I

PROBLEMS WITH PERIODIC SOLUTIONS

COPYRIGHTED MATERIAL

1

MODEL EQUATIONS

In this chapter, we examine several model equations to introduce some basic properties of differential equations and difference approximations by example. Generalizations of these ideas are discussed throughout the remainder of this book.

1.1. PERIODIC GRIDFUNCTIONS AND DIFFERENCE OPERATORS

Let $h = 2\pi/(N+1)$, where N is a natural number, denote a grid interval. A grid on the x-axis is defined to be the set of gridpoints

$$x_j = jh, \quad j = 0, \pm 1, \pm 2, \ldots$$

A discrete, possibly complex valued, function u defined on the grid is called a *gridfunction* (see Figure 1.1.1). Here, we are only interested in 2π-periodic gridfunctions, that is,

$$u_j = u(x_j) = u(x_j + 2\pi) = u_{j+N+1}.$$

Clearly, the product and sum of gridfunctions are again gridfunctions. Their gridvalues are

$$(uv)_j = u_j v_j, \quad (u+v)_j = u_j + v_j.$$

We denote the set of all 2π-periodic gridfunctions by P_h. If $u, v \in P_h$, then uv, $u + v \in P_h$.

We now introduce difference operators. They play a fundamental role throughout the book. We start with the translation operator E. It is defined by

$$(Ev)_j = v_{j+1}.$$

Time-Dependent Problems and Difference Methods, Second Edition.
Bertil Gustafsson, Heinz-Otto Kreiss, Joseph Oliger.
© 2013 John Wiley & Sons, Inc. Published 2013 by John Wiley & Sons, Inc.

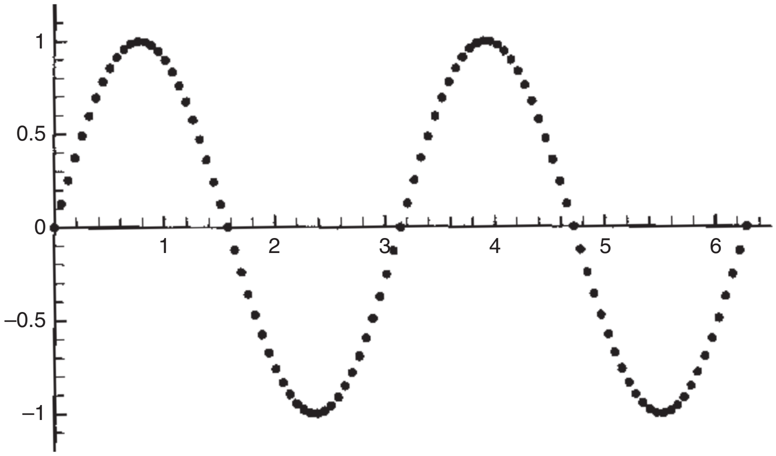


Figure 1.1.1. A gridfunction.

If $v \in P_h$, then $Ev \in P_h$. Powers of E are defined recursively,

$$E^p v = E^{p-1}(Ev).$$

Thus,

$$(E^p v)_j = v_{j+p}. \tag{1.1.1}$$

The inverse also exists and

$$(E^{-1}v)_j = v_{j-1}.$$

If we define E^0 by $E^0 v = v$, then Eq. (1.1.1) holds for all integers p. E is a linear operator and

$$(aE^p + bE^q)v = aE^p v + bE^q v.$$

The forward, backward, and central difference operators are defined by

$$\begin{aligned} D_+ &= (E - E^0)/h, \\ D_- &= (E^0 - E^{-1})/h = E^{-1}D_+, \\ D_0 &= (E - E^{-1})/(2h) = \tfrac{1}{2}(D_+ + D_-), \end{aligned} \tag{1.1.2}$$

respectively. In particular, consider these operators acting on the functions $e^{i\omega x}$. Then, we have for all $x = x_j$

$$\begin{aligned} hD_+e^{i\omega x} &= (e^{i\omega h} - 1)e^{i\omega x} = \left(i\omega h + \mathscr{O}(\omega^2h^2)\right)e^{i\omega x}, \\ hD_-e^{i\omega x} &= (1 - e^{-i\omega h})e^{i\omega x} = \left(i\omega h + \mathscr{O}(\omega^2h^2)\right)e^{i\omega x}, \\ hD_0e^{i\omega x} &= i\sin(\omega h)e^{i\omega x} = \left(i\omega h + \mathscr{O}(\omega^3h^3)\right)e^{i\omega x}. \end{aligned} \tag{1.1.3}$$

Thus,

$$\begin{aligned} \left|\left(D_+ - \frac{\partial}{\partial x}\right)e^{i\omega x}\right| &= \mathscr{O}(\omega^2h), \\ \left|\left(D_- - \frac{\partial}{\partial x}\right)e^{i\omega x}\right| &= \mathscr{O}(\omega^2h), \\ \left|\left(D_0 - \frac{\partial}{\partial x}\right)e^{i\omega x}\right| &= \mathscr{O}(\omega^3h^2). \end{aligned} \tag{1.1.4}$$

Consequently, one says that D_+ and D_- are first-order accurate approximations of $\partial/\partial x$ because the error is proportional to h. D_0 is second-order accurate.

Higher derivatives are approximated by products of the above operators. For example,

$$(D_+D_-v)_j = (D_-D_+v)_j = h^{-2}\left((E - 2E^0 + E^{-1})v\right)_j = h^{-2}(v_{j+1} - 2v_j + v_{j-1}).$$

In particular,

$$\begin{aligned} h^2D_+D_-e^{i\omega x} &= (e^{i\omega h} - 2 + e^{-i\omega h})e^{i\omega x} = -4\sin^2\left(\frac{\omega h}{2}\right)e^{i\omega x} \\ &= \left(-\omega^2h^2 + \mathscr{O}(\omega^4h^4)\right)e^{i\omega x}. \end{aligned} \tag{1.1.5}$$

Therefore,

$$\left|\left(D_+D_- - \frac{\partial^2}{\partial x^2}\right)\ e^{i\omega x}\right| = \mathscr{O}(\omega^4h^2),$$

and D_+D_- is a second-order accurate approximation of $\partial^2/\partial x^2$. Note that all of the above operators commute, because they are all defined in terms of powers of E.

We need to define norms for finite-dimensional vector spaces and discuss some of their properties. We begin with the usual Euclidean inner product and norm. Consider the m-dimensional vector space consisting of all $u = (u^{(1)}, \ldots, u^{(m)})^T$

where $u^{(j)},\ j = 1, \ldots, m$, are complex numbers. We denote the conjugate transpose of u by u^* ($u^* = u^T$ if u is real). The inner product and norm are defined by

$$\langle u, v\rangle = u * v = \sum_{j=1}^{m} \overline{u}^{(j)} v^{(j)}, \quad \text{and} \quad |u| = \langle u, u\rangle^{1/2}, \tag{1.1.6}$$

respectively. The inner product is a bilinear form that satisfies the following equalities:

$$\begin{aligned} \langle u, v\rangle &= \overline{\langle v, u\rangle}, \\ \langle u + w, v\rangle &= \langle u, v\rangle + \langle w, v\rangle, \\ \langle u, \lambda v\rangle &= \lambda\langle u, v\rangle, \quad \lambda \text{ a complex number}, \\ \langle \lambda u, v\rangle &= \overline{\lambda}\langle u, v\rangle. \end{aligned} \tag{1.1.7}$$

The following inequalities hold:

$$\begin{aligned} |\langle u, v\rangle| &\le |u|\ |v|, \\ |u + v| &\le |u| + |v|, \\ \| u| - |v\| &\le |u - v|, \\ \langle u, v\rangle &\le |u| \cdot |v| \le \delta |u|^2 + \frac{1}{4\delta}\ |v|^2 \quad \text{for any } \delta > 0. \end{aligned} \tag{1.1.8}$$

Let $A = (a_{ij})$ be a complex $m \times m$ matrix. Then, its transpose is denoted by $A^T = (a_{ji})$ and its conjugate transpose by $A^* = (\overline{a}_{ji})$. The Euclidean norm of the matrix A is defined by

$$|A| = \max_{|u|=1}\ |Au|,$$

where the norm on the right-hand side is the vector norm defined above. If A and B are matrices, then

$$\begin{aligned} |Au| &\le |A|\ |u|, \\ |A + B| &\le |A| + |B|, \\ |AB| &\le |A|\ |B|. \end{aligned} \tag{1.1.9}$$

If the scalar λ and vector $u \neq 0$ satisfy $Au = \lambda u$, then λ is an eigenvalue of A and u is the corresponding eigenvector. The spectral radius, $\rho(A)$, of a matrix A is defined by

$$\rho(A) = \max_j |\lambda_j|,$$

where the λ_j are the eigenvalues of A. The spectral radius satisfies the inequality

$$\rho(A) \le |A|. \tag{1.1.10}$$

We next define a scalar product and norm for our periodic gridfunctions of length $N + 1$. For fixed h and $N + 1$, these functions form a vector space. However, we are interested in these functions as $h \to 0$ and $N(h) + 1 \to \infty$. The Euclidean inner product and norm defined above would not necessarily be finite in this limit, so we must use a different definition.

We define a discrete scalar product and norm for periodic gridfunctions by

$$(u, v)_h = \sum_{j=0}^{N} \overline{u}_j v_j h \quad \text{and} \quad \|u\|_h = \sqrt{(u, u)_h}, \tag{1.1.11}$$

respectively.

The scalar product is also a bilinear form and satisfies the same equalities as the Euclidean inner product for vectors in Eq. (1.1.7):

$$\begin{aligned} (u, v)_h &= \overline{(v, u)_h}, \\ (u + w, v)_h &= (u, v)_h + (w, v)_h, \\ (u, \lambda v)_h &= \lambda (u, v)_h, \quad \lambda \text{ a complex number}, \\ (\lambda u, v)_h &= \overline{\lambda}(u, v)_h. \end{aligned} \tag{1.1.12}$$

The following inequalities also hold in analogy with Eq. (1.1.8):

$$\begin{aligned} |(u, v)_h| &\le \|u\|_h \|v\|_h, \\ |(u, av)_h| &\le \|a\|_\infty \|u\|_h \|v\|_h, \quad \|a\|_\infty = \max_j |a_j|, \\ \|u + v\|_h &\le \|u\|_h + \|v\|_h, \\ |\|u\|_h - \|v\|_h| &\le \|u - v\|_h. \end{aligned} \tag{1.1.13}$$

For periodic functions $f(x)$, $g(x)$ defined everywhere, the L_2 scalar product and norm are defined by

$$(f, g) = \int_0^{2\pi} \overline{u}(x) v(x)\, dx\,, \qquad ||f|| = \sqrt{(f, f)}.$$

A function $f(x)$ with finite norm $||f||$ is called an L_2 *function*.

If u, v are the projections of continuous functions onto the grid, then

$$\lim_{h \to 0} (u, v)_h = (u, v), \quad \lim_{h \to 0} \|u\|_h^2 = \|u\|^2,$$

converge to the L_2 scalar product and norm. Therefore, the above-mentioned inequalities are also valid for the L_2 scalar product and norm applied to C^1 functions. Because any function $\in L_2$ can be approximated arbitrarily well by a C^1 function, they are valid for all L_2 functions as well.

The norm of an operator is defined in the usual way,

$$\|Q\|_h = \sup_{u\neq 0} \frac{\|Qu\|_h}{\|u\|_h} = \sup_{\|u\|_h=1} \|Qu\|_h.$$

From this definition, it follows that $\|Qu\|_h \le \|Q\|_h\|u\|_h$. Thus,

$$\|E^p u\|_h^2 = \sum_{j=0}^{N} |u_{j+p}|^2 h = \sum_{j=0}^{N} |u_j|^2 h = \|u\|_h^2$$

implies

$$\|E^p\|_h = 1, \quad p = 0, \pm 1, \pm 2, \ldots \tag{1.1.14}$$

Also,

$$\|D_+ u\|_h = \frac{1}{h} \|(E - E^0)u\|_h \le \frac{2}{h} \|u\|_h,$$

that is,

$$\|D_+\|_h \le 2/h.$$

The general inequalities

$$\|P + Q\|_h \le \|P\|_h + \|Q\|_h, \qquad \|PQ\|_h \le \|P\|_h\|Q\|_h \tag{1.1.15}$$

give us

$$\|D_-\|_h = \|E^{-1}D_+\|_h \le \frac{2}{h}, \qquad \|D_0\|_h = \frac{1}{2h} \|E - E^{-1}\|_h \le \frac{1}{h}.$$

Actually, these inequalities for the norms of D_+, D_-, and D_0 can be replaced by equalities. For D_+, we define $u_j = (-1)^j$ and obtain

$$\begin{aligned} \|u\|_h^2 &= (N+1)h, \\ \|D_+u\|_h^2 &= \sum_{j=0}^{N} \left((-1)^{j+1} - (-1)^j\right)^2 h^{-1} = 4(N+1)h^{-1} = \frac{4}{h^2} \|u\|_h^2, \end{aligned}$$

which yields

$$\|D_+\|_h = 2/h. \tag{1.1.16}$$

Using the same gridfunction u_j again, we get

$$\|D_-\|_h = 2/h. \tag{1.1.17}$$

For D_0, we choose $u_j = i^j$ (where $i = \sqrt{-1}$) and obtain

$$\|u\|_h^2 = (N+1)h,$$

$$\|D_0 u\|_h^2 = \sum_{j=0}^{N} \frac{1}{4h}\left((-1)^{j+1} - (-i)^{j-1}\right)(i^{j+1} - i^{j-1}) = \frac{N+1}{h} = \frac{1}{h^2}\ \|u\|_h^2,$$

so

$$\|D_0\|_h = 1/h. \tag{1.1.18}$$

We now consider systems of partial differential equations and consequently need to define a norm and scalar product for vector-valued gridfunctions $u = (u^{(1)}, \ldots, u^{(m)})^T$. Let u and v be two such vector-valued gridfunctions, then we define

$$(u, v)_h = \sum_{j=0}^{N} \langle u_j, v_j \rangle h, \qquad \|u\|_h = \sqrt{(u, u)_h}. \tag{1.1.19}$$

The properties shown in Eqs. (1.1.12) and (1.1.13) are still valid. We can also generalize the second inequality in Eq. (1.1.13) when a is replaced by an $(m \times m)$ matrix A. If A is a constant matrix, we have

$$|(Au, v)_h| \le |A|\ \|u\|_h \|v\|_h, \tag{1.1.20}$$

If $A = A_j$ is a matrix-valued gridfunction, then

$$|(Au, v)_h| \le \max_j |A_j|\ \|u\|_h \|v\|_h. \tag{1.1.21}$$

EXERCISES

1.1.1. Derive estimates for

$$\left|\left(D - \frac{\partial^3}{\partial x^3}\right) e^{i\omega x}\right|,$$

where $D = D_+^3,\ D_- D_+^2,\ D_-^2 D_+,\ D_-^3,\ D_0 D_+ D_-$.

