2, \quad n = 0, 1, \ldots, M$, where

$$h_1 = \frac{L_1}{N_1}, \quad h_2 = \frac{L_2}{N_2}, \quad \tau = \frac{T}{M}.$$

Let w_{kj}^n be the discrete approximation to $u_{kj}^n \equiv u(x_k, y_j, t_n)$. Employing the central difference formula for u_{xx} and u_{yy} and the two-point forward-difference formula for $u_t t$, we obtain the following difference equation

$$\frac{w_{kj}^{n+1} - w_{kj}^n}{\tau} - K\left(\frac{\delta_x^2}{h_1^2} + \frac{\delta_y^2}{h_2^2}\right) w_{kj}^n = f_{kj}^n, \tag{2.135}$$

where $f_{kj}^n \equiv f(x_k, y_j, t_n)$ and

$$\delta_x^2 w_{kj}^n = w_{k+1,j}^n - 2w_{kj}^n + w_{k-1,j}^n, \quad \delta_y^2 w_{kj}^n = w_{k,j+1}^n - 2w_{kj}^n + w_{k,j-1}^n.$$

Equation (2.135) is a straightforward generalisation of the explicit forward difference scheme for the one-dimensional heat equation. We will discuss only the case of homogeneous (zero) boundary conditions when function $g(x, y, t)$ in Eq. (2.133) is identically zero. (A problem with non-zero boundary conditions can be reduced to the problem with zero conditions in the same manner as it was done for the one-dimensional heat equation.) Then, we have

$$w_{0,j}^n = w_{N_1,j}^n = 0 \quad (j = 0, 1, \ldots, N_2), \quad w_{k,0}^n = w_{k,N_2}^n = 0 \quad (k = 0, 1, \ldots, N_1), \tag{2.136}$$

and

$$w_{k,j}^0 = u_0(x_k, y_j). \tag{2.137}$$

Since the forward difference formula for u_t has the truncation error $O(\tau)$ and the central difference formulae for u_{xx} and u_{yy} have errors $O(h_1^2)$ and $O(h_2^2)$, respectively, the local truncation error of Eq. (2.135) is $O(\tau + h_1^2 + h_2^2)$.

The stability of the method (2.135) can be studied using the Fourier method. Let $w_{k,j}^n$ and $\tilde{w}_{k,j}^n$ be two solutions of Eqs. (2.135) and (2.136) corresponding to slightly different initial conditions and let $z_{k,j}^n = w_{k,j}^n - \tilde{w}_{k,j}^n$ be the perturbation at the grid point (x_k, y_j, t_n) for each $k = 0, 1, 2, \ldots, N_1$, $j = 0, 1, 2, \ldots, N_2$ and $n = 0, 1, \ldots, M$. Then z_{kj}^n satisfies the difference equation

$$\frac{z_{kj}^{n+1} - z_{kj}^n}{\tau} - K\left(\frac{\delta_x^2}{h_1^2} + \frac{\delta_y^2}{h_2^2}\right) z_{kj}^n = 0, \tag{2.138}$$

which is the homogeneous version of Eq. (2.135). We seek a particular solution of (2.138) in the form

$$z_{k,j}^n = \rho^n e^{iqx_k + ipy_j}. \tag{2.139}$$

for $q, p \in \mathbb{R}$ and $n = 0, 1, \ldots$ The finite-difference method is stable with respect to initial condition, if

$$|\rho| \leq 1 \quad \text{for all} \quad q, p \in \mathbb{R}.$$

Substituting (2.139) in (2.138), we obtain

$$e^{iqx_k + ipy_j}\left(\rho^{n+1} - \rho^n\right) - \frac{K\tau}{h_1^2}\rho^n e^{ipy_j}\left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right) - \frac{K\tau}{h_2^2}\rho^n e^{iqx_k}\left(e^{ipy_{j+1}} - 2e^{ipy_j} + e^{ipy_{j-1}}\right) = 0$$

or, equivalently,

$$\rho - 1 - \frac{K\tau}{h_1^2}\left(e^{iqh_1} - 2 + e^{-iqh_1}\right) - \frac{K\tau}{h_2^2}\left(e^{iph_2} - 2 + e^{-iph_2}\right) = 0.$$

Since

$$e^{iqh_1} - 2 + e^{-iqh_1} = -4\sin^2\frac{qh_1}{2}, \quad e^{iph_2} - 2 + e^{-iph_2} = -4\sin^2\frac{ph_2}{2},$$

we obtain

$$\rho = 1 - 4\gamma_1 \sin^2\frac{qh_1}{2} - 4\gamma_2 \sin^2\frac{ph_2}{2}.$$

where

$$\gamma_1 = \frac{K\tau}{h_1^2}, \quad \gamma_2 = \frac{K\tau}{h_2^2}.$$

It follows that $|\rho| \leq 1$ if $-1 \leq \rho$ or

$$-1 \leq 1 - 4\gamma_1 \sin^2\frac{qh_1}{2} - 4\gamma_2 \sin^2\frac{ph_2}{2}$$

The last inequality holds for all p and q provided that $\gamma_1 + \gamma_2 \leq 1/2$ or

$$K\tau\left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right) \leq \frac{1}{2}. \tag{2.140}$$

Thus, the method (2.135) is conditionally stable. Note that if $h_1 = h_2 = h$, then the stability condition (2.140) becomes $K\tau/h^2 \leq 1/4$. With m space variables and equal space step sizes the stability condition becomes $K\tau/h^2 \leq 1/2m$.

To avoid stability problems, we can employ the backward difference formula to approximate u_t. This yields the implicit scheme

$$\frac{w_{kj}^n - w_{kj}^{n-1}}{\tau} - K\left(\frac{\delta_x^2}{h_1^2} + \frac{\delta_y^2}{h_2^2}\right) w_{kj}^n = f_{kj}^n, \tag{2.141}$$

which has the same local truncation error and is unconditionally stable. If we let

$$\mathbf{w}^n = \begin{bmatrix} w_{1,1}^n \\ w_{2,1}^n \\ \vdots \\ w_{N_1-1,1}^n \\ w_{1,2}^n \\ w_{2,2}^n \\ \vdots \\ w_{N_1-1,2}^n \\ w_{1,3}^n \\ \vdots \\ \vdots \\ w_{N_1-1,N_2-1}^n \end{bmatrix}, \quad \mathbf{F}^n = \begin{bmatrix} f_{1,1}^n \\ f_{2,1}^n \\ \vdots \\ f_{N_1-1,1}^n \\ f_{1,2}^n \\ f_{2,2}^n \\ \vdots \\ f_{N_1-1,2}^n \\ f_{1,3}^n \\ \vdots \\ \vdots \\ f_{N_1-1,N_2-1}^n \end{bmatrix}, \tag{2.142}$$

then the linear system (2.141) can be written in the matrix form

$$A\mathbf{w}^n = \mathbf{w}^{n-1} + \tau\mathbf{F}^n, \tag{2.143}$$

where A is the $(N_1-1)(N_2-1)\times(N_1-1)(N_2-1)$ matrix given by

$$A=\begin{bmatrix} Q & -\gamma_2 I & 0 & \dots & \dots & 0 \\ -\gamma_2 I & Q & -\gamma_2 I & \ddots & & \vdots \\ 0 & -\gamma_2 I & Q & -\gamma_2 I & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -\gamma_2 I \\ 0 & \dots & \dots & 0 & -\gamma_2 I & Q \end{bmatrix}, \tag{2.144}$$

and where I is the identity matrix and Q is the $(N_1-1)\times(N_1-1)$ matrix having the form

$$Q=\begin{bmatrix} 1+2(\gamma_1+\gamma_2) & -\gamma_1 & 0 & \dots & 0 \\ -\gamma_1 & 1+2(\gamma_1+\gamma_2) & -\gamma_1 & \ddots & \vdots \\ 0 & -\gamma_1 & 1+2(\gamma_1+\gamma_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -\gamma_1 \\ 0 & \dots & 0 & -\gamma_1 & 1+2(\gamma_1+\gamma_2) \end{bmatrix}. \tag{2.145}$$

Here

$$\gamma_1=\frac{K\tau}{h_1^2},\quad \gamma_2=\frac{K\tau}{h_2^2}.$$

The linear algebraic system (2.143) has the block-tridiagonal matrix A, but, unfortunately, it is not tridiagonal, which makes it difficult to solve if the step sizes are small.

The Crank-Nicolson scheme can be extended to the two-dimensional problem in the form

$$\frac{w_{kj}^{n+1}-w_{kj}^n}{\tau}-\frac{K}{2}\left(\frac{1}{h_1^2}\delta_x^2+\frac{1}{h_2^2}\delta_y^2\right)\left(w_{kj}^n+w_{kj}^{n+1}\right)=f_{kj}^{n+\frac{1}{2}}, \tag{2.146}$$

where

$$f_{kj}^{n+\frac{1}{2}}=f(x_k,y_j,t_n+\tau/2)=\frac{f(x_k,y_j,t_n)+f(x_k,y_j,t_{n+1})}{2}+O(\tau^2).$$

This method is always stable, and its local truncation error is $O(\tau^2+h_1^2+h_2^2)$. Again, the linear system (2.146) is not tridiagonal.

The alternating-direction implicit (ADI) method. The ADI method is intended to simplify the solution of the algebraic equations while preserving the stability and accuracy requirements. The main idea of the method is a reformulation of the finite difference equations so that the algebraic problem consists of a set of linear equations possessing a tridiagonal matrix. We then solve this set of equations in each coordinate direction in turn by the double-sweep method. We illustrate the basic concepts for the case of equal step sizes in the x and y directions: $h_1=h_2=h$.

The idea of the ADI method is to divide each time step into two steps of size $\tau/2$. In each substep, a different dimension is treated implicitly:

$$\frac{w_{k,j}^{n+\frac{1}{2}}-w_{k,j}^n}{\tau}=\frac{K}{2h^2}\left(\delta_x^2 w_{k,j}^{n+\frac{1}{2}}+\delta_y^2 w_{k,j}^n\right)+\frac{1}{2}f_{k,j}^{n+\frac{1}{2}}, \tag{2.147}$$

$$\frac{w_{k,j}^{n+1}-w_{k,j}^{n+\frac{1}{2}}}{\tau}=\frac{K}{2h^2}\left(\delta_x^2 w_{k,j}^{n+\frac{1}{2}}+\delta_y^2 w_{k,j}^{n+1}\right)+\frac{1}{2}f_{k,j}^{n+\frac{1}{2}}. \tag{2.148}$$

The advantage of this method is that each substep requires only the solution of a simple tridiagonal system.

