MATH2103: Lecture Note on Numerical Solution of Partial Differential Equations

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2025 年 9 月 29 日

Contents

Chapter 1

Parabolic Equations

1.1 Method of Separation of Variables

A linear parabolic equation takes the general form

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(b(x,t)\frac{\partial u}{\partial x}\right) + c(x,t)u + d(x,t),$$

where b is strictly positive. If $x = 0$ is a left-hand boundary, the boundary condition will be of the form

$$\alpha_0(t)u + \alpha_1(t)\frac{\partial u}{\partial x} = \alpha_2(t),$$

where

$$\alpha_0 \geq 0, \alpha_1 \leq 0 \text{ and } \alpha_0 - \alpha_1 > 0.$$

If $x = 1$ is a right-hand boundary, the boundary condition will be of the form

$$\beta_0(t)u + \beta_1(t)\frac{\partial u}{\partial x} = \beta_2(t),$$

where

$$\beta_0 \geq 0, \beta_1 \geq 0 \text{ and } \beta_0 + \beta_1 > 0.$$

At this moment, we consider a simple model,

$$\begin{aligned} u_t &= u_{xx}, \quad \text{for } t > 0,\ 0 < x < 1, \\ u(0,t) &= u(1,t) = 0, \quad \text{for } t > 0, \\ u(x,0) &= u^0(x), \quad \text{for } 0 \leq x \leq 1. \end{aligned} \tag{1.1}$$

Using the method of separation of variables, we look for a solution of the special form $u(x,t) = f(x)g(t)$; substituting into the PDE to obtain

$$\frac{g'}{g} = \frac{f''}{f} = -k^2.$$

We arrive at the solution

$$u(x,t) = e^{-k^2 t}\sin kx.$$

Solving the SL problem for f, we obtain

$$k = m\pi, \quad m = 1, 2, \ldots$$

We eventually write the solution as a linear combination

$$u(x,t) = \sum_{m=1}^{\infty} a_m e^{-(m\pi)^2 t} \sin m\pi x. \tag{1.2}$$

Writing $t = 0$ we obtain

$$\sum_{m=1}^{\infty} a_m \sin m\pi x = u^0(x).$$

This show that a_m are the coefficients in the Fourier sine series expansion of $u^0(x)$, and are therefore given by

$$a_m = 2\int_0^1 u^0(x) \sin m\pi x \; dx.$$

This analytic solution is closely related to the numerical solution given by the **spectral method**.

1.2 An Explicit Scheme

We focus on the closed domain $[0,1] \times [0, t_F]$ where t_F can be as large as we like. We shall write Δx and Δt for the line spacings. The crossing points

$$x_j = j\Delta x, \; t_n = n\Delta t, \quad j = 0, 1, \ldots, J, \; n = 0, 1, \ldots,$$

where $\Delta x = 1/J$ are called the grid points or mesh points. We seek approximations of the solution at these mesh points; these approximate values will be denoted by

$$U_j^n \approx u(x_j, t_n).$$

The simplest different scheme uses a **forward difference** for the time derivative

$$\frac{v(x_j, t_{n+1}) - v(x_j, t_n)}{\Delta t} \approx \frac{\partial v}{\partial t}(x_j, t_n),$$

for any function v. The scheme uses a centred second difference for the second order space derivative

$$\frac{v(x_{j+1}, t_n) - 2v(x_j, t_n) + v(x_{j-1}, t_n)}{(\Delta x)^2} \approx \frac{\partial^2 v}{\partial x^2}(x_j, t_n).$$

Writing the numerical solution as U_j^n, the approximation to the PDE becomes

$$\begin{aligned} \frac{U_j^{n+1} - U_j^n}{\Delta t} &= \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{(\Delta x)^2}, \\ U_j^{n+1} &= U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right), \quad \mu := \frac{\Delta t}{(\Delta x)^2}. \end{aligned} \tag{1.3}$$

The pattern of grid points involved in (1.3) is shown in Fig. 1.1. Clearly each value at time level t_{n+1} can be independently calculated from values at time level t_n; for this reason this is called an **explicit difference scheme**. From the initial and boundary values

$$U_j^0 = u^0(x_j), \quad j = 1, 2, J-1, \tag{1.4}$$

$$U_0^n = U_J^n = 0, \quad n = 0, 1, 2, \ldots, \tag{1.5}$$

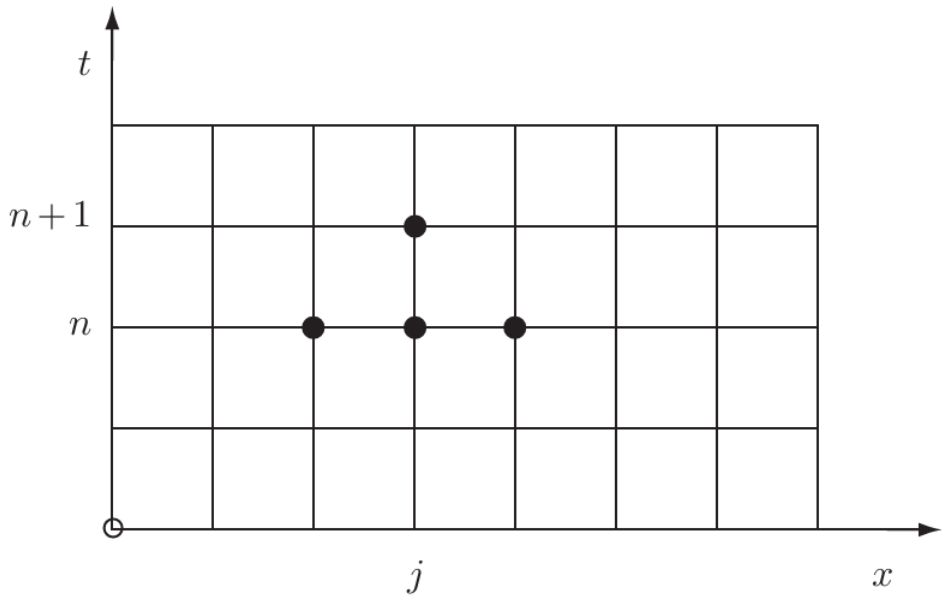

Figure 1.1: An explicit scheme.

we can calculate all the interior values for successive values of n. We shall assume for the moment that the initial and boundary data are consistent at the two corners; this means that

$$u^0(0) = u^0(1) = 0,$$

so that the solution does not have a discontinuity at the corners of the domain.

However, if we carry out a calculation using (1.3), (1.4) and (1.5), we soon discover that the numerical results depend critically on the value of μ, which relates the sizes of the time step and the space step. In Fig. 1.2 we show results corresponding to initial data in the form of a 'hat function',

$$u^0(x) = \begin{cases} 2x & \text{if } 0 \le x \le \frac{1}{2}, \\ 2 - 2x & \text{if } \frac{1}{2} \le x \le 1. \end{cases} \tag{1.6}$$

Two sets of results are displayed; both use $J = 20$, $\Delta x = 0.05$. The first set uses $\Delta t = 0.0012$, and the second uses $\Delta t = 0.0013$. The former clearly gives quite an accurate result, while the latter exhibits oscillations which grow rapidly with increasing values of t. This is a typical example of **stability or instability** depending on the value of the mesh ratio μ. The difference between the behaviour of the two numerical solutions is quite striking; these solutions use time steps which are very nearly equal, but different enough to give quite different forms of numerical solution.

Homework:

1. reproduce the above results.
2. $\mu = 1/6$, super-convergence.
3. What if we lose the compatibility of initial and boundary condition or we lose the smoothness of the initial condition.

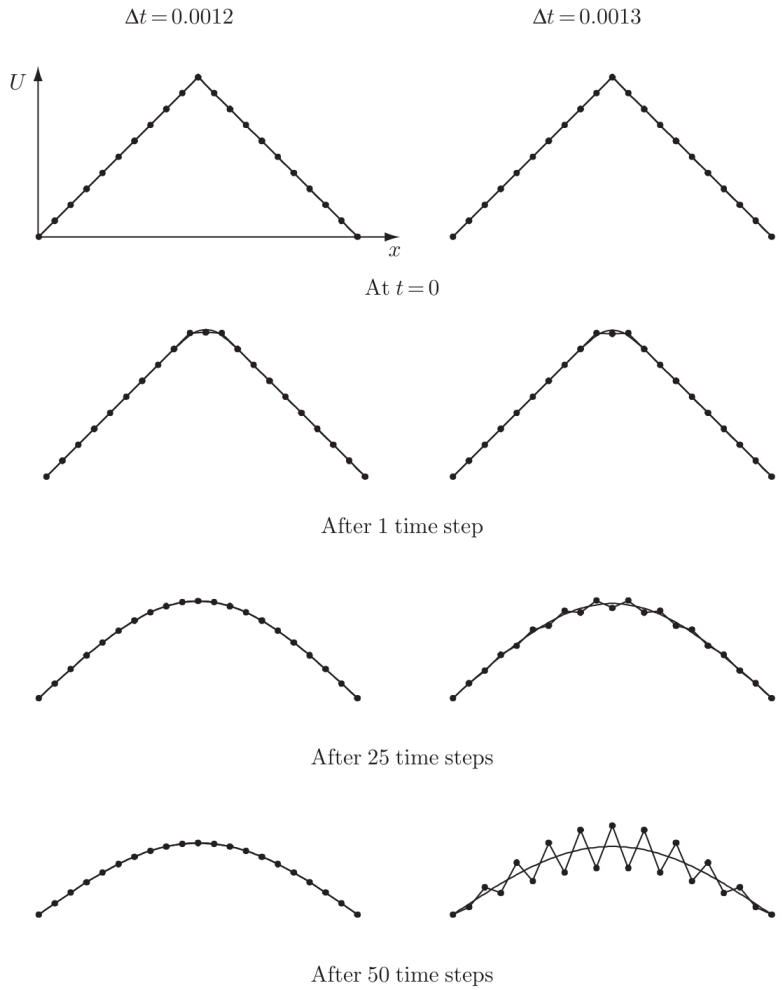

Fig. 2.2. Results obtained for the data of (2.24) with the explicit method; $J = 20$, $\Delta x = 0.05$. The exact solution is shown by the full curved line.