1.1.2. Both the difference operators D_+ and D_0 approximate $\partial/\partial x$, but they have different norms. Explain why this is not a contradiction.

1.1.3. Compute $\|D_+ D_-\|_h$.

1.2. FIRST-ORDER WAVE EQUATION, CONVERGENCE, AND STABILITY

The equation $u_t = u_x$ is the simplest *hyperbolic* equation; the general definition of the class of hyperbolic equations is given in Section 3.3. We consider the initial value problem

$$\begin{aligned} u_t &= u_x, && -\infty < x < \infty,\ t \geq 0, \\ u(x,0) &= f(x), && -\infty < x < \infty, \end{aligned} \tag{1.2.1}$$

where $f(x) = f(x + 2\pi)$ is a smooth 2π-periodic function. To begin, we assume that the initial function

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{i\omega x} \hat{f}(\omega)$$

consists of one wave. The integer ω is called the *wave number* or the *frequency*. We try to find a solution of the same type

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\, e^{i\omega x} \hat{u}(\omega, t) \tag{1.2.2}$$

with $\hat{u}(\omega, 0) = \hat{f}(\omega)$. Substituting Eq. (1.2.2) into Eq. (1.2.1) yields an initial value problem for the ordinary differential equation

$$\begin{aligned} \frac{d\hat{u}}{dt} &= i\omega\hat{u}, \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned}$$

which is called the *Fourier transform* of Eq. (1.2.1). Therefore,

$$\hat{u}(\omega, t) = e^{i\omega t} \hat{u}(\omega, 0) = e^{i\omega t} \hat{f}(\omega).$$

It follows that

$$u(x,t) = \frac{1}{\sqrt{2\pi}}\, e^{i\omega(x+t)} \hat{f}(\omega) = f(x+t) \tag{1.2.3}$$

is a solution to our problem.

Now consider the general case

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x} \hat{f}(\omega), \tag{1.2.4}$$

which is the Fourier series representation as described in Section A.1. By the superposition principle,

$$u(x,t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega(x+t)} \hat{f}(\omega) = f(x+t) \tag{1.2.5}$$

is a solution to our problem. For every fixed t, Parseval's relation (A.1.9) yields

$$\|u(\cdot,t)\|^2 = \sum_{\omega=-\infty}^{\infty} |e^{i\omega t}\hat{f}(\omega)|^2 = \sum_{\omega=-\infty}^{\infty} |\hat{f}(\omega)|^2 = \|f(\cdot)\|^2. \tag{1.2.6}$$

The squared norm $\|u\|^2$ is often called the energy of u. Therefore, the differential equation in Eq. (1.2.1) is said to be energy conserving; the obvious phrase norm conserving is often used in this context as well. Clearly, any method of approximation must be nearly norm conserving to be useful. We also note that there is a *finite speed of propagation* associated with this problem. The expression (1.2.5) shows that the solution is constant along the lines $x + t =$ const, which are called *characteristics* (see Figure 1.2.1).

Any particular feature of the initial data, such as a wave crest, is propagated along these characteristics. In our case, the speed of propagation (or wave speed) is $dx/dt = -1$. For general hyperbolic systems, there may be many families of characteristics corresponding to different wave speeds of different components. The important thing is that these speeds are always finite.

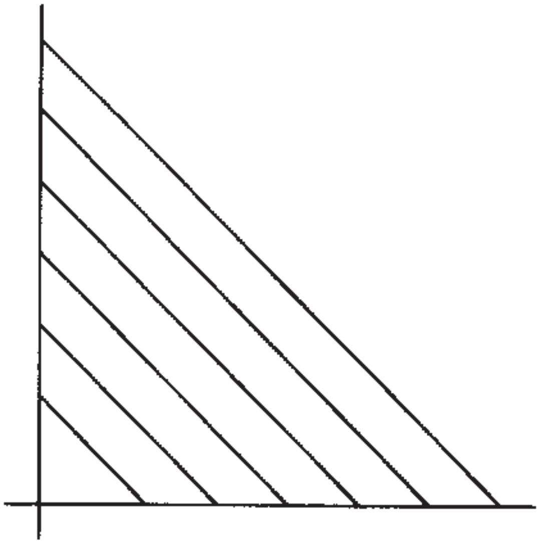


Figure 1.2.1. Characteristics.

We now solve the problem using a difference approximation. We introduce a space step $h = 2\pi/(N+1)$, with N a natural number, and a time step $k > 0$. The space and time steps h, k define a grid in x, t space, consisting of the gridpoints $(x_j, t_n) := (jh, nk)$. Gridfunctions will be denoted by $u_j^n = u(x_j, t_n)$. A simple approximation based on forward differences in time and centered differences in space is

$$\begin{aligned} v_j^{n+1} &= (I + kD_0)v_j^n =: Qv_j^n, \qquad j = 0,\ \pm1,\ \pm2,\ \ldots \\ v_j^0 &= f_j. \end{aligned} \tag{1.2.7}$$

If v^n is known at time $t_n = nk$, then we can use Eq. (1.2.7) to calculate v_j^{n+1} for all j. Thus, the initial data determine a unique solution, and we call such a method a *one-step method*. Also, if v^n is 2π-periodic, then v^{n+1} is too. Therefore, we can restrict the calculation to $j = 0, 1, 2, \ldots, N$ and use periodicity conditions to extend the solution and provide the extra needed values for Eq. (1.2.7) at $j = 0, N$, that is, $v_{-1}^n = v_N^n,\ v_{N+1}^n = v_0^n$.

We will now calculate the solution analytically. First, consider the case where f consists of one single wave, that is,

$$f_j = \frac{1}{\sqrt{2\pi}}\, e^{i\omega x_j}\hat{f}(\omega), \quad j = 0, 1, 2, \ldots, N.$$

As in the continuous case, we make the ansatz

$$v_j^n = \frac{1}{\sqrt{2\pi}}\ \hat{v}^n(\omega)e^{i\omega x_j}, \tag{1.2.8}$$

that is, we assume that the solution can also be expressed in terms of one single Fourier component. Substituting Eq. (1.2.8) into Eq. (1.2.7) yields

$$e^{i\omega x_j}\hat{v}^{n+1}(\omega) = \left(e^{i\omega x_j} + \frac{\lambda}{2}\ (e^{i\omega x_{j+1}} - e^{i\omega x_{j-1}})\right)\ \hat{v}^n(\omega),$$

where $\lambda = k/h$. This equation can be rewritten as

$$e^{i\omega x_j}\hat{v}^{n+1}(\omega) = (1 + i\lambda \sin\ \xi)e^{i\omega x_j}\hat{v}^n(\omega),$$

where $\xi = \omega h$, and we get

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \qquad \hat{Q} = 1 + i\lambda \sin\ \xi. \tag{1.2.9}$$

The complex number $\hat{Q}$ is the *Fourier transform* of $(I + kD_0)$, and Eq. (1.2.9) is the Fourier transform of Eq. (1.2.7). We also call $\hat{Q}$ the *symbol*, or the *amplification factor*. Actually, it is the *discrete* Fourier transform which is further discussed in Appendix A. The solution of Eq. (1.2.9) is

$$\hat{v}^n(\omega) = \hat{Q}^n\hat{v}^0(\omega) = \hat{Q}^n\hat{f}(\omega),$$

and it is clear that

$$v_j^n = \frac{1}{\sqrt{2\pi}}\ \hat{Q}^n e^{i\omega x_j} \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}}\ \left(1 + i\ \frac{k}{h} \sin(\omega h)\right)^n\ e^{i\omega x_j} \hat{f}(\omega)$$

solves our problem.

Now, we consider a sequence of grid intervals $k, h \to 0$. We want to show that v_j^n converges to the corresponding solution of the differential equation. We have

$$\begin{aligned}\left(1 + i\ \frac{k}{h} \sin(\omega h)\right)^n &= \left(1 + i\omega k + \mathscr{O}(kh^2\omega^3)\right)^n = \left(e^{i\omega k} + \mathscr{O}(k^2\omega^2 + kh^2\omega^3)\right)^n \\ &= \left(1 + \mathscr{O}\left((k\omega^2 + h^2\omega^3)t_n\right)\right) e^{i\omega t_n}.\end{aligned}$$

Therefore,

$$v_j^n = \frac{1}{\sqrt{2\pi}}\ \left(1 + \mathscr{O}\left((k\omega^2 + h^2\omega^3)t_n\right)\right) e^{i\omega(x_j + t_n)} \hat{f}(\omega).$$

Thus, for every fixed ω, we obtain

$$\lim_{k,h\to 0}\ v_j^n = u(x_j, t_n)$$

in any finite interval $0 \leq t \leq T$.

Now assume that the initial data are represented by a trigonometric polynomial

$$u(x, 0) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-M}^{M} e^{i\omega x} \hat{f}(\omega).$$

By the superposition principle, the above result implies that the solution of the difference approximation will converge to the solution of the differential equation as $k, h \to 0$. Thus, one might think that the approximation could be useful in practice. However, consider the problem (1.2.1) with initial data $f(x) \equiv 0$ which has the trivial solution $u(x, t) \equiv 0$. Now consider the problem with perturbed data

$$\hat{f}(\omega) = \begin{cases} \varepsilon, & \text{for } \omega = N/4, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding solution of the transformed difference approximation is

$$\hat{v}^n(N/4) = \left(1 + i\frac{k}{h}\ \sin\left(\frac{2\pi}{N+1}\frac{N}{4}\right)\right)^n \varepsilon \sim \left(1 + i\frac{k}{h}\right)^n \varepsilon,$$

that is,

$$|\hat{v}^{t_n/k}(N/4)|^2 \sim \left(1+\frac{k^2}{h^2}\right)^{t_n/k}\varepsilon^2.$$

For $t_n = 1$, that is, $n = 1/k$

$$|\hat{v}^{1/k}(N/4)|^2 \sim \left(1+\frac{k^2}{h^2}\right)^{1/k}\varepsilon^2.$$

Now consider any sequence $k, h \to 0$ with $k/h = \lambda > 0$ fixed. Then,

$$\lim_{k\to 0}|\hat{v}^{1/k}(N/4)| = \infty.$$

This "explosion," or growth, can be arbitrarily fast. For example, if we consider $\lambda = 10, k = 10^{-5}$, then

$$|\hat{v}^{1/k}(N/4)|^2 \sim 100^{10^5}\varepsilon^2.$$

The numerical calculation is therefore worthless. In Figure 1.2.2, we have calculated the maximum of the solutions of the difference approximation (1.2.7) with initial data

$$f_j = \begin{cases} x_j, & \text{for } 0 \le x_j \le \pi, \\ 2\pi - x_j, & \text{for } \pi \le x_j \le 2\pi, \end{cases}$$

and stepsizes $h = 0.01,\ k = 0.01$ and $h = 0.01,\ k = 0.1$, respectively.

The analytic results lead us to expect that the solutions will grow like $2^{n/2}$ and $101^{n/2}$, respectively. The numerical results confirm that prediction.

In realistic computations, one must always expect perturbations, either from measurement errors in the data or from rounding errors due to the finite representation of numbers in the computer. Therefore, we must require that $|\hat{Q}^n|$ is bounded independently of h and k, and we call such methods *stable*. (We make the formal definition of this concept later.)