To find the local truncation error of the ADI method, we first eliminate the intermediate values from Eqs. (2.147), (2.148). Adding the two equations, we obtain

$$\frac{w_{k,j}^{n+1}-w_{k,j}^{n}}{\tau}=\frac{K}{2h^2}\left(2\delta_x^2 w_{k,j}^{n+\frac{1}{2}}+\delta_y^2\left[w_{k,j}^{n}+w_{k,j}^{n+1}\right]\right)+f_{k,j}^{n+\frac{1}{2}}.$$

Subtracting (2.148) from (2.147), we find that

$$\frac{2}{\tau}w_{k,j}^{n+\frac{1}{2}}=\frac{w_{k,j}^{n+1}+w_{k,j}^{n}}{\tau}+\frac{K}{2h^2}\delta_y^2\left[w_{k,j}^{n}-w_{k,j}^{n+1}\right].$$

It follows that

$$\frac{w_{k,j}^{n+1}-w_{k,j}^{n}}{\tau}=\frac{K}{2h^2}\left(\delta_x^2+\delta_y^2\right)\left(w_{k,j}^{n}+w_{k,j}^{n+1}\right)+f_{k,j}^{n+\frac{1}{2}}+\frac{K^2\tau}{4h^4}\delta_x^2\delta_y^2\left[w_{k,j}^{n}-w_{k,j}^{n+1}\right]. \tag{2.149}$$

If the last term on the right side of this equation were absent, the equation would coincide with the Crank-Nicolson method whose local truncation error is $O(\tau^2+h^2)$.

We will show that the last term in (2.149), evaluated on the exact solution $u(x,y,t)$, is $O(\tau^2)$. To do this, we first observe that

$$\begin{aligned}\frac{1}{h^2}\delta_x^2\,u_{kj}^{n}&=u_{xx}(x_k,y_j,t_n)+\frac{h^2}{12}\,u_{xxxx}(x_k,y_j,t_n)+O(h^4),\\ \frac{1}{h^2}\delta_y^2\,u_{kj}^{n}&=u_{yy}(x_k,y_j,t_n)+\frac{h^2}{12}\,u_{yyyy}(x_k,y_j,t_n)+O(h^4).\end{aligned} \tag{2.150}$$

It follows from (2.150) that

$$\begin{aligned}\frac{1}{h^4}\delta_x^2\delta_y^2\,u_{k,j}^{n}&=u_{xxyy}(x_k,y_j,t_n)+O(h^2),\\ \frac{1}{h^4}\delta_x^2\delta_y^2\,u_{k,j}^{n+1}&=u_{xxyy}(x_k,y_j,t_{n+1})+O(h^2).\end{aligned}$$

Further, we have

$$\begin{aligned}\frac{1}{h^4}\delta_x^2\delta_y^2\left(u_{k,j}^{n}-u_{k,j}^{n+1}\right)&=u_{xxyy}(x_k,y_j,t_n)-u_{xxyy}(x_k,y_j,t_{n+1})+O(h^2),\\ &=-\tau u_{xxyyt}(x_k,y_j,t_n)+O(\tau^2)+O(h^2).\end{aligned} \tag{2.151}$$

Hence,

$$\frac{K^2\tau}{4h^4}\,\delta_x^2\delta_y^2\left(u_{k,j}^{n}-u_{k,j}^{n+1}\right)=\frac{K^2\tau^2}{4}\left[-u_{xxyyt}(x_k,y_j,t_n)+O(\tau)\right]+O(\tau h^2)=O(\tau^2).$$

Therefore, the local truncation error of the ADI method is $O(\tau^2+h^2)$.

Let us investigate the stability of the ADI method. If $z_{k,j}^{n}=w_{k,j}-\tilde{w}_{k,j}$ is the perturbation at the grid point (x_k,y_j,t_n) for each $k,j=0,1,2,\ldots,N$ and $n=0,1,\ldots$, then it satisfies the difference equations

$$\begin{aligned}z_{k,j}^{n+\frac{1}{2}}&=z_{k,j}^{n}+\frac{\gamma}{2}\left(\delta_x^2 z_{k,j}^{n+\frac{1}{2}}+\delta_y^2 z_{k,j}^{n}\right),\\ z_{k,j}^{n+1}&=z_{k,j}^{n+\frac{1}{2}}+\frac{\gamma}{2}\left(\delta_x^2 z_{k,j}^{n+\frac{1}{2}}+\delta_y^2 z_{k,j}^{n+1}\right),\end{aligned} \tag{2.152}$$

for $k = 1, 2, \ldots, N$ and $j = 1, 2, \ldots$. We seek a particular solution of (2.152) in the form

$$z_{k,j}^n = \rho^n e^{iqx_k + ipy_j}$$

for $q, p \in \mathbb{R}$ and $n = 0, 1, \ldots$ The method is stable, if $|\rho| \leq 1$ for all $q, p \in \mathbb{R}$.

To deal with the substeps, we assume that

$$z_{k,j}^{n+\frac{1}{2}} = z_{k,j}^n \rho^{(1)} \quad \text{and} \quad z_{k,j}^{n+1} = z_{k,j}^{n+\frac{1}{2}} \rho^{(2)}$$

(so that $\rho = \rho^{(1)}\rho^{(2)}$). Substituting these in (2.152), we find that

$$\begin{aligned}
\rho^{(1)} &= 1 + \frac{\gamma}{2}\rho^{(1)}\left(e^{iqh} - 2 + e^{-iqh}\right) + \frac{\gamma}{2}\left(e^{iph} - 2 + e^{-iph}\right), \\
\rho^{(2)} &= 1 + \frac{\gamma}{2}\left(e^{iqh} - 2 + e^{-iqh}\right) + \rho^{(2)}\frac{\gamma}{2}\left(e^{iph} - 2 + e^{-iph}\right).
\end{aligned}$$

Since

$$e^{iqh} - 2 + e^{-iqh} = \left(e^{iqh/2} - e^{-iqh/2}\right)^2 = -4\sin^2\frac{qh}{2},$$

we obtain

$$\begin{aligned}
\rho^{(1)} &= \frac{1 - 2\gamma\sin^2\frac{ph}{2}}{1 + 2\gamma\sin^2\frac{qh}{2}}, \\
\rho^{(2)} &= \frac{1 - 2\gamma\sin^2\frac{qh}{2}}{1 + 2\gamma\sin^2\frac{ph}{2}}.
\end{aligned}$$

It follows that

$$\rho = \rho^{(1)}\rho^{(2)} = \frac{\left(1 - 2\gamma\sin^2\frac{ph}{2}\right)\left(1 - 2\gamma\sin^2\frac{qh}{2}\right)}{\left(1 + 2\gamma\sin^2\frac{ph}{2}\right)\left(1 + 2\gamma\sin^2\frac{qh}{2}\right)}.$$

Evidently, $|\rho| \leq 1$ and, therefore, the method is unconditionally stable. (Note that for some p, q and γ, we can have $|\rho^{(1)}| > 1$. This, however, is compensated by $|\rho^{(2)}| < 1$, yielding $|\rho| = |\rho^{(1)}\rho^{(2)}| \leq 1$ for all p, q and γ.)

3 Elliptic partial differential equations

3.1 Poisson equation

We will illustrate the finite-difference method for elliptic PDEs with the Poisson equation

$$\Delta u = f(x, y) \tag{3.1}$$

for $(x, y) \in \mathcal{D}$ and

$$u(x, y) = g(x, y) \quad \text{for} \quad (x, y) \in S, \tag{3.2}$$

where

$$\Delta \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \quad \mathcal{D} = \{\, (x, y) \,|\, a < x < b,\ c < y < d \,\},$$

and S is the boundary of $\mathcal{D}$. In what follows, we assume that both f and g are continuous in $\mathcal{D}$ and that a unique solution of the boundary-value problem (3.1), (3.2) exists.

The first step in the finite-difference method is to choose integers N_1 and N_2 and define step sizes h_1 and h_2 by $h_1 = (b-a)/N_1$ and $h_2 = (d-c)/N_2$. Partitioning of the interval $[a, b]$ into N_1 equal parts of width h_1 and the interval $[c, d]$ into N_2 equal parts of width h_2 provides a grid on the rectangle $\mathcal{D}$ by drawing vertical and horizontal lines through the points with coordinates (x_k, y_j) where $x_k = a + kh_1$ for each $k = 0, 1, \ldots, N_1$ and $y_j = c + jh_2$ for each $j = 0, 1, \ldots, N_2$.

Let w_{kj} be an approximation to $u(x_k, y_j)$. For the interior mesh points (x_k, y_j) (with $k = 1, 2, \ldots, N_1 - 1$ and $j = 1, 2, \ldots, N_2 - 1$), we approximate the second partial derivatives with respect to x and y by the three-point central-difference formulas. As a result, we obtain the following difference equations:

$$\frac{w_{k+1,j} - 2w_{k,j} + w_{k-1,j}}{h_1^2} + \frac{w_{k,j+1} - 2w_{k,j} + w_{k,j-1}}{h_2^2} = f_{kj}$$

for each $k = 1, 2, \ldots, N_1 - 1$ and $j = 1, 2, \ldots, N_2 - 1$. Here $f_{kj} = f(x_k, y_j)$. Equivalently, we can rewrite this as

$$2\left[1 + \left(\frac{h_1}{h_2}\right)^2\right] w_{kj} - (w_{k+1,j} + w_{k-1,j}) - \left(\frac{h_1}{h_2}\right)^2 (w_{k,j+1} + w_{k,j-1}) = -h_1^2 f_{kj}, \tag{3.3}$$

for each $k = 1, 2, \ldots, N_1 - 1$ and $j = 1, 2, \ldots, N_2 - 1$, The boundary conditions at the exterior mesh points yield

$$\begin{aligned} &w_{0,j} = g(x_0, y_j), \quad w_{N_1,j} = g(x_{N_1}, y_j) \quad \text{for each} \quad j = 1, \ldots, N_2 - 1 \\ &w_{k,0} = g(x_k, y_0), \quad w_{k,N_2} = g(x_k, y_{N_2}) \quad \text{for each} \quad k = 1, 2, \ldots, N_1 - 1. \end{aligned} \tag{3.4}$$

Note that we did not specify the approximations $w_{0,0}$, $w_{N_1,0}$, w_{N_1,N_2} and w_{0,N_2} corresponding to the vertices of the rectangle $\mathcal{D}$, because these points take no part in Eq. (3.3).