Figure 1.2: Results obtained with the explicit method.

1.3 Local Truncation Error

We define three kinds of finite differences:

forward differences

$$\begin{aligned} \Delta_{+t} v(x,t) &:= \quad v(x, t+\Delta t) - v(x,t), \\ \Delta_{+x} v(x,t) &:= \quad v(x+\Delta x, t) - v(x,t); \end{aligned}$$

backward differences

$$\begin{aligned} \Delta_{-t} v(x,t) &:= \quad v(x,t) - v(x, t-\Delta t), \\ \Delta_{-x} v(x,t) &:= \quad v(x,t) - v(x-\Delta x, t); \end{aligned}$$

central differences

$$\begin{aligned}\delta_t v(x,t) := &\quad v(x,t+\frac{1}{2}\Delta t) - v(x,t-\frac{1}{2}\Delta t),\\ \delta_x v(x,t) := &\quad v(x+\frac{1}{2}\Delta x,t) - v(x-\frac{1}{2}\Delta x,t);\end{aligned}$$

When the first order central difference operator is applied twice, we obtain the very useful second order central differnce

$$\delta_x^2 v(x,t) := v(x+\Delta x,t) - 2v(x,t) + v(x-\Delta x,t).$$

We now go back to see the simple diffusion equation. A Taylor series expansion of the forward difference in t gives for the solution of (1.1)

$$\begin{aligned}\Delta_{+t}u(x,t) &= u(x,t+\Delta t) - u(x,t)\\ &= u_t\Delta t + \frac{1}{2}u_{tt}(\Delta t)^2 + \frac{1}{6}u_{ttt}(\Delta t)^3 + \cdots\\ &= u_t\Delta t + \frac{1}{2}u_{tt}(x,\eta)(\Delta t)^2\\ &= u_t\Delta t + \frac{1}{2}u_{tt}(\Delta t)^2 + \frac{1}{6}u_{ttt}(x,\eta_2)(\Delta t)^3.\end{aligned}$$

By adding together the Taylor series expansions in the x variable for $\Delta_{+x}u$ and $\Delta_{-x}u$, we see that all the odd powers of Δx cancel, giving

$$\begin{aligned}\delta_x^2 u(x,t) &= u_{xx}(\Delta x)^2 + \frac{1}{12}u_{xxxx}(\Delta x)^4 + \frac{2}{6!}u_{xxxxxx}(\Delta x)^6 + \cdots\\ &= u_{xx}(\Delta x)^2 + \frac{1}{12}u_{xxxx}(\xi,t)(\Delta x)^4.\end{aligned} \tag{1.7}$$

We can now define the truncation error of the scheme (1.3). The truncation error is then the difference between the two sides of the equation, when **the approximation U_j^n is replaced throughout by the exact solution** $u(x_j,t_n)$ of the differential equation. Indeed, at any point away from the boundary we can define the

truncation error $T(x,t)$

$$T(x,t) := \frac{\Delta_{+t}u(x,t)}{\Delta t} - \frac{\delta_x^2 u(x,t)}{(\Delta x)^2}$$

so that

$$\begin{aligned}T(x,t) &= (u_t - u_{xx}) + (\frac{1}{2}u_{tt}\Delta t - \frac{1}{12}u_{xxxx}(\Delta x)^2 + \cdots)\\ &= \frac{1}{2}u_{tt}\Delta t - \frac{1}{12}u_{xxxx}(\Delta x)^2 + \cdots\end{aligned}$$

where these leading terms are called the **principal part** of the truncation error, and we have used the fact that u satisfies the differential equation.

Using Taylor expansion, the truncation error becomes

$$T(x,t) = \frac{1}{2}u_{tt}(x,\eta)\Delta t - \frac{1}{12}u_{xxxx}(\xi,t)(\Delta x)^2,$$

from which it follows that

$$\begin{aligned}|T(x,t)| &\le \frac{1}{2}M_{tt}\Delta t + \frac{1}{12}M_{xxxx}(\Delta x)^2\\ &= \frac{1}{2}\Delta t\left[M_{tt} + \frac{1}{6\mu}M_{xxxx}\right],\end{aligned} \tag{1.8}$$

where M_{tt} is a bound for $|u_{tt}|$ and M_{xxxx} is a bound for $|u_{xxxx}|$.

Remark 1.3.1 *It is now clear why we assumed that the initial and boundary data for u were consistent, and why it is helpful if we can also assume that the initial data are sufficiently smooth. For then we can assume that the bounds M_{tt} and M_{xxxx} hold uniformly over the closed domain $[0,1]\times[0,t_F]$. Otherwise we must rely on the smoothing effect of the diffusion operator to ensure that for any $\tau > 0$ we can find bounds of this form which hold for the domain $[0,1]\times[\tau,t_F]$. This sort of difficulty can easily arise in problems which look quite straightforward. For example, suppose the boundary conditions specify that u must vanish on the boundaries $x = 0$ and $x = 1$, and that u must take the value 1 on the initial line, where $t = 0$. Then the solution u(x,t) is obviously discontinuous at the corners, and in the full domain defined by $0 < x < 1, t > 0$ all its derivatives are unbounded, so our bound for the truncation error is useless over the full domain. We shall see later how this problem can be treated by* ***Fourier analysis****.*

For the simple heat equation problem we see that

$$T(x,t) \to 0 \text{ as } \Delta x,\ \Delta t \to 0\ \forall (x,t) \in (0,1)\times[\tau,t_F),$$

independently of any relation between the two mesh sizes. We say that the scheme is **unconditionally consistent** with the differential equation. For a fixed ratio μ we also see from (1.8) that $|T|$ will behave asymptotically like $O(\Delta t)$ as $\Delta t \to 0$: except for special values of μ this will be the highest power of Δt for which such a statement could be made, so that the scheme is said to have **first order accuracy**.

Remark 1.3.2 *However, it is worth noting here that, since u satisfies $u_t = u_{xx}$ everywhere, we also have $u_{tt} = u_{xxxx}$ and hence*

$$T(x,t) = \frac{1}{2}(1 - \frac{1}{6\mu})u_{xxxx}\Delta t + O((\Delta t)^2).$$

Thus for $\mu = 1/6$ the scheme is ***second order accurate****. This however is a rather special case. Not only does it apply just for this particular choice of μ but also for more general equations with variable coefficients it cannot hold. For example, in the solution of the equation $u_t = b(x,t)u_{xx}$ it would require choosing a different time step Δt at each point.*

1.4 Convergence of the Explicit Scheme

Now suppose that we carry out a sequence of calculations using the same initial data and the same value of $\mu = \frac{\Delta t}{(\Delta x)^2}$, but with successive refinement of the two meshes, such that $\Delta t \to 0$ and $\Delta x \to 0$. Then we say that the scheme is **convergent** if, for any fixed point (x^*, t^*) in a given domain $(0,1)\times(\tau,t_F)$,

$$x_j \to x^*, \quad t_n \to t^* \quad \text{implies} \quad U_j^n \to u(x^*,t^*).$$

We shall prove that the explicit scheme for our problem is convergent if $\mu \le \frac{1}{2}$ (**more to say from view of stability**).