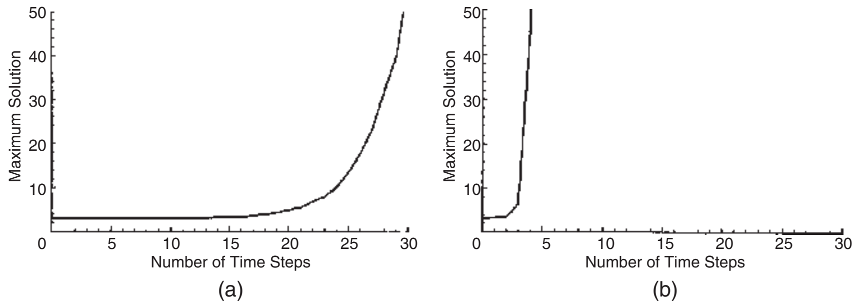


Figure 1.2.2. $\max_j |v_j^n|$, v_j^n solution of Eq. (1.2.7). (a) $h = 0.01$; $k = 0.01$ and (b) $h = 0.01$; $k = 0.1$.

Next, we modify our previous difference approximation by adding *artificial viscosity*, that is, we consider

$$v_j^{n+1} = (I + kD_0)v_j^n + \sigma k h D_+ D_- v_j^n, \qquad v_j^0 = f_j. \tag{1.2.10}$$

Here, $\sigma > 0$ is a constant, which we will choose later. We can write Eq. (1.2.10) in the form

$$\frac{v_j^{n+1} - v_j^n}{k} = D_0 v_j^n + \sigma h D_+ D_- v_j^n, \tag{1.2.11}$$

which approximates the differential equation

$$u_t = u_x + \sigma h u_{xx}.$$

As $h \to 0$, we obtain Eq. (1.2.1). Thus, Eq. (1.2.10) is a *consistent* difference approximation of (1.2.1), that is, the difference approximation converges formally to the differential equation as $k, h \to 0$.

We will now choose σ, k and h so that

$$|\hat{Q}| \le 1. \tag{1.2.12}$$

In this case, all powers $|\hat{Q}^n|$ are certainly bounded as required for stability.

From Eqs. (1.1.3) and (1.1.5), $\hat{Q}$ is of the form

$$\hat{Q} = 1 + i\lambda \sin\xi - 4\sigma\lambda \sin^2\frac{\xi}{2}, \qquad \xi = \omega h,\ \lambda = k/h.$$

Therefore,

$$\begin{aligned}
|\hat{Q}|^2 &= \left(1 - 4\sigma\lambda \sin^2\frac{\xi}{2}\right)^2 + \lambda^2 \sin^2\xi \\
&= 1 - 8\sigma\lambda \sin^2\frac{\xi}{2} + 16\sigma^2\lambda^2 \sin^4\frac{\xi}{2} + 4\lambda^2 \sin^2\frac{\xi}{2}\left(1 - \sin^2\frac{\xi}{2}\right) \\
&= 1 - (8\sigma\lambda - 4\lambda^2)\sin^2\frac{\xi}{2} + (16\sigma^2 - 4)\lambda^2 \sin^4\frac{\xi}{2}.
\end{aligned}$$

There are two ways to satisfy Eq. (1.2.12):

1. Suppose $2\sigma \le 1$. If $0 \le 8\sigma\lambda - 4\lambda^2$, that is,

$$0 < \lambda \le 2\sigma \le 1, \tag{1.2.13}$$

then $|\hat{Q}| \le 1$. By letting $|\xi|$ be small, we see that these conditions are also necessary.

2. Suppose $1 \le 2\sigma$. If we replace $\sin^4(\xi/2)$ by $\sin^2(\xi/2)$, it follows that $|\hat{Q}| \le 1$ if

$$0 \le 8\sigma\lambda - 4\lambda^2 - 16\sigma^2\lambda^2 + 4\lambda^2,$$

that is,

$$1 \le 2\sigma, \qquad 2\sigma\lambda \le 1. \tag{1.2.14}$$

By letting $\sin(\xi/2) = 1$, we see that these conditions are also necessary.

There are two particular schemes of the above-mentioned type that have been used extensively:

1. *The Lax–Friedrichs method* ($\sigma = h/2k = 1/(2\lambda)$).

$$v_j^{n+1} = \tfrac{1}{2}(v_{j+1}^n + v_{j-1}^n) + kD_0v_j^n = (I + kD_0)v_j^n + \tfrac{1}{2}h^2D_+D_-v_j^n. \tag{1.2.15}$$

In this case, Eq. (1.2.14) is satisfied if $k/h \le 1$, that is, $|\hat{Q}| \le 1$. It is remarkable that the simple change $v_j^n \to \frac{1}{2}(v_{j+1}^n + v_{j-1}^n)$ has such an effect on the solution.

2. *The Lax–Wendroff method* ($\sigma = k/2h = \lambda/2$).

$$v_j^{n+1} = v_j^n + kD_0v_j^n + \frac{k^2}{2}\,D_+D_-v_j^n. \tag{1.2.16}$$

Now Eq. (1.2.13) is satisfied if $k/h \le 1$.

In Figure 1.2.3, we have used the Lax–Friedrichs method and the Lax–Wendroff method to calculate the solution of Eq. (1.2.1) with initial data

$$f(x) = \begin{cases} x, & \text{for } 0 \le x \le \pi, \\ 2\pi - x, & \text{for } \pi \le x \le 2\pi, \end{cases}$$

and $k/h = 1/2,\ h = 2\pi/10,\ 2\pi/100$, respectively.

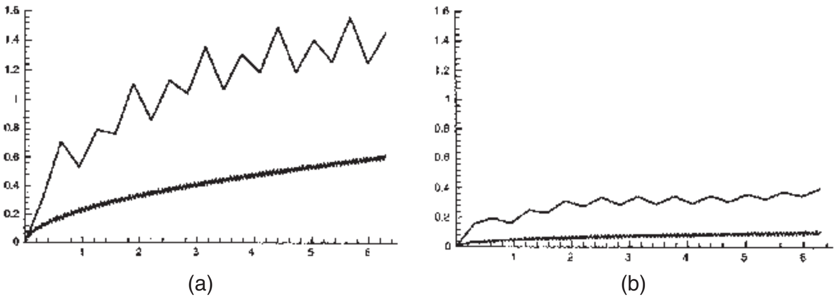
We show here the absolute maximum error plotted against time. The accuracy is not impressive, but there is no explosion.

We now consider a rather general difference approximation of the problem (1.2.1):

$$\begin{aligned} v_j^{n+1} &= Qv_j^n, \qquad Q = \sum_{\nu=-r}^{s} A_\nu(k,h)E^\nu, \\ v_j^0 &= f_j. \end{aligned} \tag{1.2.17}$$

Here, the A_ν are rational functions of k and h, and r, $s \ge 0$ are integers. Thus, we use the $s + r + 1$ values $v_{j-r}^n, \dots, v_{j+s}^n$ to calculate v_j^{n+1}. We again consider simple wave solutions

$$v_j^n = \frac{1}{\sqrt{2\pi}}\,e^{i\omega x_j}\,\hat{v}^n(\omega).$$



(a)

(b)

Figure 1.2.3. (a) The Lax–Friedrichs and (b) the Lax–Wendroff methods for Eq. (1.2.1).

By observing that $Ee^{i\omega x} = e^{i\xi}e^{i\omega x}$, we obtain

$$\hat{v}^{n+1}(\omega) = \hat{Q}(\xi)\hat{v}^n(\omega), \qquad \hat{Q} = \sum_{\nu=-r}^{s} A_\nu e^{i\nu\xi},$$

that is,

$$\hat{v}^n(\omega) = \hat{Q}^n(\xi)\hat{v}^0(\omega), \tag{1.2.18}$$

where $\hat{Q}$ is the symbol of Q.

We assume that the initial data belongs to L_2, that is, $f(x)$ can be expanded as a Fourier series

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} \hat{f}(\omega)e^{i\omega x}, \qquad \sum_{\omega} |\hat{f}(\omega)|^2 < \infty. \tag{1.2.19}$$

For the difference approximation, we use the restriction of $f(x)$ to the grid. We denote by

$$\text{Int}_N f = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \tilde{f}(\omega)e^{i\omega x} \tag{1.2.20}$$

the trigonometric interpolant of the gridfunction (see Section A.2). We assume that

$$\lim_{N\to\infty} \|\text{Int}_N f - f\| = 0. \tag{1.2.21}$$

From Theorem A.2.4, this convergence condition is satisfied if f is a smooth function.

If the differential equation is modified to

$$u_t = u_x + au,$$

the corresponding Fourier transform is

$$\frac{d\hat{u}}{dt} = i\omega\hat{u} + a\hat{u},$$

which has the solution

$$\hat{u}(\omega, t) = e^{(i\omega+a)t}\hat{u}(\omega, 0).$$

In such a case, we must allow an exponential growth of the approximate solution as well. Therefore, we define stability in the following way:

Definition 1.2.1. *The approximation* (*1.2.1*) *is called stable if there are constants* K, α *independent of* k, h *such that the symbol satisfies*

$$|\hat{Q}^n| \leq K e^{\alpha t_n}. \tag{1.2.22}$$

We now want to prove the following theorem:

Theorem 1.2.1. *Consider the difference approximation shown in Eq.* (*1.2.17*) *on a finite interval* $0 \leq t \leq T$ *for a sequence* $h, k \to 0$. *Assume that*

1. *The initial data satisfy Eqs.* (*1.2.19*) *and* (*1.2.21*).
2. *The approximation is stable, and*

$$\sup_{0\leq t_n \leq T} |\hat{Q}^n| \leq K_S.$$

3. *The approximation is consistent, that is, for every fixed* ω,

$$\lim_{k,h\to 0} \sup_{0\leq t_n\leq T} |\hat{Q}^n(\xi) - e^{i\omega t_n}| = 0.$$

Then, the trigonometric interpolant $Int_N v$ *of the solution of the difference approximation converges to the solution of the differential equation,*

$$\lim_{k,h\to 0} \sup_{0\leq t_n\leq T} \|u(\cdot, t_n) - Int_N(v_j^n)\| = 0.$$

Proof. For every fixed t_n, we can represent the solution of the difference approximation by its trigonometric interpolant and, therefore, we can think of the solution as being represented in terms of simple waves. From Eq. (1.2.18), we obtain

$$\text{Int}_N(v_j^n) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-N/2}^{N/2} \hat{Q}^n(\xi)\tilde{f}(\omega)e^{i\omega x}.$$

Let $0 < M < N/2$ be a fixed integer. From Eq. (1.2.5) and Parseval's relation, we obtain

$$\|u(\cdot, t_n) - \mathrm{Int}_N(v_j^n)\|^2 = \sum_{\omega=-N/2}^{N/2} |e^{i\omega t_n}\hat{f}(\omega) - \hat{Q}^n(\xi)\tilde{f}(\omega)|^2$$

$$+ \sum_{|\omega|>N/2} |\hat{f}(\omega)|^2 \leq I + II + III,$$

where

$$I = \sum_{\omega=-M}^{M} |\hat{Q}^n(\xi)\tilde{f}(\omega) - e^{i\omega t_n}\hat{f}(\omega)|^2,$$

$$II = 2\sum_{|\omega|>M} |\hat{f}(\omega)|^2,$$

$$III = 2\sum_{|\omega|>M} |\tilde{f}(\omega)|^2 |\hat{Q}^n(\xi)|^2.$$

By Eq. (1.2.19),

$$\lim_{M\to\infty} II = 0.$$

From Eq. (1.2.21) and the second assumption,

$$\lim_{M\to\infty} III \leq 4K_s^2 \lim_{M\to\infty} \sum_{|\omega|>M} \left(|\tilde{f}(\omega) - \hat{f}(\omega)|^2 + |\hat{f}(\omega)|^2\right) = 0.$$

Finally, for every fixed M, the second and third assumptions together with Eq. (1.2.21) imply that

$$\lim_{N\to\infty} I \leq 2 \lim_{N\to\infty} \sum_{\omega=-M}^{M} \left(|\hat{Q}^n(\xi)\left(\tilde{f}(\omega) - \hat{f}(\omega)\right)|^2 + |\left(\hat{Q}^n(\xi) - e^{i\omega t_n}\right)\hat{f}(\omega)|^2\right) = 0.$$

Now convergence follows easily. Let $\varepsilon > 0$ be an arbitrarily small constant. Choose M so large that $II + III < \varepsilon/2$. For sufficiently large N, we also have $I \leq \varepsilon/2$ and, therefore, convergence follows. This proves the theorem.

This theorem tells us that the solution of the difference approximation converges to the solution of the differential equation if the approximation is stable and consistent. In actual calculations, one uses fixed values of k and h. Convergence of the solution should therefore be considered as a guarantee that a certain approximation becomes more accurate if we choose a smaller stepsize.

EXERCISES

1.2.1. The convergence of the solutions in Figure 1.2.3 is rather slow. Explain why that is so and find which one of the terms *I, II*, or *III* is large for this example in the proof of Theorem 1.2.1.

1.2.2. Modify the scheme (1.2.10) such that it approximates $u_t = -u_x$. Prove that the conditions (1.2.13) and (1.2.14) are also necessary for stability in this case.