Evidently, the local truncation error of the finite difference method, given by (3.3)–(3.4), is $O(h_1^2 + h_2^2))$.

The typical equation in (3.3) involves approximations to $u(x_k, y_j)$ at the five points

$$(x_{k-1}, y_j), \quad (x_k, y_j), \quad (x_{k+1}, y_j), \quad (x_k, y_{j-1}), \quad (x_k, y_{j+1}).$$

If we use the information from the boundary conditions (3.4) whenever appropriate in the system (3.3), we have an $(N_1 - 1)(N_2 - 1)$ by $(N_1 - 1)(N_2 - 1)$ linear system with the unknowns w_{ij} at the interior mesh points.

For simplicity, consider the square domain ($c = a$, $d = b$) with the same number of grid lines in x and y. Then $N_1 = N_2 \equiv N$ and $h_1 = h_2 \equiv h$. Equation (3.3) takes the form

$$4w_{k,j} - (w_{k+1,j} + w_{k-1,j} + w_{k,j+1} + w_{k,j-1}) = -h^2 f_{k,j}, \tag{3.5}$$

for each $k, j = 1, 2, \ldots, N-1$.

Example. Consider the boundary value problem

$$\begin{aligned} &\Delta u = f(x,y) \quad \text{for} \quad 0 < x < 1, \quad 0 < y < 1; \\ &u(0,y) = 0, \quad u(1,y) = 1 \quad \text{for} \quad 0 < y < 1; \\ &u(x,0) = x, \quad u(x,1) = x \quad \text{for} \quad 0 < x < 1. \end{aligned}$$

Let $N = 3$, so that $h = 1/3$ and $x_1 = y_1 = 1/3$, $x_2 = y_2 = 2/3$. Equations (3.4) become

$$\begin{aligned} &w_{0,1} = w_{0,2} = 0, \quad w_{3,1} = w_{3,2} = 1, \\ &w_{1,0} = w_{1,3} = x_1, \quad w_{2,0} = w_{2,3} = x_2. \end{aligned} \tag{3.6}$$

Equations (3.5) take the form

$$\begin{aligned} 4w_{1,1} - (w_{2,1} + w_{0,1} + w_{1,2} + w_{1,0}) &= -h^2 f_{1,1}, \\ 4w_{2,1} - (w_{3,1} + w_{1,1} + w_{2,2} + w_{2,0}) &= -h^2 f_{2,1}, \\ 4w_{1,2} - (w_{2,2} + w_{0,2} + w_{1,3} + w_{1,1}) &= -h^2 f_{1,2}, \\ 4w_{2,2} - (w_{3,2} + w_{1,2} + w_{2,3} + w_{2,1}) &= -h^2 f_{2,2}, \end{aligned}$$

Substitution of (3.6) in these equations yields the system

$$\begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} w_{1,1} \\ w_{2,1} \\ w_{1,2} \\ w_{2,2} \end{pmatrix} = \begin{pmatrix} x_1 - h^2 f_{1,1} \\ 1 + x_2 - h^2 f_{2,1} \\ x_1 - h^2 f_{1,2} \\ 1 + x_2 - h^2 f_{2,2} \end{pmatrix}. \tag{3.7}$$

Thus, our problem is reduced to solving system (3.7) of 4 linear equations for 4 unknowns.

In general, if we let

$$\mathbf{w} = \begin{bmatrix} w_{1,1} \\ w_{2,1} \\ \vdots \\ w_{N-1,1} \\ w_{1,2} \\ w_{2,2} \\ \vdots \\ w_{N-1,2} \\ w_{1,3} \\ \vdots \\ \vdots \\ w_{N-1,N-1} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} w_{0,1} + w_{1,0} - h^2 f_{1,1} \\ w_{2,0} - h^2 f_{2,1} \\ \vdots \\ w_{N-1,0} + w_{N,1} - h^2 f_{N-1,1} \\ w_{0,2} - h^2 f_{1,2} \\ -h^2 f_{22} \\ \vdots \\ w_{N,2} - h^2 f_{N-1,2} \\ w_{0,3} - h^2 f_{1,3} \\ \vdots \\ \vdots \\ w_{N,N-1} + w_{N-1,N} - h^2 f_{N-1,N-1} \end{bmatrix}, \tag{3.8}$$

then the linear system (3.5) can be written in the matrix form

$$A\mathbf{w} = \mathbf{f}, \tag{3.9}$$

where

$$A = \begin{bmatrix} A_1 & B & 0 & \dots & \dots & 0 \\ B & A_1 & B & \ddots & & \vdots \\ 0 & B & A_1 & B & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & B \\ 0 & \dots & \dots & 0 & B & A_1 \end{bmatrix}, \tag{3.10}$$

and where A_1 and B are $(N-1)\times(N-1)$ matrices having the form

$$A_1 = \begin{bmatrix} 4 & -1 & 0 & \dots & \dots & 0 \\ -1 & 4 & -1 & \ddots & & \vdots \\ 0 & -1 & 4 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & \dots & 0 & -1 & 4 \end{bmatrix}, \quad B = -I = \begin{bmatrix} -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 \end{bmatrix}. \tag{3.11}$$

Thus, we have the linear algebraic system (3.9) with the block-tridiagonal matrix A. Now we need to answer two questions: (i) does system (3.9) have a unique solution and (ii) how to solve (3.9) in practice?

The existence of a unique solution of the linear system $A\mathbf{w} = \mathbf{f}$. The existence of a unique solution of this system is equivalent to non-existence of nontrivial (i.e. nonzero) solutions of the homogeneous equation

$$A\mathbf{w} = 0.$$

To show this, we will use the maximum principle. The homogeneous version of the difference equations derived in the last lecture can be written as

$$4w_{kj} - (w_{k+1,j} + w_{k-1,j} + w_{k,j+1} + w_{k,j-1}) = 0, \tag{3.12}$$

for each $k, j = 1, 2, \dots, N-1$. Equation (3.12) is supplemented with homogeneous boundary conditions

$$w_{0j} = w_{N,j} = 0 \quad \text{for} \quad j = 0, 1, \dots, N \quad \text{and} \quad w_{k0} = w_{k,N} = 0 \tag{3.13}$$

for $k = 1, 2, \dots, N-1$. Let (m, n) be a point at which the maximum of w_{kj} is attained. There may be several points at which the maximum is attained. If the maximum is attained a boundary point, then $\max_{0\le k,j\le N} w_{kj} = 0$. Assume that the maximum is attained at an interior point (m, n). Since, again, there may be several such points, we choose a point (m, n) such than it corresponds to the maximum value of index m, i.e.

$$w_{m,n} = \max_{0\le k,j\le N} w_{kj} \quad \text{and} \quad w_{m,n} > w_{m+1,n}.$$

Then, Eq. (3.12) yields

$$w_{mn} - w_{m+1,n} + (w_{mn} - w_{m-1,n}) + (w_{mn} - w_{m,n+1}) + (w_{mn} - w_{m,n-1}) \\ \ge w_{mn} - w_{m+1,n} > 0. \tag{3.14}$$

Evidently, (3.14) is in contradiction with (3.12). Thus, our assumption that the maximum is attained at an interior mesh point is wrong. Thus, we have

$$\max_{0\le k,j\le N} w_{kj} = 0. \tag{3.15}$$

Similarly, it can be shown that

$$\min_{0\le k,j\le N} w_{kj} = 0. \tag{3.16}$$

It follows from Eqs. (3.15) and (3.16) that $w_{kj} = 0$ for each $k, j = 1, 2, \ldots, N-1$. Thus, the homogeneous system has only zero solution, which proves that system $A\mathbf{w} = \mathbf{f}$ has a unique solution.

The maximum principle can also be used to prove the convergence of the discrete approximations w_{kj} to the solution. Namely, it is shown in Appendix A that if $u \in C^4(D)$ (where $D = \{(x, y) \mid 0 < x < 1,\ 0 < y < 1\}$) is the exact solution of the boundary-value problem

$$u_{xx} + u_{yy} = f(x, y), \quad 0 < x < 1, \quad 0 < y < 1; \tag{3.17}$$

$$u(0, y) = u(1, y) = 0, \quad u(x, 0) = u(x, 1) = 0. \tag{3.18}$$

and w_{kj} $(k, j = 1, 2, \ldots, N-1)$ satisfy

$$4w_{k,j} - (w_{k+1,j} + w_{k-1,j} + w_{k,j+1} + w_{k,j-1}) = -h^2 f_{k,j}, \tag{3.19}$$

for each $k, j = 1, 2, \ldots, N-1$, then

$$|w_{kj} - u(x_k, y_j)| \le Ah^2,$$

where A is independent of h.

Iterative techniques for solving the linear system $A\mathbf{w} = \mathbf{f}$. For large systems ($N \gg 1$), an iterative method (e.g., the Jacobi method, the Gauss-Seidel method or the SOR method) should be used to solve system (3.19). All these methods can be viewed as *relaxation methods*. Consider the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - f(x, y), \tag{3.20}$$

subject to boundary conditions (3.18). An initial distribution $u_0(x, y)$ will *relax* to an equilibrium solution as $t \to \infty$, and this equilibrium solution will be the solution of the original elliptic problem (3.17), (3.18).

Suppose that to solve (3.20) we employ the explicit forward-difference scheme

$$\frac{w_{kj}^{n+1} - w_{kj}^n}{\tau} - \frac{1}{h^2}\left(\delta_x^2 + \delta_y^2\right) w_{kj}^n = -f_{kj}. \tag{3.21}$$

We want to obtain an approximation to $u(x, y, T)$ where T is as large as possible, so we need to choose the step size in time τ at large as possible. We know that the scheme (3.21) is stable only if $\tau/h^2 \le 1/4$, so we choose

$$\tau = \frac{h^2}{4}.$$

Substituting this in (3.21), we can rewrite it in the form

$$w_{kj}^{n+1} = \frac{1}{4}\left(w_{k+1,j}^n + w_{k-1,j}^n + w_{k,j+1}^n + w_{k,j-1}^n\right) - \frac{h^2}{4} f_{kj}. \tag{3.22}$$

This represent the Jacobi iterative technique for solving linear system (3.19).

Another classical method, the Gauss-Seidel method, is obtained from Eq. (3.22) by using w_{kj}^{n+1} as soon as they become available. This leads to the formula

$$w_{kj}^{n+1} = \frac{1}{4}\left(w_{k+1,j}^n + w_{k-1,j}^{n+1} + w_{k,j+1}^n + w_{k,j-1}^{n+1}\right) - \frac{h^2}{4} f_{kj}. \tag{3.23}$$

Both the Jacobi and Gauss-Seidel methods are slowly converging and rarely used in practice. A more appropriate method is the successive over-relaxation (SOR) method.