We need consider only points (x^*, t^*) which coincide with mesh points for sufficiently refined meshes; for convergence at all other points will follow from the continuity of $u(x,t)$. We also suppose that we can introduce an upper bound $\bar{T}(\Delta x, \Delta t)$ for the truncation error, which holds for all mesh points on a given mesh, and use the notation T_j^n for $T(x_j, t_n)$:

$$|T_j^n| \le \bar{T}. \tag{1.9}$$

We denote by e the error $U - u$ in the approximation; more precisely

$$e_j^n := U_j^n - u(x_j, t_n). \tag{1.10}$$

Now U_j^n satisfies the equation (1.3) exactly, while $u(x_j, t_n)$ leaves the remainder $T_j^n \Delta t$; this follows immediately from the definition of T_j^n. In detail,

$$\begin{aligned} U_j^{n+1} &= U_j^n + \mu \left(U_{j+1}^n - 2U_j^n + U_{j-1}^n \right), \\ u_j^{n+1} &= u_j^n + \mu \left(u_{j+1}^n - 2u_j^n + u_{j-1}^n \right) + T_j^n \Delta t. \end{aligned}$$

Hence by subtraction we obtain

$$e_j^{n+1} = e_j^n + \mu \delta_x^2 e_j^n - T_j^n \Delta t \tag{1.11}$$

which is in detail

$$e_j^{n+1} = (1 - 2\mu) e_j^n + \mu e_{j+1}^n + \mu e_{j-1}^n - T_j^n \Delta t. \tag{1.12}$$

The important point for the proof is that if $\mu \leq \frac{1}{2}$ the coefficients of the three terms e^n on the right of this equation are all positive, and add up to unity. If we introduce the maximum error at a time step by writing

$$E^n := \max |e_j^n|, j = 0, 1, \ldots, J, \tag{1.13}$$

the fact that the coefficients are positive means that we can omit the modulus signs in the triangle inequality to give

$$\left| e_j^{n+1} \right| \leq (1 - 2\mu) E^n + \mu E^n + \mu E^n + |T_j^n| \Delta t \leq E^n + \bar{T} \Delta t. \tag{1.14}$$

Since this inequality holds for all values of j from 1 to $J - 1$, we have

$$E^{n+1} \leq E^n + \bar{T} \Delta t. \tag{1.15}$$

Suppose for the moment that the bound (1.9) holds on the finite interval $[0, t_F]$; and since we are using the given initial values for U_j^n we know that $E^0 = 0$. A very simple induction argument then shows that $E^n \leq n \bar{T} \Delta t$. Hence we obtain from (1.8)

$$E^n \leq \frac{1}{2} \Delta t \left[M_{tt} + \frac{1}{6\mu} M_{xxxx} \right] t_F \to 0 \quad \text{as} \quad \Delta t \to 0. \tag{1.16}$$

Remark 1.4.1 *In practical engineering problem, $E^0 \neq 0$ may happen due to the* ***measurement of the initial condition****.*

Theorem 1.4.2 *If a refinement path satisfies $\mu_i \leq \frac{1}{2}$ for all sufficiently large values of i, and the positive numbers n_i, j_i are such that*

$$n_i (\Delta t)_i \to t > 0, \quad j_i (\Delta x)_i \to x \in [0, 1],$$

and if $|u_{xxxx}| \leq M_{xxxx}$ uniformly in $[0, 1] \times [0, t_F]$, then the approximations $U_{j_i}^{n_i}$ generated by the explicit difference scheme (1.3) for $i = 0, 1, 2 \ldots$ converge to the solution $u(x, t)$ of the differential equation, uniformly in the region.

Remark 1.4.3 *Such a convergence theorem is the least that one can expect of a numerical scheme; it shows that arbitrarily high accuracy can be attained by use of a sufficiently fine mesh. Of course, it is also somewhat impractical. As the mesh becomes finer, more and more steps of calculation are required, and* ***the effect of rounding errors in the calculation would become significant and would eventually completely swamp the truncation error.***

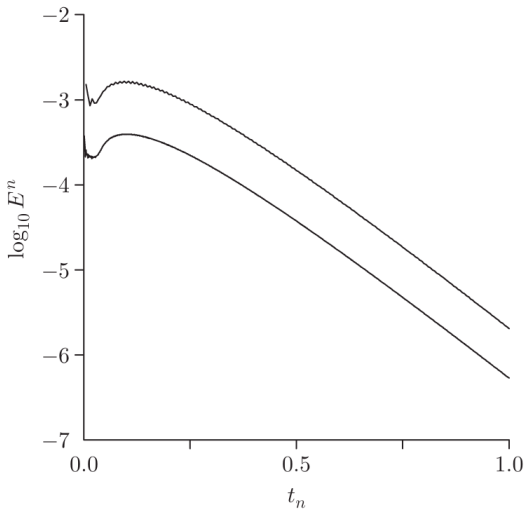

Fig. 2.4. Error decay for the explicit method applied to the heat equation with initial condition $u(x,0) = x(1-x)$. The top curve is for $\Delta x = 0.1$, $\mu = 0.5$, and the bottom curve is for $\Delta x = 0.05, \mu = 0.5$.

Figure 1.3: Error decay.

As an example with smoother properties than is given by the data of (1.6), consider the solution of the heat equation with

$$\begin{aligned} u(x,0) &= x(1-x), \\ u(0,t) &= u(1,t) = 0, \end{aligned} \tag{1.17}$$

on the region $[0,1] \times [0,1]$. Errors obtained with the explicit method are shown in Fig. 1.3. This shows a graph of $\log_{10} E^n$ against t_n, where E^n is given by (1.13). Two curves are shown; one uses $J = 10$, $\Delta x = 0.1$, and the other uses $J = 20$, $\Delta x = 0.05$. Both have $\mu = \frac{1}{2}$, which is the largest value consistent with stability.

Obs 1: The two curves show clearly how the error behaves as the grid size is reduced: they are very similar in shape, and for each value of t_n, the ratio of the two values of E^n is close to 4, the ratio of the values of $\Delta t = \frac{1}{2}(\Delta x)^2$.

Obs 2: Notice also that after some early variation the error tends to zero as t increases; our error bound in (1.15) is pessimistic, as it continues to increase with t.

Obs 3: The early variation in the error results from the lack of smoothness in the corners of the domain already referred to. We will discuss this in more detail in the next section and in the following Section.

1.5 Fourier Analysis

We have already expressed the exact solution of the differential equation as a Fourier series in equation (1.2); this expression is based on the observation that a particular set of Fourier modes are exact solutions.

We can now easily show that a similar Fourier mode is an exact solution of the difference equations (1.3). Suppose we substitute

$$U_j^n = (\lambda)^n e^{ik(j\Delta x)},$$

into the difference equation (1.3), putting $U_j^{n+1} = \lambda U_j^n$ and similarly for the other terms. We can then divide by U_j^n and see that this Fourier mode is a solution for all values of n and j provided that

$$\begin{aligned} \lambda &\equiv \lambda(k) = 1 + \mu\left(e^{ik\Delta x} - 2 + e^{-ik\Delta x}\right) \\ &= 1 - 2\mu(1 - \cos k\Delta x) \\ &= 1 - 4\mu \sin^2 \frac{1}{2} k\Delta x; \end{aligned} \tag{1.18}$$

$\lambda(k)$ is called the **amplification factor** for the mode. By taking $k = m\pi$ as in (1.2), we can therefore write our numerical approximation in the form

$$U_j^n = \sum_{-\infty}^{\infty} A_m e^{im\pi(j\Delta x)} [\lambda(m\pi)]^n. \tag{1.19}$$

Remark 1.5.1 *The low frequency terms in this expansion give a good approximation to the exact solution of the differential equation, given by (1.2), because the series expansions of* ***one-step marching*** *for* $\lambda(k)$ *and* $\exp(-k^2\Delta t)$ *match reasonably well:*

$$\begin{aligned} \exp(-k^2\Delta t) &= 1 - k^2\Delta t + \frac{1}{2}k^4(\Delta t)^2 - \cdots, \\ \lambda(k) &= 1 - 2\mu\left[\frac{1}{2}(k\Delta x)^2 - \frac{1}{24}(k\Delta x)^4 + \cdots\right] \\ &= 1 - k^2\Delta t + \frac{1}{12}k^4\Delta t(\Delta x)^2 - \cdots. \end{aligned} \tag{1.20}$$

Indeed these expansions provide an alternative means of investigating the truncation error of our scheme. It is easy to see that we will have at least first order accuracy. In fact it is quite easy to show that there exists a constant $C(\mu)$ *depending only on the value of* μ *such that*

$$|\lambda(k) - e^{-k^2\Delta t}| \le C(\mu)k^4(\Delta t)^2 \quad \forall k, \Delta t > 0. \tag{1.21}$$

(When in the special case of $(\Delta x)^2 = 6\Delta t$ *we shall have second order accuracy.)*