1.2.3. Choose σ in Eq. (1.2.10) such that Q uses only two gridpoints. What is the stability condition?

1.3. LEAP-FROG SCHEME

The difference approximations that we discussed Section 1.2 were all one-step methods, that is, v_j^{n+1} could be expressed as a linear combination of neighboring values $v_{j-r}^n, \dots, v_{j+s}^n$ at the previous time level. The leap-frog scheme

$$v_j^{n+1} = v_j^{n-1} + \lambda(v_{j+1}^n - v_{j-1}^n), \qquad \lambda = k/h, \tag{1.3.1}$$

is a two-step method, which is a special case of *multistep methods*. To determine the new values v_j^{n+1}, we need values at two previous time levels. To start the calculation, we have to specify v_j^0 and v_j^1. The initial data yields $v_j^0 = f_j$, whereas v_j^1 can be determined by a one-step method. It does not need to be stable because we use it only once. The simplest one is Eq. (1.2.7), that is,

$$v_j^1 = (I + kD_0)v_j^0 = (I + kD_0)f_j. \tag{1.3.2}$$

We again seek simple wave solutions

$$v_j^n = \frac{1}{\sqrt{2\pi}}\, e^{i\omega x_j}\, \hat{v}^n(\omega)$$

and obtain

$$\hat{v}^{n+1}(\omega) = \hat{v}^{n-1}(\omega) + 2i\lambda(\sin\xi)\hat{v}^n(\omega). \tag{1.3.3}$$

To solve Eq. (1.3.3), we make the ansatz

$$\hat{v}^n(\omega) = z^n, \tag{1.3.4}$$

where z is a complex number. Substituting Eq. (1.3.4) into Eq. (1.3.3) gives us

$$z^{n+1} = z^{n-1} + 2i\lambda(\sin\xi)z^n,$$

and, therefore, Eq. (1.3.4) is a solution of Eq. (1.3.3) if, and only if, z satisfies the so-called *characteristic equation*

$$z^2 = 1 + 2i\lambda z \sin \xi. \tag{1.3.5}$$

For $0 < \lambda < 1$, Eq. (1.3.5) has two distinct solutions with

$$|z_j| = 1,$$

given by

$$z_{1,2} = i\lambda \sin \xi \pm \sqrt{1 - \lambda^2 \sin^2 \xi}. \tag{1.3.6}$$

The general solution of Eq. (1.3.3) is

$$\hat{v}^n = \sigma_1 z_1^n + \sigma_2 z_2^n. \tag{1.3.7}$$

The parameters σ_1 and σ_2 are determined by the initial data. If $\hat{v}^0(\omega) = \hat{f}(\omega)$, then by Eq. (1.3.2)

$$\hat{v}^1(\omega) = (1 + i\lambda \sin \xi)\hat{f}(\omega),$$

and we obtain the linear system of equations

$$\begin{aligned} \sigma_1 + \sigma_2 &= \hat{f}(\omega), \\ \sigma_1 z_1 + \sigma_2 z_2 &= (1 + i\lambda \sin \xi)\hat{f}(\omega). \end{aligned} \tag{1.3.8}$$

As in the one-step case, we consider the low frequencies with $|\omega h| \ll 1$. Then, if $\lambda = k/h = \text{const}$,

$$\begin{aligned} z_1 &= 1 + i\omega k - \tfrac{1}{2}\omega^2 k^2 + \mathcal{O}(\omega^3 k^3) = e^{i\omega k\left(1+\mathcal{O}(\omega^2 k^2)\right)}, \\ z_2 &= -\left(1 - i\omega k - \tfrac{1}{2}\omega^2 k^2 + \mathcal{O}(\omega^3 k^3)\right) = -e^{-i\omega k\left(1+\mathcal{O}(\omega^2 k^2)\right)}. \end{aligned}$$

After a simple calculation, Eq. (1.3.8) gives us

$$\sigma_1 = \hat{f}(\omega)\left(1 + \mathcal{O}(\omega^2 k^2)\right), \qquad \sigma_2 = \mathcal{O}(\omega^2 k^2)\hat{f}(\omega),$$

and, therefore,

$$\begin{aligned} \hat{v}^n(\omega) = \hat{f}(\omega)&\left(1 + \mathcal{O}(\omega^2 k^2)\right) e^{i\omega t_n\left(1+\mathcal{O}(\omega^2 k^2)\right)} \\ &+ \mathcal{O}(\omega^2 k^2)\hat{f}(\omega)(-1)^n e^{-i\omega t_n\left(1+\mathcal{O}(\omega^2 k^2)\right)}. \end{aligned}$$

Thus, the solution consists of two parts. The first part approximates the corresponding solution $\hat{u}(\omega, t_n) = \hat{f}(\omega)e^{i\omega t_n}$, and the error in the exponent (called the *phase error*) is $\mathcal{O}(\omega^3 k^2 t_n)$. The second part oscillates rapidly and is independent of the differential equation. It is often called a *parasitic* solution. Luckily, the amplitude is small for $\omega^2 k^2 \ll 1$ and does not grow with time.

Because the leap-frog scheme uses three time levels, Theorem 1.2.1 does not apply as formulated. However, using the form of $\hat{v}^n(\omega)$ derived above, we can again use trigonometric interpolation to prove convergence to the solution of the differential equation. Smoothness of the initial data and the *stability condition*

$$\lambda = k/h \le 1 - \delta, \quad \delta > 0 \tag{1.3.9}$$

for any sequence $k, h \to 0$ are required for convergence (see Exercise 1.3.1).

We now discuss a property of the leap-frog scheme that can cause practical difficulties. Let $a > 0$ be a constant, and consider the differential equation

$$u_t = u_x - au.$$

The simple wave solutions are now of the form

$$u = \frac{1}{\sqrt{2\pi}} e^{i\omega(x+t)} e^{-at} \hat{f}(\omega),$$

which clearly decay exponentially with time. We use the approximation

$$v_j^{n+1} = v_j^{n-1} + 2kD_0 v_j^n - 2kav_j^n. \tag{1.3.10}$$

The simple wave solutions again have the form

$$v_j^n = (\sigma_1 z_1^n + \sigma_2 z_2^n) e^{i\omega x_j},$$

where now $z_{1,2}$ are the solutions of

$$z^2 = 1 + (2i\lambda \sin\ \xi - 2ka)z,$$

that is,

$$z_{1,2} = i\lambda \sin\ \xi - ka \pm \sqrt{1 + (i\lambda \sin\ \xi - ka)^2}.$$

Consider the special case $\omega = 0$. For $ka \ll 1$, we have

$$z_1 = 1 - ka + \frac{k^2a^2}{2} + \mathcal{O}(k^3a^3) = e^{-ka+\mathcal{O}(k^3a^3)},$$

$$z_2 = -e^{ka+\mathcal{O}(k^3a^3)},$$

and, as before,

$$\hat{v}^n(0) = \hat{f}(0)\left(1 + \mathcal{O}(k^2a^2)\right) e^{-at_n\left(1+\mathcal{O}(k^2a^2)\right)} + \mathcal{O}(k^2a^2)\hat{f}(\omega)(-1)^n e^{at_n\left(1+\mathcal{O}(k^2a^2)\right)}.$$

Now the parasitic solution grows exponentially and can obliterate the exponentially decaying solution. Therefore, the (unmodified) leap-frog scheme cannot be used for long time intervals. It is easy to modify the scheme and suppress this behavior. Instead of Eq. (1.3.10), we use

$$(1+ka)v_j^{n+1} = (1-ka)v_j^{n-1} + 2kD_0v_j^n. \tag{1.3.11}$$

Now, $z_{1,2}$ are the solutions of

$$(1+ka)z^2 = 1 - ka + 2i\lambda z \sin\xi,$$

and

$$z_{1,2} = \frac{i\lambda \sin\xi}{1+ka} \pm \sqrt{\frac{1-\lambda^2\sin^2\xi - k^2a^2}{(1+ka)^2}}.$$

Therefore,

$$|z_{1,2}| = \frac{(1-k^2a^2)^{1/2}}{1+ka} \simeq e^{-ka} \quad \text{for } \lambda^2 < 1 - k^2a^2,$$

and both $z_{1,2}^n$ decay like e^{-at_n}, that is, the solutions have the same decay rates as the solution of the differential equation.

We close this section by noting that the condition $k \le h$, found necessary for the explicit schemes (1.2.10) and (1.3.1), is very natural. Recall that the solution $u(x,t)$ of the problem (1.2.1) at any point $(\tilde{x},\tilde{t})$ is determined by the value of $f(x)$ at the point $\tilde{x}+\tilde{t}$ on the x-axis, because $u(x,t)$ is constant along the characteristic $x+t = \tilde{x}+\tilde{t}$ going through $(\tilde{x},\tilde{t})$ and $(\tilde{x}+\tilde{t},0)$. Now assume that $(\tilde{x},\tilde{t})$ is a gridpoint. Then, the solution of the difference approximation at $(\tilde{x},\tilde{t})$ depends on the initial data in the interval $\tilde{x}-\tilde{t}/\lambda \le x \le \tilde{x}+\tilde{t}/\lambda$ (see Figure 1.3.1).

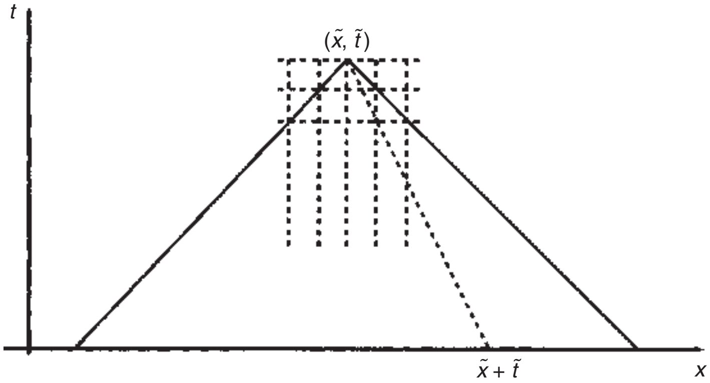


Figure 1.3.1. Domain of dependence for an explicit difference scheme.

If $\tilde{x}+\tilde{t}$ does not belong to this interval, that is, if $\lambda > 1$, then we cannot hope to obtain an accurate approximation. The condition that the domain of dependence of the difference approximation include the domain of dependence of the differential equation is known as the *Courant–Friedrichs–Lewy condition*, usually called the *CFL condition*.

In our case, the domain of dependence for the differential equation consists of one single point. This is not the case for a general hyperbolic differential equation, where the domain of dependence at a certain point $(\tilde{x}, \tilde{t})$ consists of a set of points or a whole interval.

EXERCISES

1.3.1. Prove that the solution of the leap-frog scheme converges to the solution of the differential equation, if $\lambda \leq 1-\delta, \delta > 0$.

1.3.2. Derive the explicit form of the leap-frog approximation (1.3.1) for $\lambda = 1$. Is the scheme suitable for computation?

1.3.3. Let $a = 10$. Estimate the time interval $[0, T]$, where the approximation (1.3.10) can be used. Does T depend on ω and/or k?

1.4. IMPLICIT METHODS

There is another way to stabilize the approximation (1.2.7). If we replace the forward difference in time by a backward difference, we get the *backward Euler method*

$$(I - kD_0)v_j^{n+1} = v_j^n, \quad j = 0, 1, \dots, N. \tag{1.4.1}$$

If we introduce the vector $\mathbf{v} = (v_0, \dots, v_N)^T$, then we can write Eq. (1.4.1) in matrix-vector form

$$A\mathbf{v}^{n+1} = \mathbf{v}^n,$$

$$A = \begin{bmatrix} 1 & -k/2h & 0 & \cdots & 0 & k/2h \\ k/2h & 1 & -k/2h & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & k/2h & 1 & -k/2h \\ -k/2h & 0 & \cdots & 0 & k/2h & 1 \end{bmatrix}. \tag{1.4.2}$$

This is called an *implicit* scheme because it couples the solution values of all points at the new time level. This means that the solution on the new time level depends on all values on the previous level. A linear system of $N+1$ equations must be solved to advance the scheme at each time step, and it, therefore, may seem to be an inefficient method. However, as we will see later, these schemes are often efficient and, in fact, the only realistic choice.

The now familiar way of introducing a Fourier component yields, for Eq. (1.4.1),

$$(1 - i\lambda \sin\xi)\hat{v}^{n+1}(\omega) = \hat{v}^n(\omega),$$

that is,

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \qquad \hat{Q} = \frac{1}{1 - i\lambda \sin\xi}.$$

Obviously, $|\hat{Q}| \leq 1$, and, again, there is damping of all frequencies except for $\omega = 0, \pi/h$. Note the important difference between Eq. (1.4.1) and the explicit scheme in Eq. (1.2.10). The stability condition $|\hat{Q}| \leq 1$ is satisfied for *all* values of λ for the implicit method. In other words, the scheme is stable for an arbitrary time step. Such schemes are called *unconditionally stable*. This is typical for implicit schemes.