Another relaxation method for solving problem (3.17)–(3.18) is to employ the alternating direction implicit (ADI) method to solve the two-dimensional heat equation (3.20). We know that this method is unconditionally stable, so that there is no restriction of the step size in time. It can be shown however that the convergence of the solution to the equilibrium solution is slower for large τ, so we cannot let it be too large. There exists an optimal value of τ and it can be shown that for large N it is given by the formula

$$\tau \approx \frac{1}{\pi}\frac{L^2}{N},$$

where L is the size of the (square) domain $\mathcal{D}$. Note that the time step corresponding to the Jacobi method is much smaller (for large N) and is given by

$$\tau = \frac{h^2}{4} = \frac{L^2}{4N^2}.$$

Direct methods for solving the linear system $A\mathbf{w} = \mathbf{f}$. To describe two direct (rather than iterative) methods for solving Eq. (3.19), we rewrite it in the form

$$\begin{aligned}
&A_1\mathbf{v}_1 + B\mathbf{v}_2 = \mathbf{f}_1,\\
&B\mathbf{v}_1 + A_1\mathbf{v}_2 + B\mathbf{v}_3 = \mathbf{f}_2,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&B\mathbf{v}_{k-1} + A_1\mathbf{v}_k + B\mathbf{v}_{k+1} = \mathbf{f}_k,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&B\mathbf{v}_{N-2} + A_1\mathbf{v}_{N-1} = \mathbf{f}_{N-1},
\end{aligned} \tag{3.24}$$

where $\mathbf{v}_k$ is the vector with components $w_{1,k}, w_{2,k}, \ldots, w_{N-1,k}$.

To solve equations (3.24), we can use the matrix form of the double-sweep method described below.

Matrix double-sweep method. Consider the following system:

$$A_i\mathbf{v}_{i-1} - C_i\mathbf{v}_i + B_i\mathbf{v}_{i+1} = \mathbf{F}_i \quad \text{for} \quad i = 1, \ldots, N-1; \quad \mathbf{v}_0 = \mathbf{V}_0, \quad \mathbf{v}_N = \mathbf{V}_N \tag{3.25}$$

where the coefficients A_i, B_i and C_i are given $(N-1)\times(N-1)$ real matrices, $\mathbf{F}_i, \mathbf{V}_0, \mathbf{V}_N \in \mathbb{R}^{N-1}$ are given vectors and $\mathbf{v}_i \in \mathbb{R}^{N-1}$ are unknowns.

Let

$$\mathbf{v}_{i-1} = \alpha_i\mathbf{v}_i + \beta_i \quad \text{for} \quad i = 1, 2, \ldots, N, \tag{3.26}$$

where α_i are $(N-1)\times(N-1)$ real matrices and $\beta_i \in \mathbb{R}^{N-1}$.

Eliminating $\mathbf{v}_{i-1}$, we find

$$(A_i\alpha_i - C_i)v_i + B_iv_{i+1} + A_i\beta_i - \mathbf{F}_i = 0 \quad \text{for} \quad i = 1, \ldots, N-1.$$

We also have

$$\mathbf{v}_i = \alpha_{i+1}\mathbf{v}_{i+1} + \beta_{i+1} \quad \text{for} \quad i = 0, 1, \ldots, N-1.$$

Using this to eliminate $\mathbf{v}_i$, we obtain

$$[(A_i\alpha_i - C_i)\alpha_{i+1} + B_i]\mathbf{v}_{i+1} + [(A_i\alpha_i - C_i)\beta_{i+1} + A_i\beta_i - \mathbf{F}_i] = 0 \quad \text{for} \quad i = 1, \ldots, N-1.$$

This equation is satisfied if the two expressions in the square brackets are both zero:

$$(A_i\alpha_i - C_i)\alpha_{i+1} + B_i = O \quad \text{and} \quad (A_i\alpha_i - C_i)\beta_{i+1} + A_i\beta_i - \mathbf{F}_i = \mathbf{0}$$

for $i = 1, \ldots, N-1$.

Therefore,

$$\alpha_{i+1} = -(A_i\alpha_i - C_i)^{-1}B_i, \quad \beta_{i+1} = (A_i\alpha_i - C_i)^{-1}(\mathbf{F}_i - A_i\beta_i) \tag{3.27}$$

for $i = 1, \ldots, N-1$.

If α_1 and β_1 are known, then α_i and β_i for $i = 2, 3, \ldots, N$ can be computed using (3.27). Formulae (3.27) are well defined provided $A_i\alpha_i - C_i$ is a nonsingular matrix for all i.

α_1 and β_1 can be determined as follows:

$$\mathbf{v}_0 = \alpha_1\mathbf{v}_1 + \beta_1 \quad \text{and} \quad \mathbf{v}_0 = \mathbf{V}_0 \quad \Rightarrow \quad \alpha_1\mathbf{v}_1 + \beta_1 = \mathbf{V}_0.$$

To satisfy this, we choose

$$\alpha_1 = O \quad \text{and} \quad \beta_1 = \mathbf{V}_0.$$

Once we know all α_i and β_i, we compute $\mathbf{v}_{N-1}, \mathbf{v}_{N-2}, \ldots, \mathbf{v}_1$ using formula (3.26). This method works in practice but requires $N-1$ inversion of $(N-1)\times(N-1)$ matrices which can be quite expensive computationally.

Another type of boundary conditions. Consider the Laplace equation

$$u_{xx} + u_{yy} = 0 \tag{3.28}$$

in the unit square $(0 < x < 1,\ 0 < y < 1)$ with boundary conditions for normal derivative (Neumann problem):

$$u_x(0, y) = g_0(y), \quad u_x(1, y) = g_1(y), \quad u_y(x, 0) = h_0(x), \quad u_y(x, 1) = h_1(x). \tag{3.29}$$

At interior grid points (x_k, y_j), $k, j = 1, 2, \ldots, N-1$, we use the standard difference equations

$$4w_{k,j} - (w_{k+1,j} + w_{k-1,j} + w_{k,j+1} + w_{k,j-1}) = 0, \tag{3.30}$$

To approximate boundary conditions (3.29), we introduce false boundaries at $x = x_{-1} = -h$, $x = x_{N+1} = x_N + h$, $y = y_{-1} = -h$ and $y = y_{N+1} = y_N + h$ and the corresponding false grid points (x_{-1}, y_j), (x_{N+1}, y_j) for $j = 0, 1, \ldots, N$ and (x_k, y_{-1}), (x_k, y_{N+1}) for $k = 0, 1, \ldots, N$. Then we use the central difference formula for the 1st derivatives with respect to x and y to approximate boundary condition (3.29):

$$\frac{w_{1,j} - w_{-1,j}}{2h} = g_0(y_j), \qquad \frac{w_{N+1,j} - w_{N-1,j}}{2h} = g_1(y_j),$$
$$\frac{w_{k,1} - w_{k,-1}}{2h} = h_0(x_k), \qquad \frac{w_{k,N+1} - w_{k,N-1}}{2h} = h_1(x_k)$$

or

$$w_{-1,j} = w_{1,j} - 2hg_0(y_j), \qquad w_{N+1,j} = w_{N-1,j} + 2hg_1(y_j),$$
$$w_{k,-1} = w_{k,1} - 2hh_0(x_k), \qquad w_{k,N+1} = w_{k,N-1} + 2hh_1(x_k). \tag{3.31}$$

Assuming that Eq. (3.28) is satisfies at the boundary, we extend Eq. (3.30) to the boundary grid points and use Eqs. (3.31) to eliminate $w_{-1,j}$, $w_{N+1,j}$, $w_{k,-1}$ and $w_{N+1,j}$ wherever it is necessary. This yields

$$4w_{0,j} - (2w_{1,j} + w_{0,j+1} + w_{0,j-1}) = -2hg_0(y_j), \tag{3.32}$$
$$4w_{N,j} - (2w_{N-1,j} + w_{N,j+1} + w_{N,j-1}) = 2hg_1(y_j), \tag{3.33}$$
$$4w_{k,0} - (w_{k+1,0} + w_{k-1,0} + 2w_{k,1}) = -2hh_0(x_k), \tag{3.34}$$
$$4w_{k,N} - (w_{k+1,N} + w_{k-1,N} + 2w_{k,N-1}) = 2hh_1(x_k). \tag{3.35}$$

Equations (3.30) and (3.32)–(3.35) represent the system of $(N+1)^2$ linear equations for $(N+1)^2$ unknowns w_{kj} $(k, j = 0, 1, \ldots, N)$. In vector form, we have

$$A\mathbf{w} = \mathbf{f}.$$

However, it can be shown that matrix A here is singular. This is a consequence of the fact that the boundary value problem (3.28), (3.29) is solvable only under certain condition (a solvability condition). Indeed, integrating Eq. (3.28) over the unit square and taking account of boundary conditions (3.29), we obtain

$$\int_0^1 \int_0^1 (u_{xx} + u_{yy})\, dxdy = 0 \quad \Rightarrow \quad \int_0^1 (g_1(y) - g_0(y))dy + \int_0^1 (h_1(x) - h_0(x))dx = 0. \tag{3.36}$$

Thus, problem (3.28), (3.29) is solvable not for arbitrary functions $g_{0,1}(y)$, $h_{0,1}(x)$ but only for functions that satisfy the solvability condition (3.36). Fortunately, many elliptic boundary value problems which arise in practice do satisfy the solvability condition[6]. So, if we know that a solution exists and want to find a numerical approximation to it, we need to solve a system of linear algebraic equation with singular matrix. In this case, the direct methods do not work. We may try to use iterative techniques described above. Sometimes this may work. But it is better to use some different numerical technique (such as Galerkin method or finite-element method).