Remark 1.5.2 *There is a relation between the real space and Fourier space for Fourier transform or Fourier series. In general,* ***finite-length interval in real space corresponds to equal-distance infinite-length grids. If there are finite number of real space grids, then the number of Fourier modes is also truncated to be finite.*** *For the Fourier method, we have*

$$\begin{aligned} U_j^{n+1} &= U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right), \\ \sum_k \hat{U}^{n+1}(k)e^{ikx} &= \sum_k \hat{U}^n(k)e^{ikx} + \mu\left(\sum_k \hat{U}^n(k)e^{ik(x+\Delta x)} - 2\sum_k \hat{U}^n(k)e^{ikx} + \sum_k \hat{U}^n(k)e^{ik(x-\Delta x)}\right), \\ \sum_k \hat{U}^{n+1}(k)e^{ikx} &= \sum_k \left[1 + \mu\left(e^{ik\Delta x} - 2 + e^{-ik\Delta x}\right)\right]\hat{U}^n(k)e^{ikx}, \\ \hat{U}^{n+1}(k) &= \left[1 + \mu\left(e^{ik\Delta x} - 2 + e^{-ik\Delta x}\right)\right]\hat{U}^n(k). \end{aligned}$$

The term $1 + \mu\left(e^{ik\Delta x} - 2 + e^{-ik\Delta x}\right)$ *is called the* ***amplification factor****.*

Theorem 1.4.2 establishes convergence and an error bound under the restriction $\mu \le 1/2$, but it does not show what happens if this condition is not satisfied. Our analysis of the Fourier modes shows what

happens to the high frequency components in this case. In (1.18), $k\Delta x = m\pi\Delta x$ can take value of π which corresponds to the maximum value of $\sin^2 \frac{1}{2}k\Delta x$. For stability, we require that

$$\begin{aligned} |\lambda| &= \left|1 - 4\mu \sin^2 \frac{1}{2}k\Delta x\right| \leq 1, \\ 0 &\leq 4\mu \sin^2 \frac{1}{2}k\Delta x \leq 2, \\ 0 &\leq \mu \leq \frac{1}{2}. \end{aligned}$$

case 1: When $\mu \leq 1/2$, the scheme is stable and each mode of the solution decays or keeps unchanged.
case 2: When $\mu > 1/2$, the scheme is unstable and some mode of the solution grows unboundedly.

Definition 1.5.3 *For the present model problem we shall say that a method is stable if there exists a constant K, independent of k, such that*

$$|[\lambda(k)]^n| \leq K, \quad for \quad n\Delta t \leq t_F, \forall k. \tag{1.22}$$

Theorem 1.5.4 *For stability we require the condition, due to* ***von Neumann****,*

$$|\lambda(k)| \leq 1 + K'\Delta t \tag{1.23}$$

to hold for all k. We shall find that such a stability condition is necessary and sufficient for the convergence of a consistent difference scheme approximating a single differential equation. Thus for the present model problem the method is unstable when $\mu > \frac{1}{2}$ and stable when $\mu \leq \frac{1}{2}$. We shall formulate a general definition of stability in a later chapter.

Remark 1.5.5 *We have used a representation for U_j^n as the infinite Fourier series (1.19), since it is easily comparable with the exact solution. However on the discrete mesh there are only a finite number of distinct modes; modes with wave numbers k_1 and k_2 are indistinguishable if $(k_1 - k_2)\Delta x$ is a multiple of 2π. It may therefore be more convenient to expand U_j^n as a linear combination of the distinct modes corresponding to*

$$k = m\pi, \quad m = -(J-1), -(J-2), \ldots, -1, 0, 1, \ldots, J. \tag{1.24}$$

*(**Here in fact only half of above modes corresponding to sine modes are Fourier basis functions of the solution.**) The highest mode which can be carried by the mesh has $k = J\pi$, or $k\Delta x = \pi$; this mode has the values ± 1 at alternate points on the mesh. We see from (1.18) that it is also the most unstable mode for this difference scheme, as it often is for many difference schemes, and has the amplification factor $\lambda(J\pi) = 1 - 4\mu$. It is the fastest growing mode when $\mu > \frac{1}{2}$, which is why it eventually dominates the solutions shown in Fig. 1.2.*

We can also use this Fourier analysis to extend the convergence theorem to the case where the initial data $u^0(x)$ are continuous on $[0, 1]$, but may not be smooth, in particular at the corners. We no longer have to assume that the solution has sufficient bounded derivatives that u_{xxxx} and u_{tt} are uniformly bounded on the region considered. Instead we just assume that the Fourier series expansion of $u^0(x)$ is absolutely

convergent. We suppose that μ is fixed, and that $\mu \leq \frac{1}{2}$. Consider the error, as before,

$$\begin{aligned} e_j^n &= U_j^n - u(x_j, t_n) \\ &= \sum_{-\infty}^{\infty} A_m e^{im\pi j\Delta x} \left\{[\lambda(m\pi)]^n - e^{-m^2\pi^2 n\Delta t}\right\}, \end{aligned} \tag{1.25}$$

where we have also used the full Fourier series for $u(x,t)$ instead of the sine series as in the particular case of (1.2); this will allow treatment other than of the simple Dirichlet boundary conditions. We now split this infinite sum into two parts. Given an arbitrary positive ϵ, we choose m_0 such that

$$\sum_{|m|>m_0} |A_m| \leq \frac{1}{4}\epsilon. \tag{1.26}$$

We know that this is possible, because of the absolute convergence of the series. If both $|\lambda_1| \leq 1$ and $|\lambda_2| \leq 1$, then

$$|(\lambda_1)^n - (\lambda_2)^n| \leq n|\lambda_1 - \lambda_2|; \tag{1.27}$$

so from (1.21) we have

$$\begin{aligned} |e_j^n| &\leq \frac{1}{2}\epsilon + \sum_{|m|\leq m_0} |A_m| \left| [\lambda(m\pi)]^n - e^{-m^2\pi^2 n\Delta t} \right| \\ &\leq \frac{1}{2}\epsilon + \sum_{|m|\leq m_0} |A_m| nC(\mu) \left(m^2\pi^2\Delta t\right)^2. \end{aligned} \tag{1.28}$$

We can thus deduce that

$$|e_j^n| \leq \frac{1}{2}\epsilon + t_F C(\mu)\pi^4 \left[\sum_{|m|\leq m_0} |A_m| m^4 \right] \Delta t \tag{1.29}$$

and by taking Δt sufficiently small we can obtain $|e_j^n| \leq \epsilon$ for all (x_j, t_n) in $[0,1] \times [0, t_F]$. Note how the sum involving $A_m m^4$ plays much the same role as the bound on u_{xxxx} in the earlier analysis, but by making more precise use of the stability properties of the scheme we do not require that this sum is convergent.

1.6 An Implicit Scheme

• The stability limit $\Delta t < \frac{1}{2}(\Delta x)^2$ is a very severe restriction, and implies that very many time steps will be necessary to follow the solution over a reasonably large time interval.

• Moreover, if we need to reduce Δx to improve the accuracy of the solution the amount of work involved increases very rapidly, since we shall also have to reduce Δt.

• We shall now show how the use of a backward time difference gives a difference scheme which avoids this restriction, but at the cost of a slightly more sophisticated calculation.

• If we replace the forward time difference by the backward time difference, the space difference remaining the same, we obtain the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{(\Delta x)^2}. \tag{1.30}$$

• Let $\mu = \Delta t/(\Delta x)^2$, and has the stencil shown in Fig. 2.5.

• This is an example of an **implicit** scheme, which is not so easy to use as the explicit scheme described earlier. The scheme (1.30) involves three unknown values of U on the new time level $n+1$; we cannot immediately calculate the value of U_j^{n+1} since the equation involves the two neighbouring values U_{j+1}^{n+1} and U_{j-1}^{n+1}, which are also unknown. We must now write the equation in the form

$$-\mu U_{j-1}^{n+1} + (1+2\mu)U_j^{n+1} - \mu U_{j+1}^{n+1} = U_j^n. \tag{1.31}$$

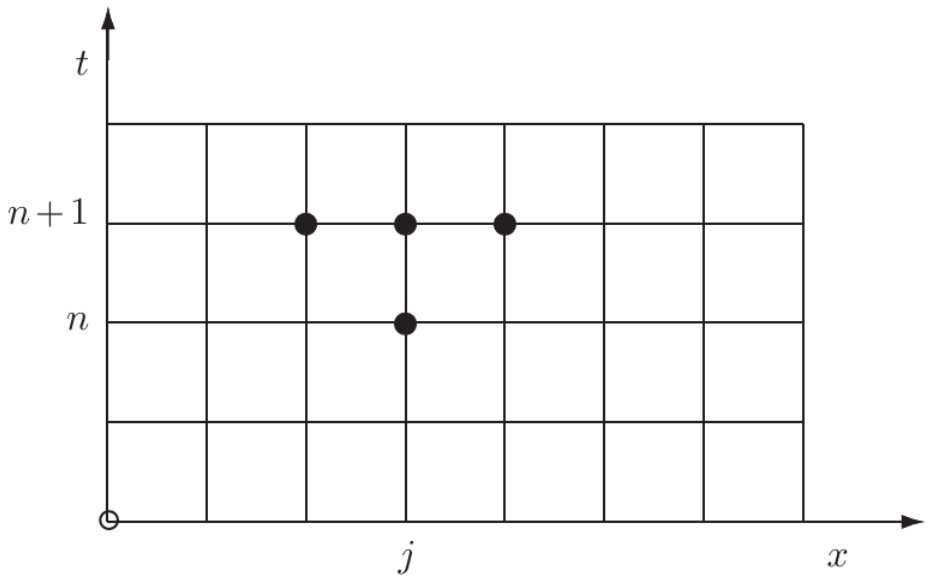

Fig. 2.5. The fully implicit scheme.