This approximation is only first-order accurate because the time differencing is not centered. Instead, we can use the *trapezoidal rule* for time differencing and obtain the *Crank–Nicholson method*

$$\left(I - \frac{k}{2}\, D_0\right) v_j^{n+1} = \left(I + \frac{k}{2}\, D_0\right) v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.4.3}$$

The amplification factor is

$$\hat{Q} = \frac{2 + i\lambda \sin\xi}{2 - i\lambda \sin\xi}. \tag{1.4.4}$$

Thus, $|\hat{Q}| = 1$ for all values of λ, that is, the scheme is unconditionally stable and, as with the leap-frog scheme, there is no damping.

The explicit and implicit approximations can be combined into the so-called *θ scheme*

$$(I - \theta k D_0)v_j^{n+1} = (I + (1-\theta)kD_0)\, v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.4.5}$$

It is unconditionally stable for $\theta \geq 1/2$. The parameter θ is usually chosen to be in the interval $1/2 \leq \theta \leq 1$. The reason for introducing such an approximation is that the damping can be controlled by adjusting θ.

The system (1.4.2) is most efficiently solved by a direct method. Let the nonzero elements of the matrix A be denoted by a_{ij}, where $i = 0, 1, \ldots, N$ and $j = 0, 1, \ldots, N$. We make a factorization $A = LR$, where L and R have the form

$$L = \begin{bmatrix} 1 & & & & & & \\ \times & 1 & & & 0 & & \\ & \times & 1 & & & & \\ & & \ddots & \ddots & & & \\ & 0 & & \ddots & \ddots & & \\ & & & & \times & 1 & \\ \times & \times & \cdots & \cdots & \cdots & \times & 1 \end{bmatrix} \quad R = \begin{bmatrix} \times & \times & & & & & \times \\ & \times & \times & & 0 & & \times \\ & & \times & \times & & & \times \\ & & & \ddots & \ddots & & \vdots \\ & 0 & & & \ddots & \ddots & \vdots \\ & & & & & \times & \times \\ & & & & & & \times \end{bmatrix}.$$

The nonzero elements l_{ij} and r_{ij} of L and R, respectively, are given by the recursive formulas

$$\begin{aligned}
r_{00} &= a_{00}, \\
r_{01} &= a_{01}, \\
&\vdots \\
r_{0N} &= a_{0N}, \\
&\left.\begin{aligned} l_{j,j-1} &= \frac{a_{j,j-1}}{r_{j-1,j-1}} \\ r_{jj} &= a_{jj} - l_{j,j-1} r_{j-1,j} \end{aligned}\right\}, \quad j = 1, \ldots, N-1, \\
&\left.\begin{aligned} r_{j,j+1} &= a_{j,j+1} \\ &\vdots \\ r_{jN} &= -l_{j,j-1} r_{j-1,N} \end{aligned}\right\}, \quad j = 1, \ldots, N-2, \\
r_{N-1,N} &= a_{N-1,N} - l_{N-1,N-2} r_{N-2,N}, \\
&\left.\begin{aligned} l_{N0} &= \frac{a_{00}}{r_{00}} \\ l_{Nj} &= -\frac{l_{N,j-1} r_{j-1,j}}{r_{jj}} \end{aligned}\right\}, \quad j = 1, \ldots, N-2, \\
l_{N,N-1} &= \frac{1}{r_{N-1,N-1}} (a_{N,N-1} - l_{N,N-2} r_{N-2,N-1}), \\
r_{NN} &= a_{NN} - \sum_{j=0}^{N-1} l_{Nj} r_{jN}.
\end{aligned}$$

The system (1.4.2) is rewritten as

$$LR\mathbf{v}^{n+1} = \mathbf{v}^n. \tag{1.4.6}$$

The solution is obtained by backward and forward substitution

$$\begin{aligned} L\mathbf{w} &= \mathbf{v}^n, \\ R\mathbf{v}^{n+1} &= \mathbf{w}. \end{aligned} \tag{1.4.7}$$

The number of arithmetic operations for the whole procedure is proportional to N. Hence, for problems in one space dimension, the work required for the implicit method is of the same order as that for an explicit method. Note, however, that on parallel computers the simpler algorithmic structure of an explicit scheme may be an advantage.

The nonzero corner elements a_{0N} and a_{N0} in the matrix A are an effect of the periodicity conditions. For other types of boundary conditions, where A is tridiagonal without the corner elements, the formulas for computing the elements of L and R still hold, and we get

$$\begin{aligned} r_{iN} &= 0, \quad i = 0, 1, \dots, N-2, \\ l_{Nj} &= 0, \quad j = 0, 1, \dots, N-2. \end{aligned}$$

For methods with more than three points on time level t_{n+1} coupled to each other, the bandwidth ν becomes larger. The same type of solution procedure can still be applied. The matrices L and R have the same number of nonzero subdiagonals and superdiagonals, respectively, as A has, and it can be shown that $\mathscr{O}(\nu^2 N)$ arithmetic operations are required for the solution.

For problems in two space dimensions on an $N \times N$ grid, the bandwidth is $\nu = \mathscr{O}(N)$, and a direct generalization of the above-mentioned method leads to an operation count of the order of N^4. In this case, iterative methods can be considerably more efficient, and they are the only realistic methods in three space dimensions.

EXERCISES

1.4.1. Prove that Eq. (1.4.5) is unconditionally stable for $\theta \geq \frac{1}{2}$.

1.4.2. Calculate the exact number of arithmetic operations required to advance by one step the implicit scheme (1.4.3). Compare it with the work required to advance by one step the explicit scheme (1.2.10).

1.4.3. Derive the direct solution algorithm for a system $A\mathbf{v} = \mathbf{b}$, where A has ν nonzero diagonals. Prove that the operation count is $\mathscr{O}(\nu^2 N)$.

1.5. TRUNCATION ERROR

In the previous sections, we have derived several difference schemes to calculate the solution u of Eq. (1.2.1). In every case, we could write their solutions v in closed form and, therefore, we could calculate the error $u - v$ explicitly. In this section, we discuss the truncation error, which is a measure of the accuracy of a given scheme. Instead of estimating the error $u - v$, we calculate how well u satisfies the difference approximation. We can then use the truncation error to estimate $u - v$. The advantage of this procedure is that it can be used when u and v are not known explicitly. It can also be used for equations with variable coefficients.

Let u be a smooth function. Using a Taylor series expansion around any point (x, t), we obtain

$$\begin{aligned} D_0 u(x,t) &= \frac{u(x+h,t) - u(x-h,t)}{2h} \\ &= u_x(x,t) + \frac{h^2}{3!}\, u_{xxx}(x,t) + \frac{h^4}{5!}\, \varphi_0(x,t), \qquad (1.5.1) \\ |\varphi_0(x,t)| &\le \max_{x-h\le\xi\le x+h} \left| \frac{\partial^5 u(\xi,t)}{\partial x^5} \right|, \\ D_+D_- u(x,t) &= \frac{u(x+h,t) - 2u(x,t) + u(x-h,t)}{h^2} \\ &= u_{xx}(x,t) + \frac{2h^2}{4!}\, u_{xxxx}(x,t) + \frac{2h^4}{6!}\, \varphi_1(x,t), \qquad (1.5.2) \\ |\varphi_1(x,t)| &\le \max_{x-h\le\xi\le x+h} \left| \frac{\partial^6 u(\xi,t)}{\partial x^6} \right|, \\ \frac{u(x,t+k) - u(x,t)}{k} &= u_t(x,t) + \frac{k}{2}\, u_{tt}(x,t) + \frac{k^2}{3!}\, \psi_0(x,t), \\ |\psi_0(x,t)| &\le \max_{t\le\xi\le t+k} \left| \frac{\partial^3 u(x,\xi)}{\partial t^3} \right|, \qquad (1.5.3) \\ \frac{u(x,t+k) - u(x,t-k)}{2k} &= u_t(x,t) + \frac{k^2}{3!}\, u_{ttt}(x,t) + \frac{k^4}{5!}\, \psi_1(x,t), \\ |\psi_1(x,t)| &\le \max_{t-k\le\xi\le t+k} \left| \frac{\partial^5 u(x,\xi)}{\partial t^5} \right|, \qquad (1.5.4) \\ \frac{u(x,t+k) - 2u(x,t) + u(x,t-k)}{k^2} & \\ &= u_{tt}(x,t) + \frac{2k^2}{4!}\, u_{tttt}(x,t) + \frac{2k^4}{6!}\, \psi_2(x,t), \qquad (1.5.5) \\ |\psi_2(x,t)| &\le \max_{t-k\le\xi\le t+k} \left| \frac{\partial^6 u(x,\xi)}{\partial t^6} \right|. \end{aligned}$$

Now assume that u is a smooth solution of the problem (1.2.1) and substitute it into the difference scheme (1.2.10). Then, we obtain from Eq. (1.5.1) to Eq. (1.5.4) and $u_t = u_x, \; u_{tt} = u_{xt} = u_{xx}$,

$$\begin{aligned}\frac{u_j^{n+1} - u_j^n}{k} - D_0 u_j^n - \sigma h D_+ D_- u_j^n &= u_t(x_j, t_n) - u_x(x_j, t_n) + \frac{k}{2}\, u_{tt}(x_j, t_n) \\ &\quad - \sigma h u_{xx}(x_j, t_n) + \mathscr{O}(h^2 + k^2) \\ &= \left(\frac{k}{2} - \sigma h\right) u_{xx}(x_j, t_n) + \mathscr{O}(h^2 + k^2) =: \tau_j^n.\end{aligned} \tag{1.5.6}$$

We call τ_j^n the *truncation error* and say that the method is accurate of order (p, q) if $\tau = \mathscr{O}(h^p + k^q)$. For $\sigma \neq k/(2h)$, the above-mentioned method is accurate of order (1, 1). For $\sigma = k/(2h)$, the Lax–Wendroff method, the order of accuracy is (2, 2).

Equation (1.5.6) implies that u satisfies

$$\begin{aligned} u_j^{n+1} &= (I + kD_0)u_j^n + \sigma k h D_+ D_- u_j^n + k\tau_j^n, \\ u_j^0 &= f_j. \end{aligned} \tag{1.5.7}$$

Subtracting Eq. (1.2.10) from Eq. (1.5.7), we obtain, for the error $e = u - v$,

$$\begin{aligned} e_j^{n+1} &= (I + kD_0)e_j^n + \sigma k h D_+ D_- e_j^n + k\tau_j^n, \\ e_j^0 &= 0. \end{aligned}$$

We will later show that e is of the same order as τ if the approximation is stable.

One can also easily derive expressions for τ for the other methods. The leap-frog and the Crank–Nicholson methods are accurate of order (2, 2), whereas the backward Euler method is accurate of order (2, 1). Thus, we expect that the error in time will dominate when using the backward Euler method unless the solution varies much slower in time than in space (in the truncation error the time step k is multiplied by time derivatives).

EXERCISES

1.5.1. When deriving the order of accuracy, Taylor expansion around some point (x_*, t_*) is used. Prove that (x_*, t_*) can be chosen arbitrarily and, in particular, that it does not have to be a gridpoint.

1.5.2. Prove that the leap-frog scheme (1.3.1) and the Crank–Nicholson scheme (1.4.3) are accurate of order (2, 2). Despite the same order of accuracy, one can expect that one scheme is more accurate than the other. Why is that so?

1.6. HEAT EQUATION

In this section, we consider the simplest *parabolic* model problem for heat conduction,

$$
\begin{aligned}
u_t &= u_{xx}, \qquad -\infty < x < \infty,\ 0 \leq t \\
u(x,0) &= f(x), \qquad -\infty < x < \infty,
\end{aligned}
\tag{1.6.1}
$$

with 2π-periodic initial data. We again use the Fourier technique to obtain the solution. The differential operator $\partial^2/\partial x^2$ corresponds to the multiplication operator $-\omega^2$ in Fourier space, and we obtain

$$
\begin{aligned}
\frac{\partial \hat{u}(\omega,t)}{\partial t} &= -\omega^2 \hat{u}(\omega,t), \\
\hat{u}(\omega,0) &= \hat{f}(\omega).
\end{aligned}
\tag{1.6.2}
$$

The solution of the problem (1.6.2) is

$$
\hat{u}(\omega,t) = e^{-\omega^2 t}\hat{f}(\omega), \tag{1.6.3}
$$

which yields

$$
u(x,t) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{-\omega^2 t} e^{i\omega x}\hat{f}(\omega), \qquad f(x) = \frac{1}{\sqrt{2\pi}} \sum_{\omega=-\infty}^{\infty} e^{i\omega x}\hat{f}(\omega). \tag{1.6.4}
$$

From Parseval's relation (A.1.9), we obtain

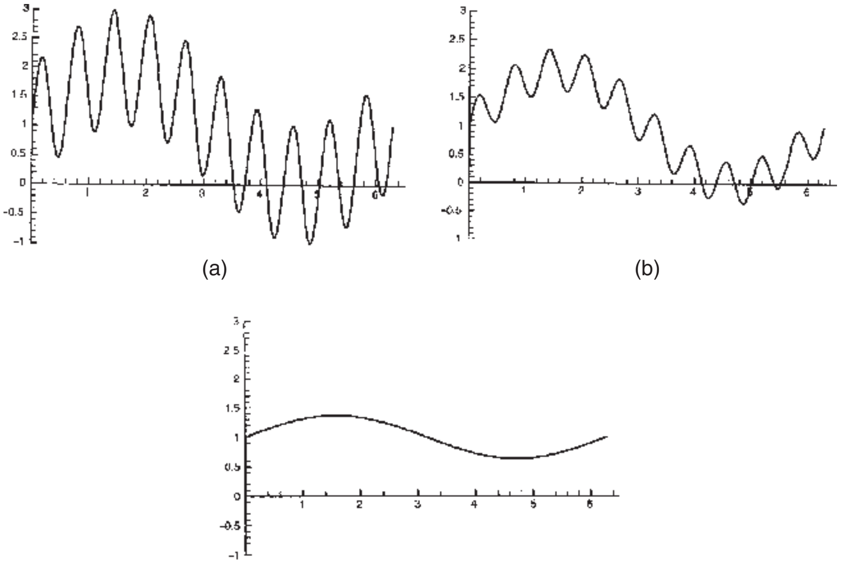
$$
\|u(\cdot,t)\|^2 = \sum_{\omega=-\infty}^{\infty} |e^{-\omega^2 t}\hat{f}(\omega)|^2 \leq \|f(\cdot)\|^2. \tag{1.6.5}
$$

Equation (1.6.3) illustrates typical parabolic behavior; each Fourier component is damped with time, and the damping is very strong for high frequencies. Even if the initial data are very rough, the solution is an analytic function for $t > 0$, that is, the Fourier coefficients decay exponentially. In Figure 1.6.1, we have plotted the solution of the problem (1.6.1) with initial data $f(x) = 1 + \sin x + \sin(10x)$ for $t = 0,\ 0.01,\ 1$.