3.2 Equations with variable coefficients

Consider now the following elliptic equation

$$A(x,y)\frac{\partial^2 u}{\partial x^2} + B(x,y)\frac{\partial^2 u}{\partial y^2} + C(x,y)\frac{\partial u}{\partial x} + D(x,y)\frac{\partial u}{\partial y} + F(x,y)\, u = G(x,y) \tag{3.37}$$

for $(x, y) \in \mathcal{D}$ where $\mathcal{D} = \{\, (x,y) \,|\, a < x < b,\ c < y < d \,\}$ and A, B, C, D, F and G are given functions of x and y in $\mathcal{D}$. Equation (3.37) is supplemented with the boundary conditions (at the boundary S of the region $\mathcal{D}$):

$$u(x,y) = g(x,y) \quad \text{for} \quad (x,y) \in S. \tag{3.38}$$

To find a finite-difference approximation to the boundary value problem (3.37), (3.38) at interior grid points (x_k, y_j) ($x_k = a + h_1 k$ for $k = 1, \ldots, N_1$, $y_j = a + h_2 j$ for $j = 1, \ldots, N_2$, $h_1 = (b-c)/N_1$, $h_2 = (b-c)/N_2$), we use the three-point central-difference formulas to approximate the first and second partial derivatives with respect to x and y As a result, we obtain the following difference equations:

$$\left(\frac{A_{kj}}{h_1^2}\delta_x^2 + \frac{B_{kj}}{h_2^2}\delta_y^2 + \frac{C_{kj}}{2h_1}\delta_x + \frac{D_{kj}}{2h_2}\delta_y + F_{kj} \right) w_{kj} = G_{kj} \tag{3.39}$$

where

$$\delta_x^2 = w_{k+1,j} - 2w_{k,j} + w_{k-1,j}, \qquad \delta_y^2 = w_{k,j+1} - 2w_{k,j} + w_{k,j-1},$$
$$\delta_x = w_{k+1,j} - w_{k-1,j}, \qquad \delta_y = w_{k,j+1} - w_{k,j-1},$$

and where $A_{kj} = A(x_k, y_j)$, $B_{kj} = B(x_k, y_j)$, etc. Since the truncation error of each formula employed to obtain Eq. (3.39) is either $O(h_1^2)$ or $O(h_2^2)$, the truncation error of this finite-difference method is $O(h_1^2 + h_2^2))$.

[6] Even in this case, the solution is not unique: there are infinitely many solutions that differ one from another by a constant. Evidently, if $u(x, y)$ is a solution of (3.28), (3.29), then $u(x, y) + c$ with any constant c is also a solution.

If we use the information from the boundary conditions (3.38) whenever appropriate in the system (3.39), we obtain an $(N_1-1)(N_2-1)$ by $(N_1-1)(N_2-1)$ linear system with the unknowns w_{ij} at the interior mesh points:

$$A\mathbf{w} = \mathbf{b}$$

The matrix A of this linear system is, in general, not symmetric even for the square domain ($c=a$, $d=b$) with the same number of grid lines in x and y because the coefficients of Eq. (3.37) depend on x and y.

If the linear elliptic equation (3.37) is self-adjoint or can be made self-adjoint, i.e. can be written in the form

$$\frac{\partial}{\partial x}\left(A\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(B\frac{\partial u}{\partial y}\right) + F\,u = G,$$

for some $A(x,y)$ and $B(x,y)$, and we have the square domain ($c=a$, $d=b$) with the same number of grid lines in x and y, then the symmetry of A can be assured if we employ approximations of the form

$$\left.\frac{\partial}{\partial x}\left(A\frac{\partial u}{\partial x}\right)\right|_{(x_k,y_j)} = \frac{A_{k+\frac{1}{2},j}}{h^2}\left(u_{k+1,j}-u_{k,j}\right) - \frac{A_{k-\frac{1}{2},j}}{h^2}\left(u_{k,j}-u_{k-1,j}\right) + O(h^2),$$

$$\left.\frac{\partial}{\partial y}\left(B\frac{\partial u}{\partial y}\right)\right|_{(x_k,y_j)} = \frac{B_{k,j+\frac{1}{2}}}{h^2}\left(u_{k,j+1}-u_{k,j}\right) - \frac{B_{k,j-\frac{1}{2}}}{h^2}\left(u_{k,j}-u_{k,j-1}\right) + O(h^2),$$

where $A_{k\pm\frac{1}{2},j} = (A_{k\pm1,j}+A_{kj})/2$ and $B_{k,j\pm\frac{1}{2}} = (B_{k,j\pm1}+B_{kj})/2$.

3.3 Arbitrary (not rectangular) domains

How to deal with the problem when region $\mathcal{D}$ is not a rectangle? In this case, we first cover the whole region by a rectangular grid, say, with the step length h_1 in the x direction and the step length h_2 in the y direction. Then all interior grid points can be divided in two groups: regular and irregular grid points. A grid point is called regular if all its neighbouring grid points are also inside region $\mathcal{D}$. At a regular point, we can use the difference equation (3.39). Irregular points require a special treatment.

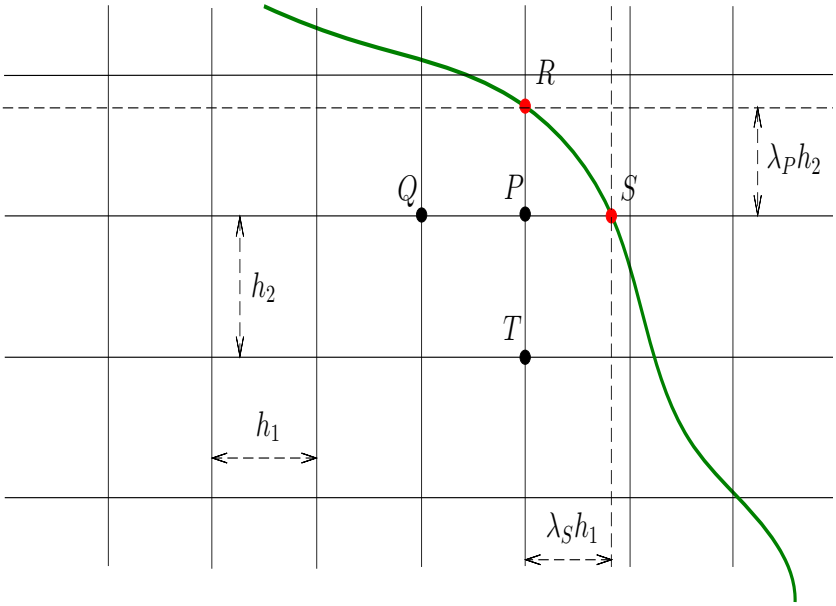

Figure 3.1:

Suppose that $P = (x, y)$ is an irregular grid point, such that $S = (x+\lambda_S h_1, y)$ and $R = (x, y+\lambda_R h_2)$ where $0 < \lambda_{S,R} < 1$ lie on the boundary of $\mathcal{D}$ and the other two neighbouring grid points, $Q = (x-h_1, y)$ and $T = (x, y-h_2)$, are inside $\mathcal{D}$ (see Fig. 1). Then the derivatives of u at point P can be approximated in the following manner. First, we expand $u_S = u|_S$ and $u_Q = u|_Q$ in Taylor series at point P:

$$u_S = u_P + \lambda_S h_1 \frac{\partial u}{\partial x}\bigg|_P + \frac{(\lambda_S h_1)^2}{2}\frac{\partial^2 u}{\partial x^2}\bigg|_P + O(h_1^3),$$
$$u_Q = u_P - h_1 \frac{\partial u}{\partial x}\bigg|_P + \frac{h_1^2}{2}\frac{\partial^2 u}{\partial x^2}\bigg|_P + O(h_1^3),$$

Multiplying the second equation by λ_S^2 and subtracting the result from the first equation yields

$$\frac{\partial u}{\partial x}\bigg|_P = \frac{1}{h_1\lambda_S(1+\lambda_S)}\left[u_S - \lambda_S^2 u_Q - (1-\lambda_S^2)u_P\right] + O(h_1^2).$$

Similarly, if we multiply the second equation by λ_S and add the result to the first equation yields, we can obtain the formula for the second derivative with respect to x:

$$\frac{\partial^2 u}{\partial x^2}\bigg|_P = \frac{2}{h_1^2\lambda_S(1+\lambda_S)}\left[u_S + \lambda_S u_Q - (1+\lambda_S)u_P\right] + O(h_1).$$

Note that the truncation error of the formula for the second derivative is $O(h_1)$. To obtain a higher order accuracy, we need to use one more grid point, e.g. point $(x - 2h_1, y)$.

Similarly, one can obtain the following formulae for the derivatives with respect to y:

$$\frac{\partial u}{\partial y}\bigg|_P = \frac{1}{h_2\lambda_R(1+\lambda_R)}\left[u_R - \lambda_R^2 u_T - (1-\lambda_R^2)u_P\right] + O(h_2^2),$$
$$\frac{\partial^2 u}{\partial y^2}\bigg|_P = \frac{2}{h_2^2\lambda_R(1+\lambda_R)}\left[u_R + \lambda_R u_T - (1+\lambda_R)u_P\right] + O(h_2).$$

4 Hyperbolic partial differential equations

4.1 Wave equation

We will illustrate the finite-difference method for hyperbolic PDEs with the linear wave equation

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \alpha^2 \frac{\partial^2 u}{\partial x^2}(x,t) = F(x,t), \quad a < x < b, \quad 0 < t < T, \tag{4.1}$$

subject to the boundary conditions

$$u(a,t) = u(b,t) = 0 \quad \text{for} \quad t \in [0,T], \tag{4.2}$$

and initial conditions

$$u(x,0) = f(x), \tag{4.3}$$

$$\frac{\partial u}{\partial t}(x,0) = g(x) \quad \text{for} \quad x \in [a,b], \tag{4.4}$$

where $f(x)$ and $g(x)$ are given functions.

First we choose integers N and M and let $h = (b-a)/N$, $\tau = T/M$. Then we define the mesh points (x_k, t_j):

$$x_k = a + hk \quad (k = 0, 1, \ldots, N), \quad t_j = \tau j \quad (j = 0, 1, \ldots, M).$$

We approximate the second partial derivatives with respect to t and x at interior grid points using the central difference formulas

$$\frac{\partial^2 u}{\partial t^2}(x_k,t_j) = \frac{u(x_k,t_{j+1}) - 2u(x_k,t_j) + u(x_k,t_{j-1})}{\tau^2} - \frac{\tau^2}{12}\frac{\partial^4 u}{\partial t^4}(x_k,\mu_j),$$

$$\frac{\partial^2 u}{\partial x^2}(x_k,t_j) = \frac{u(x_{k+1},t_j) - 2u(x_k,t_j) - u(x_{k-1},t_j)}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_k,t_j),$$

where $\mu_j \in (t_{j-1}, t_{j+1})$ and $\xi_k \in (x_{k-1}, x_{k+1})$.

With these formulas, we approximate the wave equation (4.1) at the interior mesh points (x_k, t_j) for $k = 1, 2, \ldots, N-1$ and $j = 1, 2, \ldots$ by the difference equation

$$\frac{w_{k,j+1} - 2w_{kj} + w_{k,j-1}}{\tau^2} - \alpha^2 \frac{w_{k+1,j} - 2w_{kj} + w_{k-1,j}}{h^2} = F_{kj}, \tag{4.5}$$

where w_{kj} approximates $u(x_k, t_j)$ (i.e. $w_{kj} \approx u(x_k, t_j)$).