Figure 1.4: An implicit scheme.

• Note that in the first and last of these equations, corresponding to $j = 1$ and $j = J-1$, we incorporate the known values of U_0^{n+1} and U_J^{n+1} given by the boundary conditions.

• The importance of the implicit method is, of course, that the time steps can be much larger, for, as we shall see, there is no longer any stability restriction on Δt. We construct a solution of the difference equations for Fourier modes of the same form as before,

$$U_j^n = (\lambda)^n e^{ik(j\Delta x)}. \tag{1.32}$$

This will satisfy (1.31) provided that

$$\lambda - 1 = \mu\lambda(e^{ik\Delta x} - 2 + e^{-ik\Delta x})$$

$$= -4\mu\lambda \sin^2 \frac{1}{2} k\Delta x, \tag{1.33}$$

which shows that

$$\lambda = \frac{1}{1 + 4\mu \sin^2 \frac{1}{2} k\Delta x}. \tag{1.34}$$

Evidently we have $0 < \lambda < 1$ for any positive choice of μ, so that this implicit method is **unconditionally stable**. As we shall see in the next section, the truncation error is much the same size as that of the explicit scheme, but we no longer require any restriction on μ to ensure that no Fourier mode grows as n increases.

Remark 1.6.1 *The time step is still limited by the requirement that the truncation error must stay small, but in practice it is found that in most problems the implicit method can use a much larger Δt than the explicit method; although each step takes about twice as much work, the overall amount of work required to reach the time t_F is much less.*

• The system of equations to be solved is tridiagonal: equation number j in the system only involves unknowns with numbers $j-1$, j and $j+1$, so that the matrix of the system has non-zero elements only on the diagonal and in the positions immediately to the left and to the right of the diagonal. We shall meet such systems again, and it is useful here to consider a more general system of the form

$$-a_j U_{j-1} + b_j U_j - c_j U_{j+1} = d_j, \quad j = 1, 2, \ldots, J-1,$$

with

$$U_0 = 0, \quad U_J = 0.$$

Here we have written the unknowns U_j, omitting the superscript for the moment. The coefficients a_j, b_j and c_j, and the right-hand side d_j, are given, and we assume that they satisfy the conditions

$$a_j > 0, \quad b_j > 0, \quad c_j > 0,$$

$$b_j > a_j + c_j.$$

Though stronger than necessary, these conditions ensure that the matrix is **diagonally dominant**, with the diagonal element in each row being at least as large as the sum of the absolute values of the other elements. It is easy to see that these conditions are satisfied by our difference equation system.

1.7 The weighted average or θ-method

1.7.1 The weighted average

• A natural generalisation is to an approximation which uses all six of these points. This can be regarded as taking a weighted average of the two formulae. Since the time difference on the left-hand sides is the same, we obtain the six-point scheme (see Fig. 1.5)

$$U_j^{n+1} - U_j^n = \mu \left[\theta \delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n\right], \quad j = 1, 2, \ldots, J-1. \tag{1.35}$$

• We shall assume that we are using an average with nonnegative weights, so that $0 \le \theta \le 1$; $\theta = 0$ gives the explicit scheme, $\theta = 1$ the fully implicit scheme.

• For any $\theta \neq 0$, we have a tridiagonal system to solve for U_j^{n+1}, namely,

$$-\theta\mu U_{j-1}^{n+1} + (1 + 2\theta\mu) U_j^{n+1} - \theta\mu U_{j+1}^{n+1} = [1 + (1-\theta)\mu\delta_x^2] U_j^n. \tag{1.36}$$

1.7.2 stability

• Let us consider the stability of this one-parameter family of schemes by using Fourier analysis. Substituting the mode (1.32) into equation (1.35), we obtain

$$\begin{aligned} \lambda - 1 &= \mu[\theta\lambda + (1-\theta)]\left(e^{ik\Delta x} - 2 + e^{-ik\Delta x}\right) \\ &= \mu[\theta\lambda + (1-\theta)]\left(-4\sin^2\frac{1}{2}k\Delta x\right), \end{aligned}$$

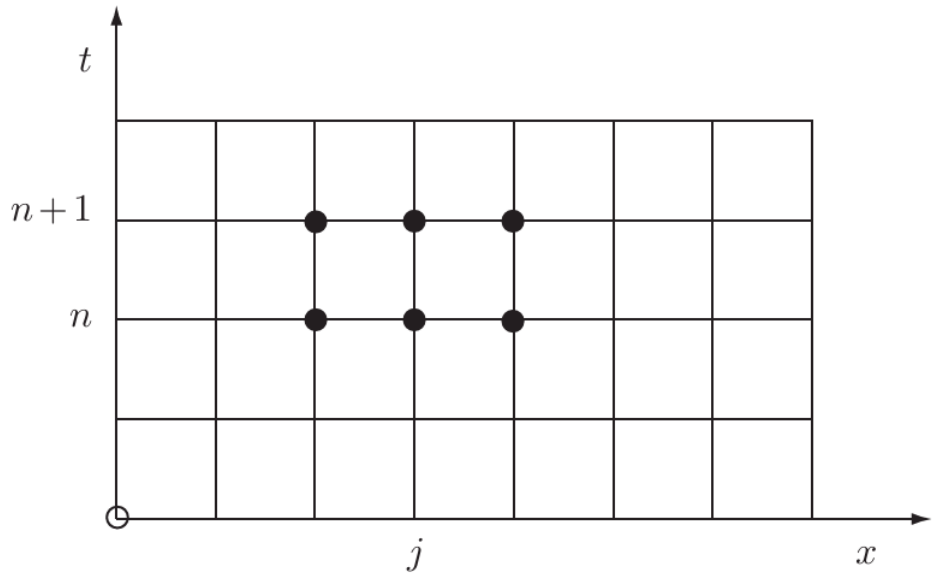

Fig. 2.6. The θ-method.

Figure 1.5: The θ-method.

i.e.,

$$\lambda = \frac{1 - 4(1-\theta)\mu \sin^2 \frac{1}{2}k\Delta x}{1 + 4\theta\mu \sin^2 \frac{1}{2}k\Delta x}. \tag{1.37}$$

- Because $\mu > 0$, and we are assuming that $0 \le \theta \le 1$, it is clear that we can never have $\lambda > 1$.
- Thus, instability arises only through the possibility that $\lambda < -1$, that is that

$$1 - 4(1-\theta)\mu \sin^2 \frac{1}{2}k\Delta x < -\left[1 + 4\theta\mu \sin^2 \frac{1}{2}k\Delta x\right], \tag{1.38}$$

i.e.,

$$4\mu(1-2\theta)\sin^2 \frac{1}{2}k\Delta x > 2. \tag{1.39}$$

The mode most liable to instability is the one for which the left side is largest: as before this is the most rapidly oscillatory mode, for which $k\Delta x = \pi$. This is an unstable mode if

$$\mu(1-2\theta) > \frac{1}{2}. \tag{1.40}$$

- This includes the earlier explicit case, $\theta = 0$: and it also shows that the fully implicit scheme with $\theta = 1$ is not unstable for any value of μ. Indeed no scheme with $\theta \ge \frac{1}{2}$ is unstable for any μ. If condition (1.40) is satisfied there can be unbounded growth over a fixed time as $\Delta t \to 0$ and hence $n \to \infty$: on the other hand if (1.40) is not satisfied, we have $|\lambda(k)| \le 1$ for every mode k, so that no mode grows at all and the scheme is stable.