One can also show that, unlike the hyperbolic case, the speed of propagation is infinite. We now consider simple difference approximations of Eq. (1.6.1) and begin with

$$
v_j^{n+1} = (I + kD_+D_-)v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.6.6}
$$

The scheme is based on forward differencing in time and is often called the *Euler method*. We recall that the corresponding approximation (1.2.7) for $u_t = u_x$ was useless because it was unstable for any sequence $k, h \to 0$ with $k/h \geq c > 0$.



(a)

(b)

(c)

Figure 1.6.1. Solution of the problem (1.6.1). (a) $t = 0$, (b) $t = 0.01$, and (c) $t = 1$.

To compute the symbol $\hat{Q}$, we use the basic trigonometric formulas of Section 1.1. From Eq. (1.1.5),

$$kD_+D_-e^{i\omega x_j} = -4\sigma \sin^2 \frac{\xi}{2}\, e^{i\omega x_j}, \tag{1.6.7}$$

where $\sigma = k/h^2 and\ \xi = \omega h.$

The transformed difference scheme is then

$$\hat{v}^{n+1}(\omega) = \hat{Q}\hat{v}^n(\omega), \qquad \hat{Q} = 1 - 4\sigma \sin^2 \frac{\xi}{2}. \tag{1.6.8}$$

The condition $|\hat{Q}| \leq 1$ is equivalent to

$$\sigma \leq \frac{1}{2}. \tag{1.6.9}$$

We have calculated approximations of the solution shown in Figure 1.6.1 using Eq. (1.6.6) with $\sigma = 1/2,\ N = 100$. In Figure 1.6.2, we have plotted the error $u - v$, for $t = 0.01,\ 1$.

The condition given in Eq. (1.6.9) implies that the time step k must be chosen proportional to h^2. This is often too restrictive. On the other hand, it is natural for an explicit scheme. As noted earlier, there is no finite speed of propagation for parabolic problems. This means that the domain of dependence of the difference

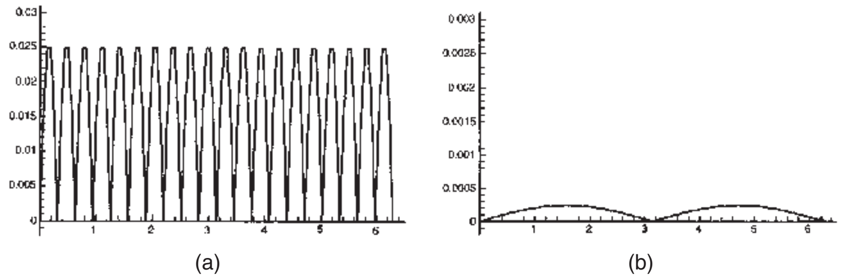


Figure 1.6.2. Error when solving the problem (1.6.1) by the method (1.6.6). (a) $t = 0.01$ and (b) $t = 1$.

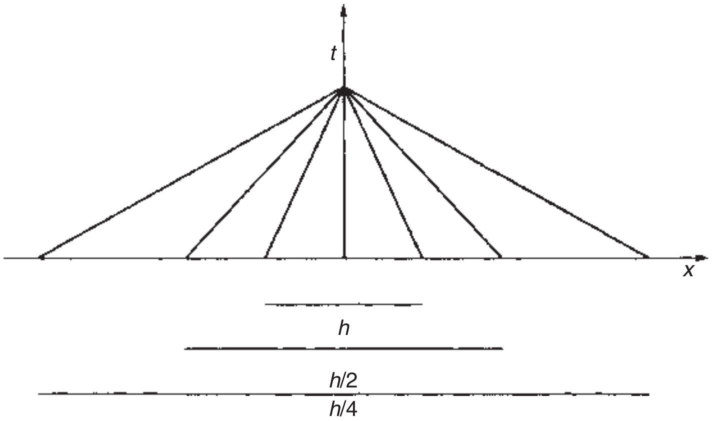


Figure 1.6.3. Domain of dependence for decreasing h and $k = \sigma h^2$.

scheme must cover the whole interval in the limit $k \to 0, h \to 0$, even for points $(\tilde{x}, \tilde{t})$ arbitrarily close to the x-axis; otherwise, the approximation cannot converge to the true solution. Figure 1.6.3 shows the expanding domain of dependence for fixed t, decreasing h and $k = \sigma h^2$.

The leap-frog scheme approximating Eq. (1.6.1) is

$$v_j^{n+1} = 2kD_+D_-v_j^n + v_j^{n-1}, \quad j = 0, 1, \dots, N. \tag{1.6.10}$$

The solution in Fourier space is of the form shown in Eq. (1.3.7), where z_1 and z_2 are the roots of the characteristic equation

$$z^2 + 8\sigma \left(\sin^2 \frac{\xi}{2}\right) z - 1 = 0, \tag{1.6.11}$$

that is,

$$z_{1,2} = -4\sigma \sin^2 \frac{\xi}{2} \pm \sqrt{1 + \left(4\sigma \sin^2 \frac{\xi}{2}\right)^2}.$$

For $\xi \neq 0$, one of $z_{1,2}$ is larger than 1 in magnitude for all values of $\sigma > 0$. The scheme is useless because it is unstable for any sequence $k, h \to 0$ with $k/h^2 \geq c > 0$.

A small modification can be made to stabilize the scheme. Equation (1.6.10) can be written as

$$v_j^{n+1} = 2\sigma(v_{j+1}^n - 2v_j^n + v_{j-1}^n) + v_j^{n-1},$$

and if we replace v_j^n by $(v_j^{n+1} + v_j^{n-1})/2$, then we obtain

$$v_j^{n+1} = 2\sigma(v_{j+1}^n - v_j^{n+1} - v_j^{n-1} + v_{j-1}^n) + v_j^{n-1}, \quad j = 0, 1, \ldots, N. \quad (1.6.12)$$

This is known as the *DuFort–Frankel method*. It is still explicit because we can solve for v_j^{n+1} and write it as

$$v_j^{n+1} = \frac{1}{1+2\sigma}\left(2\sigma(v_{j+1}^n + v_{j-1}^n) + (1-2\sigma)v_j^{n-1}\right).$$

Now the characteristic equation is

$$z^2 - \frac{4\sigma}{1+2\sigma}(\cos\xi)z - \frac{1-2\sigma}{1+2\sigma} = 0,$$

that is,

$$z_{1,2} = \frac{2\sigma}{1+2\sigma}\cos\xi \pm \frac{1}{1+2\sigma}\sqrt{A}, \quad (1.6.13)$$

where $A = 4\sigma^2\cos^2\xi + 1 - 4\sigma^2$. If $A \geq 0$, then $A \leq 1$, and

$$|z_{1,2}| \leq \frac{2\sigma}{1+2\sigma} + \frac{1}{1+2\sigma} = 1.$$

If $A < 0$, then we write

$$z_{1,2} = \frac{1}{1+2\sigma}\left(2\sigma\cos\xi \pm i\sqrt{4\sigma^2(1-\cos^2\xi) - 1}\right) \quad (1.6.14)$$

and get

$$|z_{1,2}|^2 = \frac{4\sigma^2 - 1}{(1+2\sigma)^2} = \frac{2\sigma - 1}{2\sigma + 1} < 1.$$

We shall come back to general stability conditions for multistep methods in Chapter 4. For the DuFort–Frankel approximation, it is easily seen that one of the roots z_j is always strictly inside the unit circle, and it turns out that this leads to stability as long as $\sigma > 0$ is a constant (see Corem and Ditkowski (2012) for the case where σ is not a constant). This is somewhat surprising because the scheme is explicit. The time step can be chosen independent of the space step.

This seems to contradict the conclusion that the domain of dependence must expand as h decreases. However, this apparently contradictory behavior is an illustration of the fact that stability is only a necessary condition for convergence. It does not guarantee that solutions are accurate approximations. It only guarantees that solutions remain bounded.

To investigate the order of accuracy, we calculate the truncation error. From Eq. (1.5.1) to Eq. (1.5.6),

$$\begin{aligned}\tau &= \frac{u_j^{n+1}-u_j^{n-1}}{2k} - D_+D_-u_j^n + \frac{k^2}{h^2}\frac{u_j^{n+1}-2u_j^n+u_j^{n-1}}{k^2}\\ &= u_t - u_{xx} + \frac{k^2}{h^2}\,u_{tt} + \mathscr{O}\left(k^2+h^2+\frac{k^4}{h^2}\right) = \frac{k^2}{h^2}\,u_{tt} + \mathscr{O}\left(k^2+h^2+\frac{k^4}{h^2}\right).\end{aligned} \tag{1.6.15}$$

Thus, $\lim_{k,h\to 0}\tau = 0$ only if $\lim_{k,h\to 0} k/h = 0$. Typically, one chooses

$$k = ch^{1+\delta}, \quad \delta > 0. \tag{1.6.16}$$

Then, the truncation error is $\mathscr{O}(h^{2\delta})$, and the method is only accurate of order (2,2) if $\delta = 1$, that is, $k = O(h^2)$, which is essentially the same restriction as that required for the Euler method.

We now examine analogs of the implicit schemes introduced in Section 1.5. The *backward Euler* approximation is

$$(I - kD_+D_-)v_j^{n+1} = v_j^n, \quad j = 0, 1, \ldots, N, \tag{1.6.17}$$

with the amplification factor

$$\hat{Q} = \frac{1}{1+4\sigma\sin^2\frac{\xi}{2}}, \qquad \sigma = \frac{k}{h^2}. \tag{1.6.18}$$

The magnitude of $\hat{Q}$ is never greater than 1 independent of σ, and all nonzero frequencies are damped. Note that, as for the differential equation, the damping is stronger for larger ω.

In Figure 1.6.4, we show the error of backward Euler calculations with $k = h$ and $N = 100$ for $t = 0.01,\ 1$, respectively. The initial data are the same as in Figure 1.6.1.

The Crank–Nicholson scheme

$$\left(I - \frac{k}{2}\,D_+D_-\right)v_j^{n+1} = \left(I + \frac{k}{2}\,D_+D_-\right)v_j^n, \quad j = 0, 1, \ldots, N \tag{1.6.19}$$

has the amplification factor

$$\hat{Q} = \frac{1-2\sigma\sin^2\frac{\xi}{2}}{1+2\sigma\sin^2\frac{\xi}{2}}, \qquad \sigma = \frac{k}{h^2}, \tag{1.6.20}$$

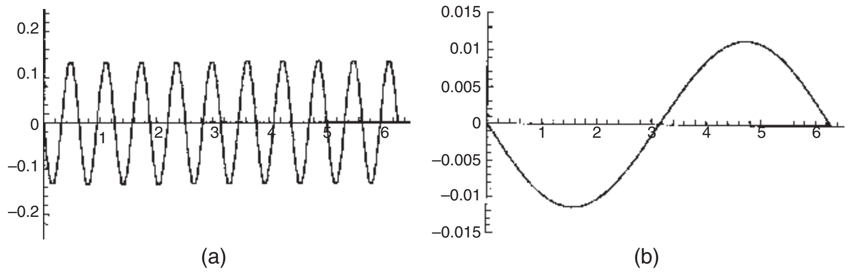


Figure 1.6.4. Error when solving the problem (1.6.1) by the backward Euler method. (a) $t = 0.01$ and (b) $t = 1$.