The local truncation error of the difference equation (4.5) is

$$\tau_{ij} = \frac{\tau^2}{12}\frac{\partial^4 u}{\partial t^4}(x_k,\mu_j) - \alpha^2 \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_k,t_j) = O(\tau^2 + h^2).$$

Let

$$\gamma \equiv \frac{\alpha\tau}{h}.$$

Then equation (4.5) can be written as

$$w_{k,j+1} = 2\left(1-\gamma^2\right) w_{kj} + \gamma^2 \left(w_{k+1,j} + w_{k-1,j}\right) - w_{k,j-1} + \tau^2 F_{kj}, \tag{4.6}$$

for each $k = 1, 2, \ldots, N-1$ and $j = 1, 2, \ldots$. The boundary conditions (4.2) imply that

$$w_{0,j} = w_{N,j} = 0 \quad \text{for each} \quad j = 1, 2, \ldots, \tag{4.7}$$

and the initial condition (4.3) yields

$$w_{k,0} = f(x_k) \quad \text{for each} \quad i = 1, 2, \ldots, N-1. \tag{4.8}$$

Equation (4.5) can be written in the matrix form

$$\mathbf{w}^{(j+1)} = A\mathbf{w}^{(j)} - \mathbf{w}^{(j-1)} + \mathbf{F}^{(j)}, \tag{4.9}$$

where

$$A = \begin{bmatrix} 2(1-\gamma^2) & \gamma^2 & 0 & \dots & \dots & 0 \\ \gamma^2 & 2(1-\gamma^2) & \gamma^2 & \ddots & & \vdots \\ 0 & \gamma^2 & 2(1-\gamma^2) & \gamma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \gamma^2 \\ 0 & \dots & \dots & 0 & \gamma^2 & 2(1-\gamma^2) \end{bmatrix}, \tag{4.10}$$

$$\mathbf{w}^{(j)} = \begin{bmatrix} w_{1,j} \\ w_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ w_{N-1,j} \end{bmatrix}, \quad \mathbf{F}^{(j)} = \begin{bmatrix} \tau^2 F_{1,j} \\ \tau^2 F_{2,j} \\ \vdots \\ \vdots \\ \vdots \\ \tau^2 F_{N-1,j} \end{bmatrix}. \tag{4.11}$$

It is clear from Eq. (4.9) that the $(j+1)$-st time step requires the values from the j-th and $(j-1)$-st time steps. For example, to compute $w_{k,2}$, we need values of $w_{k,0}$ and $w_{k,1}$ for $k = 1, 2, \dots, N-1$. The values of $w_{k,0}$ are given by Eq. (4.8), but values of $w_{k,1}$ must be obtained from the other initial condition (4.4).

The simplest way to approximate the initial condition (4.4) is to replace $\partial u/\partial t$ by the forward-difference formula

$$\frac{\partial u}{\partial t}(x_k, 0) = \frac{u(x_k, t_1) - u(x_k, 0)}{\tau} - \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}(x_k, \mu), \quad 0 < \mu < \tau.$$

It follows that

$$u(x_k, t_1) = u(x_k, 0) + \tau\frac{\partial u}{\partial t}(x_k, 0) + \frac{\tau^2}{2}\frac{\partial^2 u}{\partial t^2}(x_k, \mu) = f(x_k) + \tau g(x_k) + \frac{\tau^2}{2}\frac{\partial^2 u}{\partial t^2}(x_k, \mu).$$

Hence,

$$w_{k,1} = f(x_k) + \tau g(x_k). \tag{4.12}$$

However, this equation has local truncation error of order $O(\tau)$. A better approximation to $u(x_k, t_1)$ can be obtained as follows. Expanding $u(x_k, t_1)$ in Taylor's series in t at $(x_k, 0)$, we obtain

$$\frac{u(x_k, t_1) - u(x_k, 0)}{\tau} = \frac{\partial u}{\partial t}(x_k, 0) + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}(x_k, 0) + \frac{\tau^2}{6}\frac{\partial^3 u}{\partial t^3}(x_k, \mu) \tag{4.13}$$

for some $\mu \in (0, t_1)$. Suppose that the wave equation also holds on the initial line, i.e.

$$\frac{\partial^2 u}{\partial t^2}(x_k, 0) - \alpha^2\frac{\partial^2 u}{\partial x^2}(x_k, 0) = 0 \quad \text{for} \quad k = 0, 1, \dots, N.$$

Then

$$\frac{\partial^2 u}{\partial t^2}(x_k, 0) = \alpha^2\frac{\partial^2 u}{\partial x^2}(x_k, 0) + F((x_k, 0) = \alpha^2 f''(x_k) + F(x_k, 0).$$

This equation and Eq. (4.13) yield

$$u(x_k, t_1) = f(x_k) + \tau g(x_k) + \frac{\tau^2}{2}\left[\alpha^2 f''(x_k) + F(x_k, 0)\right] + \frac{\tau^3}{6}\frac{\partial^3 u}{\partial t^3}(x_k, \mu).$$

Hence,

$$w_{k,1} = f(x_k) + \tau g(x_k) + \frac{\tau^2}{2}\left[\alpha^2 f''(x_k) + F(x_k, 0)\right]. \tag{4.14}$$

This is an approximation with local truncation error $O(\tau^2)$ for each $k = 1, 2, \ldots, N-1$. If $f''(x_k)$ is not readily available, we can use the central difference formula to approximate it.

The finite-difference method described above is explicit and has local truncation error $O(\tau^2 + h^2)$. Now we will investigate the stability of the method. To do this, we employ the Fourier method. Since the difference equation (4.5) is linear, the perturbation z_{kj} satisfies the equation

$$\frac{z_{k,j+1} - 2z_{kj} + z_{k,j-1}}{\tau^2} - \alpha^2 \frac{z_{k+1,j} - 2z_{kj} - z_{k-1,j}}{h^2} = 0,$$

for $k = 1, 2, \ldots, N-1, j = 1, 2, \ldots$, which is a homogeneous version of Eq. (4.5). Substituting $z_{k,j} = \rho_q^j e^{iqx_k}$ into this equation, we obtain

$$e^{iqx_k}\left(\rho_q^{j+1} - 2\rho_q^j + \rho_q^{j-1}\right) - \gamma^2 \rho_q^j \left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right) = 0$$

or

$$\rho_q^2 - 2\rho_q\left[1 + \frac{\gamma^2}{2}\left(e^{iqh} - 2 + e^{-iqh}\right)\right] + 1 = 0$$

$$\Rightarrow \quad \rho_q^2 - 2\rho_q\left[1 - 2\gamma^2 \sin^2\frac{qh}{2}\right] + 1 = 0.$$

Thus, ρ_q is a root of the quadratic equation

$$\rho_q^2 - 2a\rho_q + 1 = 0, \quad a \equiv 1 - 2\gamma^2 \sin^2\frac{qh}{2}.$$

Its roots are $\rho_q^{\pm} = a \pm \sqrt{a^2 - 1}$, so that the product of the roots is equal to 1 ($\rho_q^+ \rho_q^- = 1$). It follows that the stability condition $|\rho_q| \le 1$ can be satisfied only if $|\rho_q^+| = |\rho_q^-| = 1$. This means that the roots must be complex conjugate. Hence, we must have

$$a^2 - 1 \le 0 \quad \text{or} \quad \left|1 - 2\gamma^2 \sin^2\frac{qh}{2}\right| \le 1 \quad \Rightarrow \quad -1 \le 1 - 2\gamma^2 \sin^2\frac{qh}{2}.$$

The last inequality is satisfied for all q if $\gamma \le 1$. It can be shown that if $\gamma = 1$, then for certain modes q (e.g. for q such that $qh = \pi$), the above quadratic equation has a double root ρ_q, which results in a week instability: in the limit $\tau \to 0$, their amplitudes grow linearly in j. Therefore, the stability condition is

$$\gamma < 1 \quad \text{or} \quad \alpha\tau < h.$$

Thus, the above explicit finite-difference method is conditionally stable.

To obtain an *unconditionally* stable method, we consider the difference equation in which the second derivative with respect to x is approximated by the central difference formula averaged over the three time steps: $j+1$, j and $j-1$.

Let

$$\delta_x^2 w_{kj} = w_{k+1,j} - 2w_{k,j} + w_{k-1,j}.$$

Consider the difference equation

$$w_{k,j+1} - 2w_{kj} + w_{k,j-1} - \gamma^2\left[\sigma\delta_x^2 w_{k,j+1} + (1-2\sigma)\delta_x^2 w_{k,j} + \sigma\delta_x^2 w_{k,j-1}\right] = \tau^2 F_{kj}, \tag{4.15}$$

where σ is an arbitrary number such that $0 \le \sigma \le 1/2$. It can be shown that local truncation error of Eq. (4.15) is $O(\tau^2 + h^2)$. Now we will try to choose σ such that the implicit method given by Eq. (4.15) is unconditionally stable.