Thus we can summarise the necessary and sufficient conditions for the stability of (1.35) as

$$\begin{aligned} &\text{when } 0 \le \theta < \frac{1}{2}, \text{ stable if and only if } \mu \le \frac{1}{2}(1-2\theta)^{-1} \\ &\text{when } \frac{1}{2} \le \theta \le 1, \text{ stable for all } \mu. \end{aligned} \tag{1.41}$$

1.7.3 consistency

Working from (1.35) we therefore have, using the superscript/subscript notation for u as well as U,

$$\begin{aligned} u_j^{n+1} &= \left[u + \frac{1}{2}\Delta t u_t + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2 u_{tt} + \frac{1}{6}\left(\frac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots\right]_j^{n+1/2}, \\ u_j^{n} &= \left[u - \frac{1}{2}\Delta t u_t + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2 u_{tt} - \frac{1}{6}\left(\frac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots\right]_j^{n+1/2}. \end{aligned}$$

If we subtract these two series, all the even terms of the two Taylor series will cancel, and we obtain

$$\delta_t u_j^{n+1/2} = u_j^{n+1} - u_j^n = \left[\Delta t u_t + \frac{1}{24}(\Delta t)^3 u_{ttt} + \cdots\right]_j^{n+1/2}. \tag{1.42}$$

Also from (1.7) we have

$$\delta_x^2 u_j^{n+1} = \left[(\Delta x)^2 u_{xx} + \frac{1}{12}(\Delta x)^4 u_{xxxxx} + \frac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right]_j^{n+1}. \tag{1.43}$$

We now expand each term in this series in powers of Δt, about the point $(x_j, t_{n+1/2})$. For simplicity in presenting these expansions, we omit the superscript and subscript, so it is understood that the resulting expressions are all to be evaluated at this point. This gives

$$\begin{aligned} \delta_x^2 u_j^{n+1} &= \left[(\Delta x)^2 u_{xx} + \frac{1}{12}(\Delta x)^4 u_{xxxx} + \frac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right] \\ &\quad + \frac{1}{2}\Delta t\left[(\Delta x)^2 u_{xxt} + \frac{1}{12}(\Delta x)^4 u_{xxxxt} + \cdots\right] \\ &\quad + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2\left[(\Delta x)^2 u_{xxtt} + \cdots\right] + \cdots. \end{aligned}$$

There is a similar expansion for $\delta_x^2 u_j^n$: combining these we obtain

$$\begin{aligned} \theta\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n &= \left[(\Delta x)^2 u_{xx} + \frac{1}{12}(\Delta x)^4 u_{xxxx} + \frac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right] \\ &\quad + (\theta - \frac{1}{2})\Delta t\left[(\Delta x)^2 u_{xxt} + \frac{1}{12}(\Delta x)^4 u_{xxxxt} + \cdots\right] \\ &\quad + \frac{1}{8}(\Delta t)^2(\Delta x)^2[u_{xxtt}] + \cdots. \end{aligned} \tag{1.44}$$

Here we have retained more terms than we shall normally need to calculate the principal part of the truncation error, in order to show clearly the pattern for all the terms involved. In addition we have not exploited yet the fact that u is to satisfy the differential equation, so that (2.80) and (2.82) hold for any sufficiently smooth functions. If we now use these expansions to calculate the truncation error we obtain

$$T_j^{n+1/2} := \frac{\delta_t u_j^{n+1/2}}{\Delta t} - \frac{\theta\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n}{(\Delta x)^2} \tag{1.45}$$

$$\begin{aligned} &= [u_t - u_{xx}] + \left[\left(\frac{1}{2} - \theta\right)\Delta t u_{xxt} - \frac{1}{12}(\Delta x)^2 u_{xxxx}\right] \\ &\quad + \left[\frac{1}{24}(\Delta t)^2 u_{ttt} - \frac{1}{8}(\Delta t)^2 u_{xxtt}\right] \\ &\quad + \left[\frac{1}{12}\left(\frac{1}{2} - \theta\right)\Delta t(\Delta x)^2 u_{xxxxt} - \frac{2}{6!}(\Delta x)^4 u_{xxxxxx}\right] \end{aligned} \tag{1.46}$$

where we have still not carried out any cancellations but have merely grouped terms which are ripe for cancellation.

- The first term in (1.46) always cancels, so confirming consistency for all values of θ and μ.
- The second shows that we shall normally have first order accuracy (in Δt)
- but that the symmetric average $\theta = \frac{1}{2}$ is special: this value gives the well known and popular **Crank-Nicolson scheme**, named after those two authors who in a 1947 paper[1] applied the scheme very successfully to problems in the dyeing of textiles. Since the third term in (1.46) does not cancel even when we exploit the differential equation to obtain

$$T_j^{n+1/2} = -\frac{1}{12}\left[(\Delta x)^2 u_{xxxx} + (\Delta t)^2 u_{ttt}\right]_j^{n+1/2} + \cdots \tag{1.47}$$

(when $\theta = \frac{1}{2}$), we see that the Crank-Nicolson scheme is always second order accurate in both Δt and Δx: this means that we can exploit the extra stability properties of the scheme to take larger time steps, with for example $\Delta x = O(\Delta t)$, and because then the truncation error is $O((\Delta t)^2)$ we can achieve good accuracy economically.

- Another choice which is sometimes advocated is a generalisation of that discussed in previous Section. It involves eliminating the second term in (1.46) completely by relating the choice of θ to that of Δt and Δx so that

$$\theta = \frac{1}{2} - (\Delta x)^2/12\Delta t, \tag{1.48}$$

i.e.,

$$\mu = \frac{1}{6(1-2\theta)}, \tag{1.49}$$

but note that this requires $(\Delta x)^2 \leq 6\Delta t$ to ensure $\theta \geq 0$. This gives a value of θ less than $\frac{1}{2}$ but it is easy to see that the condition (1.41) is satisfied, so that it is stable. It reduces to $\mu = \frac{1}{6}$ for the explicit case $\theta = 0$. The resulting truncation error is

$$T_j^{n+1/2} = -\frac{1}{12}\left[(\Delta t)^2 u_{ttt} + \frac{1}{20}(\Delta x)^4 u_{xxxxxx}\right]_j^{n+1/2} + \cdots \tag{1.50}$$

(when $\theta = \frac{1}{2} - \frac{1}{12\mu}$), which is $O((\Delta t)^2 + (\Delta x)^4)$. Thus again we can take large time steps while maintaining accuracy and stability: for example, with $\Delta t = \Delta x = 0.1$ we find we have $\theta = \frac{1}{2} - \frac{1}{120}$ so the scheme is quite close to the Crank-Nicolson scheme.

- There are many other possible difference schemes that could be used for the heat flow equation and in Richtmyer and Morton (1967) (pp. 189–91), some fourteen schemes are tabulated.
- See numerical comparison in Fig. 1.6.

1.8 A maximum principle and convergence for $\mu(1-\theta) \leq 1/2$

If we consider what other properties a difference approximation to $u_t = u_{xx}$ should possess beyond convergence as $\Delta t, \Delta x \to 0$ (together with the necessary stability and a reasonable order of accuracy), a natural next requirement is a maximum principle. For we know mathematically (and by common experience if u represents, say, temperature) that $u(x,t)$ is bounded above and below by the extremes attained by the

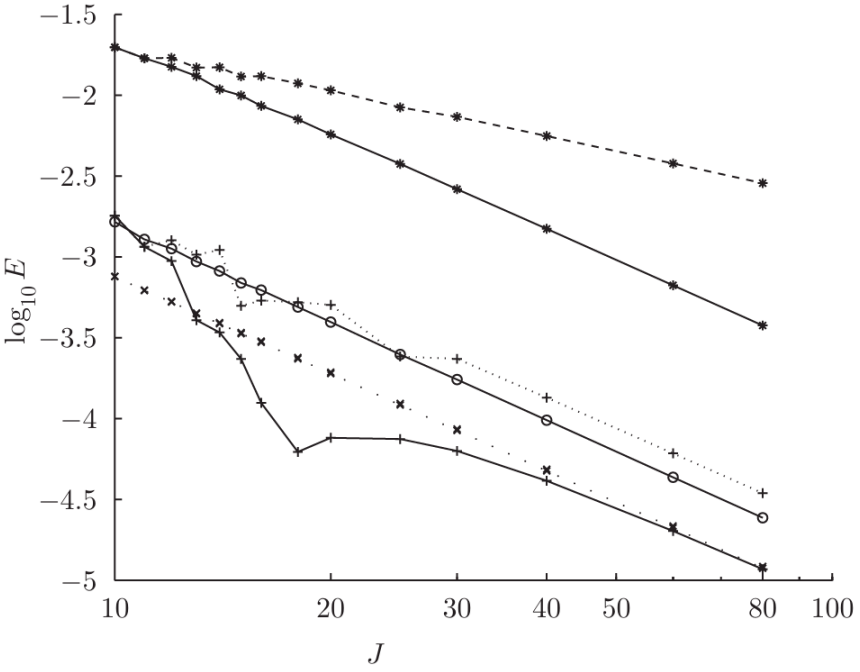

Fig. 2.7. Maximum error on $[0,1] \times [0.1,1]$ plotted against J, for various schemes.

$A:$	$\theta = 0,\ \mu = \frac{1}{2}$	—0—0—0—0
$B:$	$\theta = \frac{1}{2},\ \mu = \frac{1}{2},$	—×—×—×—
	$\theta = \frac{1}{2},\ \nu = \frac{1}{20}$	· · · × · · × · ·
$C:$	$\theta = \frac{1}{2},\ \mu = 5$	—+—+—+—
	$\theta = \frac{1}{2},\ \nu = \frac{1}{2}$	· · · + · · · + · · ·
$D:$	$\theta = 1,\ \mu = 5$	—∗—∗—∗—
	$\theta = 1,\ \nu = \frac{1}{2}$	- - - ∗ - - - ∗ - - - ∗ - - -

Figure 1.6: Various difference schemes.

initial data and the values on the boundary up to time t. Such a principle also lay behind the proof of convergence for the explicit scheme: and any engineering client for our computed results would be rather dismayed if they did not possess this property. We generalise that result by the following theorem.