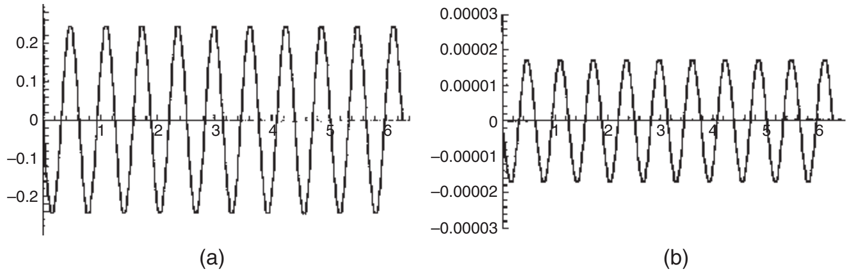


Figure 1.6.5. Error when solving the problem (1.6.1) by the Crank–Nicholson method. (a) $t = 0.01$ and (b) $t = 1$.

and, like the backward Euler method, it is unconditionally stable. However, when σ is large, $\hat{Q}$ is near -1 for $\xi \neq 0$, and there is very little damping. This is a serious drawback because one would like to use time steps of the same order as the space step. With that choice, we get $\sigma = \mathscr{O}(1/h)$ and $\hat{Q} \to -1$ as $h \to 0$ for every fixed ξ (i.e., as $\omega \to \infty$).

We have calculated the approximate solution of the problem (1.6.1) with the same initial data as before using the Crank–Nicholson method with $k = h$ and $N = 100$ for $t = 0.01,\ 1$. The error is plotted in Figure 1.6.5. There is now an oscillating error (see Exercise 1.6.3).

We can also combine the two implicit schemes for parabolic equations obtaining the θ scheme

$$\begin{aligned}(I - \theta k D_+ D_-) v_j^{n+1} &= (I + (1-\theta) k D_+ D_-)\, v_j^n, \\ &\qquad j = 0, 1, \ldots, N, \quad 0 \leq \theta \leq 1,\end{aligned} \tag{1.6.21}$$

which is unconditionally stable for $\theta \geq \frac{1}{2}$ (see Exercise 1.6.2). As in the hyperbolic case, the damping increases with θ up to $\theta = 1$ (backward Euler) but the accuracy decreases.

EXERCISES

1.6.1. Assume that the initial data for the problem (1.6.1) is a simple wave $f(x) = e^{i\omega x}$. Determine the time t_1, where $\|u(\cdot, t_1)\| = 10^{-6}$. Apply the Euler method (1.6.6), and calculate the corresponding time t_2. Determine the optimal time step for a given h.

1.6.2. Prove that the θ scheme (1.6.21) is unconditionally stable for $\theta \geq \frac{1}{2}$.

1.6.3. Derive the truncation error for the backward Euler and the Crank–Nicholson methods applied to $u_t = u_{xx}$. Prove that it is $\mathcal{O}(h^2 + k)$ and $\mathcal{O}(h^2 + k^2)$, respectively. Despite this fact, the backward Euler method is more accurate at certain times for the example computed in this section. Explain this paradox.

1.7. CONVECTION–DIFFUSION EQUATION

In many applications, the differential equations have both first- and second-order derivatives in space. We now consider the model problem for convection–diffusion

$$\begin{aligned} u_t + a u_x &= \eta u_{xx}, \qquad -\infty < x < \infty\,,\ 0 \leq t, \quad \eta = \text{const} > 0, \\ u(x, 0) &= f(x), \qquad -\infty < x < \infty, \end{aligned} \tag{1.7.1}$$

with 2π-periodic initial data. In Fourier space, the corresponding problem is

$$\begin{aligned} \frac{\partial \hat{u}(\omega, t)}{\partial t} + i a \omega \hat{u}(\omega, t) &= -\eta \omega^2 \hat{u}(\omega, t), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned} \tag{1.7.2}$$

with solutions

$$\hat{u}(\omega, t) = e^{-(i a \omega + \eta \omega^2) t} \hat{f}(\omega). \tag{1.7.3}$$

Consider the difference approximation

$$v_j^{n+1} = v_j^n + k(\eta D_+ D_- - a D_0) v_j^n, \quad j = 0, 1, \ldots, N. \tag{1.7.4}$$

The amplification factor is

$$\hat{Q} = 1 - 2\alpha \sin^2 \frac{\xi}{2} - i\lambda \sin \xi, \qquad \alpha = \frac{2\eta k}{h^2}, \quad \lambda = \frac{ak}{h}. \tag{1.7.5}$$

The "parabolic" stability condition for the case $a = 0$ is

$$\frac{k\eta}{h^2} \leq \frac{1}{2}, \quad \text{that is,}\ \alpha \leq 1. \tag{1.7.6}$$

For $a \neq 0$, we have

$$\begin{aligned} |\hat{Q}|^2 &= 1 - 4\alpha \sin^2 \frac{\xi}{2} + 4\alpha^2 \sin^4 \frac{\xi}{2} + 4\lambda^2 \sin^2 \frac{\xi}{2} \left(1 - \sin^2 \frac{\xi}{2}\right) \\ &= 1 - 4(\lambda^2 - \alpha^2)s^2 + 4(\lambda^2 - \alpha)s, \end{aligned} \tag{1.7.7}$$

where $s = \sin^2(\xi/2)$. Thus, $|\hat{Q}| \leq 1$ for all ξ if, and only if,

$$\phi(s) := -(\lambda^2 - \alpha^2)s + \lambda^2 - \alpha \leq 0, \qquad 0 \leq s \leq 1. \tag{1.7.8}$$

$\phi(s)$ is a linear function of s and, therefore, Eq. (1.7.8) holds if and only if

$$\phi(0) = \lambda^2 - \alpha \leq 0, \qquad \phi(1) = \alpha^2 - \alpha \leq 0,$$

that is,

$$\lambda^2 \leq \alpha \leq 1 \quad \text{or} \quad a^2 k \leq 2\eta \leq h^2/k. \tag{1.7.9}$$

The conditions in Eq. (1.7.9) can be interpreted in this way: The parabolic term makes it possible to stabilize the approximation of the hyperbolic part. However, the coefficient η must be large enough compared to a (or k small enough) in order to provide enough damping. Furthermore, the damping of the method is always less than that of the differential equation. The true parabolic decay rate for Eq. (1.7.1) is not preserved by the approximation. Part of it is required to stabilize the hyperbolic part. As η becomes small, this becomes more severe.

In Section 1.6, it was noted that, for parabolic problems, the stability restriction on the time step for explicit schemes (except the DuFort–Frankel method) is often too severe, and implicit approximations should be used. Implicit methods can also be used when first-order derivatives are present. For example, the Crank–Nicholson method

$$\begin{aligned} &\left(I + \frac{k}{2}\,(aD_0 - \eta D_+ D_-)\right) v_j^{n+1} \\ &\quad = \left(I - \frac{k}{2}\,(aD_0 - \eta D_+ D_-)\right) v_j^n, \quad j = 0, 1, \ldots, N \end{aligned} \tag{1.7.10}$$

is unconditionally stable. In applications, however, the hyperbolic part is often nonlinear, that is, $a = a(u)$, and a nonlinear system of equations must be solved at each step. In this case, it is convenient to use the so-called *semi-implicit* method. The simplest approximation of this kind for our problem is

$$(I - k\eta D_+ D_-)v_j^{n+1} = -2kaD_0 v_j^n + (I + k\eta D_+ D_-)v_j^{n-1}, \quad j = 0, 1, \ldots, N, \tag{1.7.11}$$

which is a combination of the leap-frog approximation and the Crank–Nicholson approximation. v_j^1 must be computed by some other one-step method. In Fourier space, Eq. (1.7.11) yields

$$\left(1 + 4\sigma \sin^2 \frac{\xi}{2}\right) \hat{v}^{n+1}(\omega) = -i2\lambda \ (\sin\xi)\hat{v}^n(\omega) + \left(1 - 4\sigma \sin^2 \frac{\xi}{2}\right) \hat{v}^{n-1}(\omega),$$

$$\sigma = \frac{k\eta}{h^2}, \qquad \lambda = \frac{ak}{h}. \tag{1.7.12}$$

The corresponding characteristic equation is

$$z^2 + \frac{i2\lambda \sin\xi}{1+\beta}\, z - \frac{1-\beta}{1+\beta} = 0, \qquad \beta = 4\sigma \sin^2 \frac{\xi}{2}, \tag{1.7.13}$$

with solutions

$$z_{1,2} = \frac{-i\lambda \sin\xi \pm \sqrt{1 - \beta^2 - \lambda^2 \sin^2 \xi}}{1+\beta}. \tag{1.7.14}$$

First, assume that the square root is real, that is,

$$\beta^2 + \lambda^2 \sin^2 \xi \le 1. \tag{1.7.15}$$

Then,

$$|z_{1,2}|^2 = \frac{1-\beta^2}{(1+\beta)^2} = \frac{1-\beta}{1+\beta} \le 1.$$

Next, assume that

$$\beta^2 + \lambda^2 \sin^2 \xi > 1. \tag{1.7.16}$$

Then, the roots are purely imaginary, and we have, for $|\lambda| < 1$,

$$|z_{1,2}| = \left| \frac{\lambda \sin\xi \pm \sqrt{\beta^2 + \lambda^2 \sin^2 \xi - 1}}{1+\beta} \right| \le \frac{|\lambda| + \beta}{1+\beta} < 1. \tag{1.7.17}$$

Thus, $|z_{1,2}| \le 1$ for $|\lambda| \le 1$, which is the same stability condition we obtained for the leap-frog approximation (1.3.1) of $u_t = u_x$. Note, however, that $z_1 = z_2$ for $\beta^2 + \lambda^2 \sin^2 \xi = 1$. Then, the representation $\hat{v}^n(\omega) = \sigma_1 z_1^n + \sigma_2 z_2^n$ becomes $\hat{v}^n(\omega) = (\sigma_1 + \sigma_2 n) z_1^n$. Because $|z_1| \le |\lambda| < 1$ in this case, we have

$$n|z_1|^n \le \text{const}$$

independent of n. Thus, $|\hat{v}^n(\omega)|$ is bounded independent of ω, n, and it is stable. We shall consider the approximation in Eq. (1.7.11) in a more general setting in Chapter 4.

The time step can be chosen to be of the same order as the space step with the semi-implicit scheme (1.7.11), which is a substantial gain in efficiency compared to an explicit scheme. This was achieved without involving the whole difference operator at the new time level.

EXERCISES

1.7.1. Write a program that computes the solutions to Eq. (1.7.4) for $N = 10, 20, 40, \ldots$. Choose the time step such that

(**a**) α, defined in Eq. (1.7.5), is a constant with $\alpha \le 1$,

(**b**) λ, defined in Eq. (1.7.5), is a constant with $|\lambda| \le 1$.

Compare the solutions and explain the difference in their behavior.

1.7.2. Newton's method for a nonlinear system $\mathbf{F}(\mathbf{v}) = 0$ is defined by

$$\mathbf{v}^{(n+1)} = \mathbf{v}^{(n)} - \mathbf{F}'(\mathbf{v}^{(n)})^{-1}\mathbf{F}(\mathbf{v}^{(n)}), \qquad n = 0, 1, \ldots,$$

where $\mathbf{F}'$ is the Jacobian matrix of $\mathbf{F}$ with respect to $\mathbf{v}$. Assume that the coefficients a and η in Eq. (1.7.1) depend on u. Prove that Newton's method applied to each step of the Crank–Nicholson scheme (1.7.10) leads to linear systems of the same structure as discussed in Section 1.4.

1.8. HIGHER ORDER EQUATIONS

In this section, we briefly discuss differential equations of the form

$$\begin{aligned} \frac{\partial u}{\partial t} &= a\frac{\partial^p u}{\partial x^p}, \qquad -\infty < x < \infty,\ t \ge 0, \\ u(x, 0) &= f(x), \qquad -\infty < x < \infty, \end{aligned} \tag{1.8.1}$$

where a is a complex number and $p \ge 1$. In Fourier space, Eq. (1.8.1) becomes

$$\begin{aligned} \frac{\partial \hat{u}(\omega, t)}{\partial t} &= a(i\omega)^p \hat{u}(\omega, t), \\ \hat{u}(\omega, 0) &= \hat{f}(\omega), \end{aligned} \tag{1.8.2}$$

that is,

$$\hat{u}(\omega, t) = e^{a(i\omega)^p t} \hat{f}(\omega),$$

with

$$|\hat{u}(\omega, t)| = |e^{\text{Re}\,[a(i\omega)^p]t} \hat{f}(\omega)|. \tag{1.8.3}$$

For the problem to be well-posed, it is sufficient that the condition

$$\text{Re}\,[a(i\omega)^p] \le 0 \tag{1.8.4}$$

be fulfilled for all ω. This ensures that the solution will satisfy the estimate

$$\|u(\cdot, t)\| \le \|f(\cdot)\|. \tag{1.8.5}$$

Because ω is real and can be positive or negative, we obtain the condition

$$\begin{aligned} \operatorname{sign}(\operatorname{Re} a) &= (-1)^{p/2+1}, \quad && \text{if } \operatorname{Re} a \neq 0 \text{ and } p \text{ is even,} \\ \operatorname{Im} a &= 0, && \text{if } p \text{ is odd.} \end{aligned} \tag{1.8.6}$$