We will investigate the stability of the method by the Fourier method. The perturbation z_{kj} satisfies the equation

$$z_{k,j+1} - 2z_{kj} + z_{k,j-1} - \gamma^2 \left[\sigma\delta_x^2 z_{k,j+1} + (1-2\sigma)\delta_x^2 z_{k,j} + \sigma\delta_x^2 z_{k,j-1}\right] = 0,$$

for $k = 1, 2, \ldots, N-1, j = 1, 2, \ldots$. This is the homogeneous version of Eq. (4.15). Substituting $z_{k,j} = \rho_q^j e^{iqx_k}$ into this equation, we obtain

$$e^{iqx_k}\left(\rho_q^{j+1} - 2\rho_q^j + \rho_q^{j-1}\right) - \gamma^2\left[\sigma\rho_q^{j+1} + (1-2\sigma)\rho_q^j + \sigma\rho_q^{j-1}\right]\left(e^{iqx_{k+1}} - 2e^{iqx_k} + e^{iqx_{k-1}}\right) = 0$$

or

$$\left(\rho_q^2 - 2\rho_q + 1\right) + 4\gamma^2 \sin^2\frac{qh}{2}\left[\sigma\rho_q^2 + (1-2\sigma)\rho_q + \sigma\right] = 0.$$

Hence, ρ_q satisfies the quadratic equation

$$\rho_q^2 - 2a\rho_q + 1 = 0, \tag{4.16}$$

where

$$a \equiv \frac{1 - 2\gamma^2(1-2\sigma)\sin^2\frac{qh}{2}}{1 + 4\gamma^2\sigma\sin^2\frac{qh}{2}}.$$

Again, the method can be stable only if Eq. (4.16) has complex conjugate roots. This implies that

$$a^2 - 1 < 0 \quad \Rightarrow \quad \left|\frac{1 - 2\gamma^2(1-2\sigma)\sin^2\frac{qh}{2}}{1 + 4\gamma^2\sigma\sin^2\frac{qh}{2}}\right| \leq 1$$

$$\Rightarrow \quad -1 \leq \frac{1 - 2\gamma^2(1-2\sigma)\sin^2\frac{qh}{2}}{1 + 4\gamma^2\sigma\sin^2\frac{qh}{2}}.$$

The last inequality is equivalent to

$$2 - 2\gamma^2(1-2\sigma)\sin^2\frac{qh}{2} + 4\gamma^2\sigma\sin^2\frac{qh}{2} \geq 0$$

or

$$1 - \gamma^2(1-4\sigma)\sin^2\frac{qh}{2} \geq 0.$$

Evidently, if $\sigma \geq 1/4$, this inequality is satisfied for all q, irrespective of the value of γ. Thus, if $\sigma \in [1/4, 1/2]$, then the above implicit method is unconditionally stable.

4.2 Hyperbolic systems of first-order partial differential equations

Since the equations of physics (e.g., fluid mechanics) are based upon conservation laws, it convenient to use a form of the equations, called the *divergence form* (or *flux-conservative* form, or *conservation-law* form). A system of equations

$$\mathbf{U}_t + [\mathbf{F}(\mathbf{U})]_x = 0, \tag{4.17}$$

where $\mathbf{U}(x,t)$ is a vector function with n components and $\mathbf{F}$ is a (in general, nonlinear) vector function (with n components) of the vector $\mathbf{U}$, is called a system of conservation laws.

The wave equation

$$\frac{\partial^2 u}{\partial t^2}(x,t) - \alpha^2\frac{\partial^2 u}{\partial x^2}(x,t) = 0 \tag{4.18}$$

can be easily written in the conservative form (4.17). Indeed, if

$$r = \alpha u_x \quad \text{and} \quad s = u_t,$$

then Eq. (4.18) is equivalent to the system:

$$\frac{\partial}{\partial t}\begin{pmatrix} r \\ s \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} -\alpha s \\ -\alpha r \end{pmatrix} = 0, \tag{4.19}$$

which has the form of Eq. (4.17).

In what follows we will discuss only the pure initial value problem (putting aside problems with boundary conditions).

Methods with error $O(\tau + h^2)$. First we consider the simplest case of Eq. (4.17) when **U** has only one component. Namely, we start with the scalar equation

$$u_t + \alpha u_x = 0, \tag{4.20}$$

where α is a constant. Use the forward-difference formula for u_t and the central difference formula for u_x, we obtain the following finite-difference approximation of Eq. (4.20):

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} + \alpha \frac{w_{k+1,j} - w_{k-1,j}}{2h} = 0, \tag{4.21}$$

where $w_{k,j}$ denotes the discrete approximation to $u(x_k, t_j)$, τ and h are the step sizes in t and x, respectively. Equation (4.21) represent an explicit method whose local truncation error is $O(\tau + h^2)$.

To find out whether this method is stable, we use the Fourier method. The error (perturbation) $z_{k,j}$ satisfies the same equation as Eq. (4.21). And we look for a solution in the form $z_{k,j} = \rho_q^j e^{iqx_k}$ where $q \in \mathbb{R}$. Substitution of $z_{k,j}$ in Eq. (4.21) yields

$$\rho_q - 1 + \frac{\gamma}{2}\left(e^{iqh} - e^{-iqh}\right) = 0$$

or

$$\rho_q = 1 - i\gamma \sin(qh),$$

where $\gamma = \alpha\tau/h$. Evidently,

$$|\rho_q|^2 = 1 + \gamma^2 \sin^2(qh) > 1.$$

for all q for which $\sin(qh) \neq 0$. Therefore, the method is unconditionally unstable.

It turns out that a stable method can be obtained simply by replacing $u_{k,j}$ in the forward-difference formula for the time derivative

$$u_t \approx \frac{u_{k,j+1} - u_{k,j}}{\tau}$$

by its average

$$\frac{1}{2}\left(u_{k+1,j} + u_{k-1,j}\right).$$

This transforms Eq. (4.21) to

$$\frac{1}{\tau}\left(w_{k,j+1} - \frac{1}{2}[w_{k+1,j} + w_{k-1,j}]\right) + \alpha \frac{w_{k+1,j} - w_{k-1,j}}{2h} = 0,$$

or, equivalently,

$$w_{k,j+1} = \frac{1}{2}[w_{k+1,j} + w_{k-1,j}] - \frac{\gamma}{2}\left[w_{k+1,j} - w_{k-1,j}\right]. \tag{4.22}$$

Equation (4.22) is called the *Lax* scheme. Let us now investigate its stability. Substitution of $z_{k,j} = \rho_q^j e^{iqx_k}$ leads to

$$\rho_q = \frac{1}{2}\left(e^{iqh} + e^{-iqh}\right) - \frac{\gamma}{2}\left(e^{iqh} - e^{-iqh}\right) \quad \Rightarrow \quad \rho_q = \cos(qh) - i\gamma\sin(qh).$$

If follows that

$$|\rho_q|^2 = \cos^2(qh) + \gamma^2\sin^2(qh) \quad \Leftrightarrow \quad |\rho_q|^2 = 1 + (\gamma^2 - 1)\sin^2(qh).$$

The stability condition $|\rho_q| \le 1$ leads to the requirement

$$\gamma \le 1 \quad \text{or} \quad \tau \le \frac{h}{\alpha}. \tag{4.23}$$

This inequality is called the *Courant* stability criterion. The surprising result that the above simple modification can stabilize an unconditionally stable method can be explained as follows. First, we can rewrite Eq. (4.22) as

$$\frac{w_{k,j+1} - w_{k,j}}{\tau} + \alpha\frac{w_{k+1,j} - w_{k-1,j}}{2h} = \frac{h^2}{2\tau}\frac{w_{k+1,j} - 2w_{k,j} + w_{k-1,j}}{h^2}. \tag{4.24}$$

If the term on the right side of this equation was zero, we would have the unconditionally unstable method (4.21). For small τ and h the difference equation (4.24) approximates the equation

$$u_t + \alpha u_x = \frac{h^2}{2\tau}u_{xx}. \tag{4.25}$$

Thus, effectively we have added a diffusion (or dissipation) term to the equation, and this made the new method stable. The Lax scheme is said to have *numerical dissipation* or *numerical viscosity*.

When we have a system of equations rather than a scalar equation, the stability analysis becomes slightly more complicated. As an example, consider the wave equation written in the form (4.19). The Lax scheme for Eq. (4.19) is

$$\begin{aligned} r_{k,j+1} &= \frac{1}{2}[r_{k+1,j} + r_{k-1,j}] + \frac{\gamma}{2}\left[s_{k+1,j} - s_{k-1,j}\right], \\ s_{k,j+1} &= \frac{1}{2}[s_{k+1,j} + s_{k-1,j}] + \frac{\gamma}{2}\left[r_{k+1,j} - r_{k-1,j}\right], \end{aligned} \tag{4.26}$$

where r_{kj} and s_{kj} denote approximations to $r(x_k, t_j)$ and $s(x_k, t_j)$, respectively. To investigate the stability of (4.26), we assume that

$$\begin{pmatrix} r_{kj} \\ s_{kj} \end{pmatrix} = \rho_q^j e^{iqx_k} \begin{pmatrix} r^{(0)} \\ s^{(0)} \end{pmatrix}, \tag{4.27}$$

where $r^{(0)}$ and $s^{(0)}$ are constants. Substituting this in Eq. (4.26), we obtain

$$\begin{pmatrix} \cos(qh) - \rho_q & i\gamma\sin(qh) \\ i\gamma\sin(qh) & \cos(qh) - \rho_q \end{pmatrix} \begin{pmatrix} r^{(0)} \\ s^{(0)} \end{pmatrix} = 0.$$

This system has a nonzero solution only if the determinant of the matrix on the left side is zero. This gives us

$$(\cos(qh) - \rho_q)^2 + \gamma^2\sin^2(qh) = 0 \quad \Rightarrow \quad \rho_q = \cos(qh) \pm i\gamma\sin(qh).$$

The stability condition is that both roots satisfy the inequality $|\rho_q| \le 1$, which again leads us to the Courant condition (4.23).

For the system of conservation laws (4.17), the Lax method is given by

$$\mathbf{U}_{k,j+1} = \frac{1}{2}\left(\mathbf{U}_{k+1,j} + \mathbf{U}_{k-1,j}\right) - \frac{\tau}{2h}\left[\mathbf{F}(\mathbf{U}_{k+1,j}) - \mathbf{F}(\mathbf{U}_{k-1,j})\right]. \tag{4.28}$$

Here $\mathbf{U}_{k,j} \approx \mathbf{U}(x_k, t_j)$.

Methods with error $O(\tau^2 + h^2)$. All the schemes discussed above have local truncation error $O(\tau + h^2)$. It is desirable to have a method whose truncation error is quadratic both in space and time. We will discuss two such methods. First of them is called the 'leapfrog' method in which we use the central-difference formula for derivatives in both x and t. For Eq. (4.20), this method is given by

$$\frac{w_{k,j+1} - w_{k,j-1}}{2\tau} + \alpha\frac{w_{k+1,j} - w_{k-1,j}}{2h} = 0. \tag{4.29}$$

The standard stability analysis yields the following equation for ρ_q:

$$\rho_q^2 + 2i\gamma\sin(qh)\rho_q - 1 = 0.$$

It follows that

$$\rho_q = -i\gamma\sin(qh) \pm \sqrt{1 - \gamma^2\sin^2(qh)}.$$

If the Courant condition (4.23) is satisfied, i.e. $\gamma \leq 1$, then $|\rho_q| = 1$, and the method is stable. If $\gamma > 1$, then for q such that $\sin(qh) = 1$, we have

$$\rho_q = -i\gamma \pm i\sqrt{\gamma^2 - 1}.$$

For the second root (with 'minus' sign), we obtain

$$|\rho_q| = \gamma + \sqrt{\gamma^2 - 1} > 1.$$

Therefore, the Courant condition is the necessary and sufficient condition for stability of the 'leapfrog' method.