Theorem 1.8.1 *The θ-method of (1.35) with $0 \le \theta \le 1$ and $\mu(1-\theta) \le \frac{1}{2}$ yields U_j^n satisfying*

$$U_{min} \le U_j^n \le U_{max} \tag{1.51}$$

where

$$U_{min} := \min U_0^m, 0 \le m \le n; U_j^0, 0 \le j \le J; U_J^m, 0 \le m \le n, \tag{1.52}$$

and

$$U_{max} := \max U_0^m, 0 \le m \le n; U_j^0, 0 \le j \le J; U_J^m, 0 \le m \le n. \tag{1.53}$$

For any refinement path which eventually satisfies this stability condition, the approximations given by (1.35) with consistent initial and Dirichlet boundary data converge uniformly on $[0,1] \times [0, t_F]$ if the initial data are smooth enough for the truncation error $T_j^{n+1/2}$ to tend to zero along the refinement path uniformly in this domain.

Proof. We write (1.35) in the form

$$(1+2\theta\mu)U_j^{n+1} = \theta\mu\left(U_{j-1}^{n+1} + U_{j+1}^{n+1}\right) + (1-\theta)\mu\left(U_{j-1}^n + U_{j+1}^n\right) + [1-2(1-\theta)\mu]U_j^n. \tag{1.54}$$

Then under the hypotheses of the theorem all the coefficients on the right are nonnegative and sum to $(1+2\theta\mu)$. Now suppose that U attains its maximum at an internal point, and this maximum is U_j^{n+1}, and let U^* be the greatest of the five values of U appearing on the right-hand side of (1.54). Then since the coefficients are nonnegative $U_j^{n+1} \le U^*$, but since this is assumed to be the maximum value, we also have $U_j^{n+1} \ge U^*$, so $U_j^{n+1} = U^*$. Indeed, the maximum value must also be attained at each neighbouring point which has a non-zero coefficient in (1.54). The same argument can then be applied at each of these points, showing that the maximum is attained at a sequence of points, until a boundary point is reached. The maximum is therefore attained at a boundary point. An identical argument shows that the minimum is also attained at a boundary point, and the first part of the proof is complete.

By the definition of truncation error (see (1.46)), the solution of the differential equation satisfies the same relation as (1.54) except for an additional term $\Delta t T_j^{n+1/2}$ on the right-hand side. Thus the error $e_j^n = U_j^n - u_j^n$ is determined from the relations

$$(1+2\theta\mu)e_j^{n+1} = \theta\mu\left(e_{j-1}^{n+1} + e_{j+1}^{n+1}\right) + (1-\theta)\mu\left(e_{j-1}^n + e_{j+1}^n\right) + [1-2(1-\theta)\mu]e_j^n - \Delta t T_j^{n+1/2} \tag{1.55}$$

for $j = 1, 2, \ldots, J-1$ and $n = 0, 1, \ldots$ together with initial and boundary conditions. Suppose first of all that these latter are zero because $U_j^0 = u_j^0$, $U_0^m = u_0^m$ and $U_j^m = u_j^m$. Then we define

$$E^n := \max_{0 \le j \le J} |e_j^n|, \quad T^{n+1/2} := \max_{1 \le j \le J-1} \left|T_j^{n+1/2}\right|. \tag{1.56}$$

Because of the nonnegative coefficients, it follows that

$$(1+2\theta\mu)E^{n+1} \le 2\theta\mu E^{n+1} + E^n + \Delta t T^{n+1/2} \tag{1.57}$$

and hence that

$$E^{n+1} \leq E^n + \Delta t T^{n+1/2} \tag{1.58}$$

so that, since $E^0 = 0$,

$$\begin{aligned} E^n &\leq \Delta t \sum_0^{n-1} T^{m+1/2}, \\ &\leq n\Delta t \max_m T^{m+1/2} \end{aligned} \tag{1.59}$$

and this tends to zero along the refinement path under the assumed hypotheses.

So far we have assumed that numerical errors arise from the truncation errors of the finite difference approximations, but that the boundary values are used exactly. Suppose now that there are errors in the initial and boundary values of U_j^n and let us denote them by e_j^0, e_0^m and e_j^m with $0 \leq j \leq J$ and $0 \leq m \leq N$, say. Then the errors e_j^n satisfy the recurrence relation (1.55) with initial and boundary values

$$e_j^0 = e_j^0, \quad j = 0, 1, \ldots, n,$$

$$e_0^m = e_0^m, \quad e_j^m = e_j^m, \quad 0 \leq m \leq N.$$

Then (by Duhamel's principle) e_j^N can be written as the sum of two terms. The first term satisfies (1.55) with zero initial and boundary values; this term is bounded by (1.59). The second term satisfies the homogeneous form of (1.55), with the term in T omitted, and with the given non-zero initial and boundary values. By the maximum principle this term must lie between the maximum and minimum values of these initial and boundary values. Thus the error of the numerical solution will tend to zero along the refinement path, as required, provided that the initial and boundary values are **consistent**; that is, the errors in the initial and boundary values also tend to zero along the refinement path.

■

Remark 1.8.2 *The condition for this theorem, $\mu(1-\theta) \leq \frac{1}{2}$, is very much more restrictive than that needed in the Fourier analysis of stability, $\mu(1-2\theta) \leq \frac{1}{2}$; for example, the Crank-Nicolson scheme always satisfies the stability condition, but only if $\mu \leq 1$ does it satisfy the condition given for a maximum principle, which in the theorem is then used to deduce stability and convergence.*

Thus the maximum principle analysis can be viewed as an alternative means of obtaining stability conditions. It has the advantage over Fourier analysis that it is easily extended to apply to problems with variable coefficients; but, as we see above, it is easy to derive only sufficient stability conditions (maximum principle $\Rightarrow$ stability).

Example 1.8.3 *These points are illustrated in Fig. 1.7. Here the model problem is solved by the Crank-Nicolson scheme. The boundary conditions specify that the solution is zero at each end of the range, and the initial condition gives the values of U_j^0 to be zero except at the mid-point; the value at the mid-point is unity. This corresponds to a function with a sharp spike at $x = \frac{1}{2}$.*

In the case $\mu = 2$ the maximum principle does not hold, and we see that at the first time level the numerical solution becomes negative at the mid-point. This would normally be regarded as unacceptable.

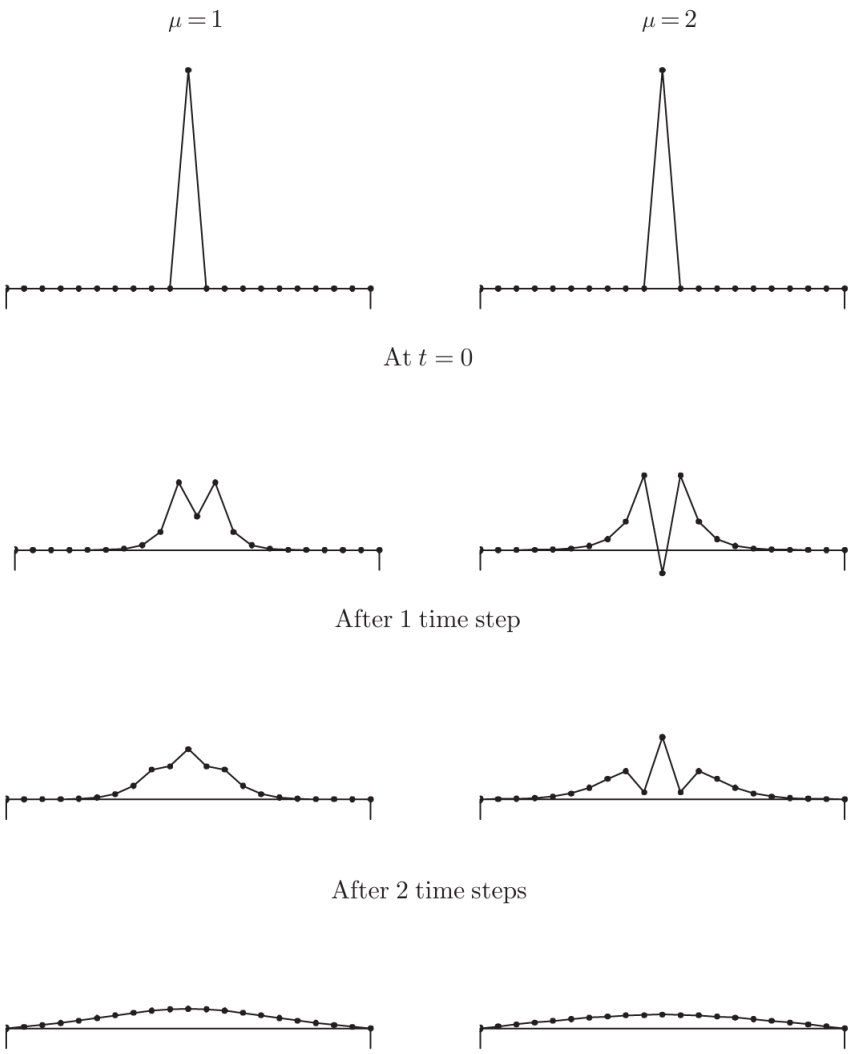

Fig. 2.9. The Crank–Nicolson method applied to the heat equation where the initial distribution has a sharp spike at the mid-point; $J = 20$, $\Delta x = 0.05$.