A special case is

$$\frac{\partial u}{\partial t} = i\alpha \frac{\partial^2 u}{\partial x^2}, \tag{1.8.7}$$

where α is real. This is the principal part of the *Schrödinger equation* governing the fundamentals of quantum mechanics. In Fourier space, we have

$$\hat{u}(\omega, t) = e^{-i\alpha\omega^2 t} \hat{f}(\omega),$$

leading to norm conservation

$$\|u(\cdot, t)\| = \|f(\cdot)\|.$$

The most natural centered difference approximation for the general equation of order p is given by

$$\frac{\partial^p}{\partial x^p} \to Q_p = \begin{cases} (D_+D_-)^{p/2}, & p \text{ even,} \\ D_0(D_+D_-)^{(p-1)/2}, & p \text{ odd,} \end{cases} \tag{1.8.8}$$

which, in Fourier space, yields

$$(i\omega)^p \to \hat{Q}_p = \begin{cases} \left(-\dfrac{4}{h^2}\sin^2\dfrac{\omega h}{2}\right)^{p/2}, & p \text{ even,} \\ \frac{i}{h}\sin(\omega h)\left(-\dfrac{4}{h^2}\sin^2\dfrac{\omega h}{2}\right)^{(p-1)/2}, & p \text{ odd.} \end{cases} \tag{1.8.9}$$

If a is real, the Euler method can always be used if p is even. The leap-frog scheme can always be used if p is odd. This follows directly from the calculations made in Sections 1.3 and 1.6. For the Euler method, we have

$$\hat{Q} = 1 + ka\hat{Q}_p, \tag{1.8.10}$$

where $\hat{Q}_p$ is defined in Eq. (1.8.9). If a is real, stability requires that the condition

$$(-1)^{p/2-1} \cdot \frac{4^{p/2}ak}{h^p} \leq 2, \qquad p \text{ even,} \tag{1.8.11}$$

be satisfied. For $p \geq 4$, the time step restriction is so severe that the method cannot be used in any realistic computation. Similarly, for the leap-frog scheme, we obtain a condition of the form

$$\frac{k}{h^p} \leq \text{const}, \quad p \text{ odd}. \tag{1.8.12}$$

[This is easily seen if we follow the calculations leading to Eq. (1.3.9).] $p = 1$ is the practical limit in several space dimensions, and we conclude that implicit methods are necessary for higher order equations. (In one space dimension, one could possibly use explicit methods for $p = 2, 3$.)

EXERCISES

1.8.1. What explicit method could be used for the Schrödinger type equation

$$u_t = iu_{xx}? \tag{1.8.13}$$

Derive the stability condition.

1.8.2. Define the Crank–Nicholson approximation for the Korteweg de Vries type equation

$$u_t = u_{xxx} + au_x. \tag{1.8.14}$$

Prove unconditional stability.

1.8.3. Define a semi-implicit approximation suitable for the efficient solution of Eq. (1.8.14) with $a = a(u)$. Derive the stability condition (for $a = \text{const}$).

1.9. SECOND-ORDER WAVE EQUATION

In Section 1.2, we discussed the simplest possible partial differential equation describing a wave propagating in one direction. In this section, we consider the second order wave equation describing the real case, where waves are propagated in two directions:

$$u_{tt} = c^2 u_{xx}, \qquad -\infty < x < \infty, \ t \geq 0.$$

Here, c is the wave propagation speed. Because we have a second derivative in time, two initial conditions are needed, and we prescribe

$$\begin{aligned} u(x, 0) &= f(x), \\ u_t(x, 0) &= g(x). \end{aligned}$$

In Fourier space, the wave equation takes the form

$$\hat{u}_{tt} = -c^2\omega^2\hat{u}, \tag{1.9.1}$$

which has the solution

$$\hat{u}(\omega, t) = \alpha e^{ic\omega t} + \beta e^{-ic\omega t}.$$

There are two constants to be determined by the initial data $\hat{f}(\omega)$ and $\hat{g}(\omega)$.

The standard difference scheme is

$$v_j^{n+1} - 2v_j^n + v_j^{n-1} = k^2c^2D_+D_-v_j^n, \qquad n = 1, 2, \ldots,$$

which requires data at two time levels t_0 and t_1 to get started. One possibility is to use

$$\begin{aligned} v_j^0 &= f_j, \\ v_j^1 &= f_j + kg_j. \end{aligned}$$

The second condition is a low order approximation of the second initial condition for the differential equation. We shall further discuss the accuracy of this type of approximations in Chapter 4.

The Fourier transform of the difference scheme is

$$\hat{v}^{n+1}(\omega) - 2\hat{v}^n(\omega) + \hat{v}^{n-1}(\omega) = -4\lambda^2 \sin^2\frac{\xi}{2}\,\hat{v}^n(\omega), \tag{1.9.2}$$

where $\lambda = kc/h$. The characteristic equation is

$$z^2 - 2(1 - 2\lambda^2 \sin^2\frac{\xi}{2})z + 1 = 0,$$

which has the solutions

$$z_{1,2} = 1 - 2\lambda^2\sin^2\frac{\xi}{2} \pm 2\lambda\sin\frac{\xi}{2}\sqrt{\lambda^2\sin^2\frac{\xi}{2} - 1}. \tag{1.9.3}$$

If $\lambda \leq 1$, we have

$$z_{1,2} = 1 - 2\lambda^2\sin^2\frac{\xi}{2} \pm 2\lambda\sin\frac{\xi}{2}i\sqrt{1 - \lambda^2\sin^2\frac{\xi}{2}}, \tag{1.9.4}$$

with $|z_{1,2}| = 1$. If $\lambda > 1$, one of the roots will exceed 1 in magnitude leading to an unstable scheme. This restriction on the time step is quite natural here as well as for the leap-frog scheme in Section 1.3. The domain of dependence for the difference approximation must include the domain of dependence for the differential equation.

We shall come back to the wave equation and its generalizations in Chapter 10.

1.10. GENERALIZATION TO SEVERAL SPACE DIMENSIONS

In two space dimensions, the hyperbolic model problem becomes

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2}, \qquad -\infty < x_1, x_2 < \infty, \quad t \geq 0, \tag{1.10.1}$$

with initial data

$$u(x, 0) = f(x), \qquad -\infty < x_1, x_2 < \infty,$$

where $x = (x_1, x_2)$. Here, we assume that $f(x)$ is 2π-periodic in x_1 and x_2. If

$$f(x) = \frac{1}{2\pi}\, e^{i\langle\omega,x\rangle} \hat{f}(\omega), \qquad \omega = (\omega_1, \omega_2),$$

we make the ansatz

$$u = \frac{1}{2\pi}\, e^{i\langle\omega,x\rangle} \hat{u}(\omega, t), \qquad \hat{u}(\omega, 0) = \hat{f}(\omega),$$

and obtain

$$\hat{u}_t(\omega, t) = i(\omega_1 + \omega_2)\hat{u}(\omega, t).$$

Thus,

$$u(x, t) = \frac{1}{2\pi}\, e^{i(\omega_1+\omega_2)t} e^{i\langle\omega,x\rangle} \hat{f}(\omega)$$

is the solution to our problem. For general f, we obtain, by the principle of superposition,

$$u = \frac{1}{2\pi} \sum_{\omega} e^{i(\omega_1+\omega_2)t} e^{i\langle\omega,x\rangle} \hat{f}(\omega) = f(x_1 + t, x_2 + t). \tag{1.10.2}$$

Thus, we can solve the problem as we did in the one-dimensional case. Also, the solution is constant along the characteristics, which are the lines $x_1 + t = \text{const}$ and $x_2 + t = \text{const}$.

We now discuss difference approximations. We introduce a time step, $k > 0$, and a two-dimensional grid by

$$x_j = (j_1 h, j_2 h), \quad j_\nu = 0, \ \pm 1, \pm 2, \ldots, \quad h = 2\pi/(N + 1),$$

and gridfunctions by

$$v_j^n = v(x_j, t_n), \qquad t_n = nk.$$

Corresponding to Eq. (1.2.7), we now have

$$v_j^{n+1} = \big(I + k(D_{0x_1} + D_{0x_2})\big)\, v_j^n. \tag{1.10.3}$$

Its Fourier transform is

$$\hat{v}^{n+1}(\omega) = (1 + i\lambda(\sin\ \xi_1 + \sin\ \xi_2))\, \hat{v}^n(\omega). \tag{1.10.4}$$

It is of the same form as Eq. (1.2.9). Therefore, the approximation is not useful. We add artificial viscosity, that is, we consider

$$v_j^{n+1} = \big(I + k(D_{0x_1} + D_{0x_2}) + \sigma kh(D_{+x_1}D_{-x_1} + D_{+x_2}D_{-x_2})\big)\, v_j^n. \tag{1.10.5}$$

As before, we can choose $\lambda = k/h,\ \sigma > 0$ such that $|\hat{Q}| \le 1$, that is, the approximation is stable.

The leap-frog and the Crank–Nicholson approximations are also easily generalized. They are

$$v_j^{n+1} = v_j^{n-1} + 2k(D_{0x_1} + D_{0x_2})v_j^n \tag{1.10.6}$$

and

$$\left(I - \frac{k}{2}\ (D_{0x_1} + D_{0x_2})\right) v_j^{n+1} = \left(I + \frac{k}{2}\ (D_{0x_1} + D_{0x_2})\right) v_j^n \tag{1.10.7}$$

respectively. The approximation (1.10.6) is stable for $k/h < 1/2$, whereas (1.10.7) is stable for all values of $\lambda = k/h$. Both methods are second-order accurate.

The parabolic model problem is

$$\begin{aligned} u_t &= u_{x_1x_1} + u_{x_2x_2}, \\ u(x,0) &= f(x). \end{aligned}$$

Like the one-dimensional problem, its solution

$$u = \frac{1}{2\pi} \sum_{\omega} e^{-|\omega|^2 t}\ e^{i\langle\omega,x\rangle}\, \hat{f}(\omega)$$

becomes "smoother" with time because the highly oscillatory waves ($|\omega| \gg 1$) are rapidly damped. We can easily construct difference approximations analogous to those used for the one-dimensional problem. We need only to replace D_+D_- in Section 1.6 by $D_{+x_1}D_{-x_1} + D_{+x_2}D_{-x_2}$. The analysis proceeds as before. The explicit Euler method in Eq. (1.6.6) is stable for $\sigma = k/h^2 \le \frac{1}{4}$, whereas the backward Euler, Crank–Nicholson, and DuFort–Frankel methods are unconditionally stable.

EXERCISES

1.10.1. Derive the stability condition for the leap-frog approximation to $u_t = u_x + u_y + u_z$, where the stepsizes Δx, Δy, and Δz may be different.

1.10.2. Derive the stability condition for the Euler approximation to $u_t = u_{xx} + u_{yy} + u_{zz}$. Prove that the DuFort–Frankel method is unconditionally stable for the same equation.

BIBLIOGRAPHIC NOTES

Most of the difference schemes introduced in this chapter were developed very early, in several cases, before the electronic computer was invented. The leap-frog scheme was discussed in a classical paper by *Courant–Friedrichs–Levy* (Courant et al., 1928). In the same paper, the so-called *CFL condition* was introduced, that is, the domain of dependence of the difference scheme must include the domain of dependence of the differential equation. Today, one often uses the term "CFL number" which, for the model equation $u_t = au_x$, means $\lambda = ka/h$.

The Lax–Friedrichs scheme was introduced for conservation laws $u_t = F(u)_x$ by Lax (1954), and the Lax–Wendroff method was presented in its original form by Lax and Wendroff (1960). Various versions have been presented later, but for the simple model equations we have been considering so far, they are identical. Any approximation of a hyperbolic equation that is a one-step explicit scheme with a centered second-order accurate approximation complemented with a damping term of second order, is usually called a Lax–Wendroff type approximation.

The Crank–Nicholson approximation was initially constructed for parabolic heat conduction problems by Crank and Nicholson (1947). The same name has later been used for other types of equations, where centered difference operators are used in space and the trapezoidal rule is used for discretization in time. The DuFort–Frankel method for parabolic problems was introduced by DuFort and Frankel (1953).

Stability analysis based on Fourier modes as presented here goes back to von Neumann, who used it at Los Alamos National Laboratory during World War II. It was first published by Crank and Nicholson (1947) and later by Charney et al. (1950), von Neumann and Richtmyer (1950), and O'Brien et al. (1951).

In this book, we use Fourier series representations of periodic functions, but one could use Fourier integral representations of general L_2 functions as well:

$$
\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}\, d\omega, \\
\hat{f}(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx,
\end{aligned}
$$

see, for example, Richtmyer and Morton (1967). The gridfunctions can be extended such that they are defined everywhere if the initial function is defined everywhere. In that way, the Fourier integrals are also defined for the solutions to the difference approximations. The Fourier transformed equations are exactly the same as they are for Fourier series and, accordingly, the stability conditions derived in Fourier space will be identical.

The stability definition 1.2.1 allows for an exponential growth, and it is satisfied if $|\hat{Q}| \leq 1 + \mathcal{O}(k)$. We notice that the condition $|\hat{Q}| \leq 1$ used in all our examples is stronger. However, if k/h is kept constant, our approximations are not explicitly dependent on k for the hyperbolic model equation. The same conclusion holds for the parabolic model problem if k/h^2 is constant. Therefore, $|\hat{Q}| \leq 1$ is the only possibility for a stable scheme in these cases.