For the system of conservation laws (4.17), the 'leapfrog' method is given by

$$\frac{\mathbf{U}_{k,j+1} - \mathbf{U}_{k,j-1}}{2\tau} + \frac{\mathbf{F}(\mathbf{U}_{k+1,j}) - \mathbf{F}(\mathbf{U}_{k-1,j})}{2h} = 0. \tag{4.30}$$

Here $\mathbf{U}_{k,j} \approx \mathbf{U}(x_k, t_j)$. Note that in order to use Eq. (4.29) or Eq. (4.30), one needs to know $w_{k,0}$ and $w_{k,1}$ (or $\mathbf{U}_{k,0}$ and $\mathbf{U}_{k,1}$).

The other method is called the two-step Lax-Wendroff scheme. First, we compute intermediate values $w_{k+\frac{1}{2},j+\frac{1}{2}}$ at the half timesteps $t_{j+\frac{1}{2}}$ and the half mesh point $x_{k+\frac{1}{2}}$ using the Lax method:

$$w_{k+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{2}\left[w_{k+1,j} + w_{k,j}\right] - \frac{\gamma}{2}\left[w_{k+1,j} - w_{k,j}\right]. \tag{4.31}$$

Then, we compute the updated values using the equation:

$$w_{k,j+1} = w_{k,j} - \gamma\left[w_{k+\frac{1}{2},j+\frac{1}{2}} - w_{k-\frac{1}{2},j+\frac{1}{2}}\right]. \tag{4.32}$$

Substituting (4.31) in (4.32), we can rewrite the method in the form:

$$\begin{aligned} w_{k,j+1} = w_{k,j} - \gamma\Bigg[&\frac{1}{2}\left(w_{k+1,j} + w_{k,j}\right) - \frac{\gamma}{2}\left(w_{k+1,j} - w_{k,j}\right) \\ &-\frac{1}{2}\left(w_{k,j} + w_{k-1,j}\right) + \frac{\gamma}{2}\left(w_{k,j} - w_{k-1,j}\right)\Bigg]. \end{aligned}$$

or, equivalently,

$$w_{k,j+1} = w_{k,j} - \gamma \left[\frac{1}{2} (w_{k+1,j} - w_{k-1,j}) - \frac{\gamma}{2} (w_{k+1,j} - 2w_{k,j} + w_{k-1,j}) \right]. \tag{4.33}$$

One can show that the local truncation error of Eq. (4.33) is $O(\tau^2 + h^2)$ (prove it!). The stability analysis leads to

$$\rho_q = 1 - i\gamma \sin(qh) - \gamma^2 [1 - \cos(qh)].$$

It follows that

$$|\rho_q|^2 = 1 - \gamma^2 (1 - \gamma^2)[1 - \cos(qh)]^2.$$

Again, the stability condition is satisfied provided that $\gamma \leq 1$, i.e. if the Courant condition (4.23) holds.

For the system of conservation laws (4.17), the Lax-Wendroff scheme has the form

$$\begin{aligned} \mathbf{U}_{k+\frac{1}{2},j+\frac{1}{2}} &= \frac{1}{2} [\mathbf{U}_{k+1,j} + \mathbf{U}_{k,j}] - \frac{\tau}{2h} [\mathbf{F}(\mathbf{U}_{k+1,j}) - \mathbf{F}(\mathbf{U}_{k,j})], \\ \mathbf{U}_{k,j+1} &= \mathbf{U}_{k,j} - \frac{\tau}{h} \left[\mathbf{F}(\mathbf{U}_{k+\frac{1}{2},j+\frac{1}{2}}) - \mathbf{F}(\mathbf{U}_{k-\frac{1}{2},j+\frac{1}{2}}) \right]. \end{aligned} \tag{4.34}$$

5 Appendix A

We want to show that if $u \in C^4(D)$ (where $D = \{(x,y) \mid 0 < x < 1,\ 0 < y < 1\}$) is the exact solution of the boundary-value problem (3.17), (3.18) and w_{ij} $(i,j = 1,2,\ldots,N-1)$ satisfy Eq. (3.19), then

$$|w_{ij} - u(x_i, y_j)| \leq Ah^2,$$

where A is independent of h.

Solution. To solve this problem, we need first to prove 2 auxiliary propositions.

Proposition 1. *Let v_{ij} for $i,j = 0,1\ldots,N$ be a set of real numbers satisfying the inequality*

$$v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{i,j} \geq 0, \tag{A.1}$$

for all $i,j = 1,2,\ldots,N-1$ (i.e at all internal mesh points). Then the maximum of v_{ij} is attained at least at one of the boundary points.

Proof. Assume that the above statement is not true, i.e. that the maximum is attained at an internal point (there may be several such points). Let (m,n) be a point at which the maximum is attained and which corresponds to the maximum value of m, i.e.

$$v_{m,n} = \max_{0 \leq i,j \leq N} v_{ij} \quad \text{and} \quad v_{m,n} > v_{m+1,n}.$$

Then,

$$v_{m+1,n} - v_{m,n} + (v_{m-1,n} - v_{m,n}) + (v_{m,n+1} - v_{m,n}) + (v_{m,n-1} - v_{m,n}) \leq v_{m+1,n} - v_{m,n} < 0. \tag{A.2}$$

Evidently, (A.2) is in contradiction with (A.1). Thus, our assumption is wrong, which proves the proposition.

Proposition 2. *Let v_{ij} for $i,j = 0,1\ldots,N$ be a set of real numbers satisfying the inequality*

$$v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{i,j} \leq 0, \tag{A.3}$$

for all $i,j = 1,2,\ldots,N-1$ (i.e at all internal mesh points). Then the minimum of v_{ij} is attained at least at one of the boundary points.

Proof. Assume that the above statement is not true, i.e. that the minimum is attained at an internal point (there may be several such points). Let (m, n) be a point at which the minimum is attained and which corresponds to the minimum value of m, i.e.

$$v_{m,n} = \min_{0 \le i,j \le N} v_{ij} \quad \text{and} \quad v_{m,n} < v_{m-1,n}.$$

Then we have

$$v_{m+1,n} - v_{m,n} + (v_{m-1,n} - v_{m,n}) + (v_{m,n+1} - v_{m,n}) + (v_{m,n-1} - v_{m,n}) \ge v_{m-1,n} - v_{m,n} > 0. \quad \text{(A.4)}$$

Evidently, (A.4) contradicts inequality (A.3). So, our assumption is wrong, which proves the proposition.

Now we are ready to prove the original statement. If $z_{ij} = w_{ij} - u_{ij}$, then, it follows from Eq. (3.5) that

$$z_{i+1,j} + z_{i-1,j} + z_{i,j+1} + z_{i,j-1} - 4z_{i,j} = h^2 f(x_j, y_j) - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} + 4u_{i,j},$$

where $u_{ij} \equiv u(x_i, y_j)$. Comparing the right hand side of this equation with the definition of local truncation error $\tau_{ij}(h)$, we find that

$$z_{i+1,j} + z_{i-1,j} + z_{i,j+1} + z_{i,j-1} - 4z_{i,j} = -h^2 \tau_{ij}(h). \quad \text{(A.5)}$$

Let $R = \operatorname{diam} \mathcal{D}/2 = \sqrt{2}/2$. We define the auxiliary function $Q(x, y)$ by the formula

$$Q(x, y) = \frac{1}{4}\left[R^2 - \left(x - \frac{1}{2}\right)^2 - \left(y - \frac{1}{2}\right)^2\right] E,$$

where

$$E \equiv \max_{1 \le i,j \le N-1} |\tau_{ij}(h)|$$

is the maximum truncation error. Q is a quadratic polynomial in x and y, and therefore

$$Q_{i+1,j} + Q_{i-1,j} + Q_{i,j+1} + Q_{i,j-1} - 4Q_{i,j} = h^2 \left(\frac{\partial^2 Q}{\partial x^2} + \frac{\partial^2 Q}{\partial x^2}\right)\Bigg|_{x=x_i, y=y_j} = -h^2 E, \quad \text{(A.6)}$$

where $Q_{ij} \equiv Q(x_i, y_j)$.

Let $v_{ij} = z_{ij} - Q_{ij}$. Then, according to (A.5) and (A.6)

$$v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{i,j} = h^2(E - \tau_{ij}) \ge 0.$$

By Proposition 1, maximum of v_{ij} is attained on the boundary. But on the boundary, we have

$$v_{ij} = z_{ij} - Q_{ij} = -Q_{ij} \le 0.$$

Thus,

$$z_{ij} \le Q_{ij} \quad \text{for} \quad i, j = 0, 1, \ldots, N.$$

Similarly, if $v_{ij} = z_{ij} + Q_{ij}$, then, according to (A.5) and (A.6)

$$v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} - 4v_{i,j} = h^2(-E - \tau_{ij}) \le 0.$$

By Proposition 2, minimum of v_{ij} is attained on the boundary. But on the boundary, we have

$$v_{ij} = z_{ij} + Q_{ij} = Q_{ij} \ge 0.$$

Hence,

$$-Q_{ij} \le z_{ij} \quad \text{for} \quad i,j = 0,1,\ldots,N.$$

Thus,

$$-Q_{ij} \le z_{ij} \le Q_{ij} \quad \Rightarrow \quad |z_{ij}| \le |Q_{ij}| \quad \Rightarrow \quad \max_{0\le i,j\le N} |z_{ij}| \le \max_{0\le i,j\le N} |Q_{ij}| = \frac{R^2}{4}E$$

It can be shown that

$$\tau_{ij}(h) = \frac{h^2}{24}\left[\frac{\partial^4 u}{\partial x^4}(\xi_i^+, y_j) + \frac{\partial^4 u}{\partial x^4}(\xi_i^-, y_j) + \frac{\partial^4 u}{\partial y^4}(x_i, \mu_j^+) + \frac{\partial^4 u}{\partial y^4}(x_i, \mu_j^-)\right].$$

Let $M = \max\{M_1, M_2\}$ where

$$M_1 = \max_{(x,y)\in\mathcal{D}} \left|\frac{\partial^4 u}{\partial x^4}\right|, \quad M_2 = \max_{(x,y)\in\mathcal{D}} \left|\frac{\partial^4 u}{\partial y^4}\right|.$$

Then

$$E = \max_{0\le i,j\le N} |\tau_{ij}(h)| \le \frac{Mh^2}{6}.$$

Finally, we obtain

$$\max_{0\le i,j\le N} |z_{ij}| \le \frac{R^2 M}{24}h^2.$$