Figure 1.7: Maximum principle.

When $\mu = 1$ the maximum principle holds, and the numerical values all lie between 0 and 1, as required. However, at the first time level the numerical solution shows two peaks, one each side of the mid-point; the exact solution of the problem (Gaussian kernel) will have only a single maximum for all t. These results correspond to a rather extreme case, and the unacceptable behaviour only persists for a few time steps; thereafter the solution becomes very smooth in each case. However, they show that in a situation where we require to model some sort of rapid variation in the solution we shall need to use a value of μ somewhat smaller than the stability limit.

1.9 three- or more-time-level scheme

Crank-Nicolson

Second order backward differentiation formula (BDF2)

1.10 More general boundary conditions

1.10.1 scheme 1

Let us now consider a more general model problem by introducing a derivative boundary condition at $x = 0$, of the form

$$\frac{\partial u}{\partial x} = \alpha(t)u + g(t), \quad \alpha(t) \geq 0. \tag{1.60}$$

By using a forward space difference for the derivative, we can approximate this by

$$\frac{U_1^n - U_0^n}{\Delta x} = \alpha^n U_0^n + g^n \tag{1.61}$$

and use this to give the boundary value U_0^n in the form

$$U_0^n = \beta^n U_1^n - \beta^n g^n \Delta x, \tag{1.62}$$

where

$$\beta^n = \frac{1}{1 + \alpha^n \Delta x}. \tag{1.63}$$

Then we can apply the θ-method (1.35) in just the same way as for Dirichlet boundary conditions.

stability: We can use (1.62) to eliminate U_0^n; the second difference at the first interior point then has the form

$$\delta_x^2 U_1^n = U_2^n - 2U_1^n + U_0^n = U_2^n - (2 - \beta^n)U_1^n - \beta^n g^n \Delta x. \tag{1.64}$$

Thus with the usual definition of truncation error, after some manipulation, the global error can be shown to satisfy, instead of (1.55), the new relation

$$[1 + \theta\mu(2 - \beta^{n+1})]e_1^{n+1} = [1 - (1 - \theta)\mu(2 - \beta^n)]e_1^n + \theta\mu e_2^{n+1} + (1 - \theta)\mu e_2^n - \Delta t T_1^{n+1/2}. \tag{1.65}$$

This equation is different from that at other mesh points, which precludes our using Fourier analysis to analyse the system. But the maximum principle arguments of the preceding section can be used: we see first

that if $\mu(1-\theta) \le \frac{1}{2}$ all the coefficients in (1.65) are nonnegative for any nonnegative value of α^n; and the sum of the coefficients on the right is no greater than that on the left if

$$\theta(1-\beta^{n+1}) \ge -(1-\theta)(1-\beta^n), \tag{1.66}$$

which again is always satisfied if $\alpha(t) \ge 0$. Hence we can deduce the bound (1.59) for the global error in terms of the truncation error as before. The importance of the assumption $\alpha(t) \ge 0$ is clear in these arguments.

consistency: It remains to estimate the truncation error $T_1^{n+1/2}$. Let us consider only the explicit case $\theta = 0$, for which we expand around the first interior point. Suppose we straightforwardly regard (1.61) as applying the boundary condition at $(0, t_n)$ and expand about this point for the exact solution to obtain

$$\frac{u_1^n - u_0^n}{\Delta x} - \alpha^n u_0^n - g^n = \left[\frac{1}{2}\Delta x u_{xx} + \frac{1}{6}(\Delta x)^2 u_{xxx} + \cdots\right]_0^n. \tag{1.67}$$

Then we write the truncation error in the following form, in which an appropriate multiple of the approximation (1.61) to the boundary condition is added to the difference equation in order to cancel the terms in u_0^n,

$$\begin{aligned} T_1^{n+1/2} &= \frac{u_1^{n+1} - u_1^n}{\Delta t} - \frac{\delta_x^2 u_1^n}{(\Delta x)^2} - \frac{\beta^n}{\Delta x}\left[\frac{u_1^n - u_0^n}{\Delta x} - \alpha^n u_0^n - g^n\right] \\ &= \left[\frac{1}{2}\Delta t u_{tt} - \frac{1}{12}(\Delta x)^2 u_{xxxx} + \cdots\right]_1^n - \beta^n\left[\frac{1}{2}u_{xx} + \cdots\right]_0^n, \end{aligned} \tag{1.68}$$

to obtain

$$T_1^{n+1/2} \approx -\frac{1}{2}\beta^n u_{xx}. \tag{1.69}$$

This does not tend to zero as the mesh size tends to zero and, although we could rescue our convergence proof by a more refined analysis, we shall not undertake this here. (check if here is consistency and convergence?)

1.10.2 scheme 2

However, a minor change can remedy the problem. We choose a new grid of points, which are still equally spaced, but with the boundary point $x = 0$ half-way between the first two grid points. The other boundary, $x = 1$, remains at the last grid point as before. We now replace the approximation to the boundary condition by the more accurate version

$$\frac{U_1^n - U_0^n}{\Delta x} = \frac{1}{2}\alpha^n(U_0^n + U_1^n) + g^n, \tag{1.70}$$

$$U_0^n = \frac{1 - \frac{1}{2}\alpha^n \Delta x}{1 + \frac{1}{2}\alpha^n \Delta x} U_1^n - \frac{\Delta x}{1 + \frac{1}{2}\alpha^n \Delta x} g^n. \tag{1.71}$$

Then (1.67) is replaced by an expansion about $j = \frac{1}{2}$, giving

$$\frac{u_1^n - u_0^n}{\Delta x} - \frac{1}{2}\alpha^n(u_0^n + u_1^n) - g^n = \left[\frac{1}{24}(\Delta x)^2 u_{xxx} - \frac{1}{8}\alpha^n(\Delta x)^2 u_{xx} + \cdots\right]_{1/2}^n \tag{1.72}$$

and hence

$$T^{n+1/2} = \left[\frac{1}{2}\Delta t u_{tt} - \frac{1}{12}(\Delta x)^2 u_{xxxx} + \cdots\right]_1^n - \frac{1}{1 + \frac{1}{2}\alpha^n \Delta x}\left[\frac{1}{24}\Delta x(uxxx - 3\alpha^n u_{xx}) + \cdots\right]_0^n = O(\Delta x) \tag{1.73}$$

Only minor modifications are necessary to (1.65) and the proof of convergence is straightforward. **Indeed, as we shall show in Chapter 6 where a sharper error analysis based on the maximum principle is presented, the error remains $O((\Delta x)^2)$ despite this $O(\Delta x)$ truncation error near the boundary.**

1.10.3 scheme 3

An alternative, and more widely used, approach is to keep the first grid point at $x = 0$ but to introduce a fictitious value U^n_{-1} outside the domain so that we can use central differences to write

$$\frac{U^n_1 - U^n_{-1}}{2\Delta x} = \alpha^n U^n_0 + g^n. \tag{1.74}$$

Then the usual difference approximation is also applied at $x = 0$ so that U^n_{-1} can be eliminated. That is, for the θ-method we take

$$\begin{aligned} \frac{U^{n+1}_0 - U^n_0}{\Delta t} &- \frac{\delta^2_x}{(\Delta x)^2}\left[\theta U^{n+1}_0 + (1-\theta)U^n_0\right] \\ &- \frac{2\theta}{\Delta x}\left[\frac{U^{n+1}_1 - U^{n+1}_{-1}}{2\Delta x} - \alpha^{n+1}U^{n+1}_0 - g^{n+1}\right] \\ &- \frac{2(1-\theta)}{\Delta x}\left[\frac{U^n_1 - U^n_{-1}}{2\Delta x} - \alpha^n U^n_0 - g^n\right] = 0. \end{aligned} \tag{1.75}$$

Clearly for the truncation error we pick up terms like

$$\frac{2\theta}{\Delta x}\left[\frac{u^{n+1}_1 - u^{n+1}_{-1}}{2\Delta x} - \alpha^{n+1}u^{n+1}_0 - g^{n+1}\right] = \theta\left[\frac{1}{3}\Delta x u_{xxx}\right]^{n+1}_0 + \cdots \tag{1.76}$$

to add to the usual truncation error terms. If we rewrite (1.75) in the form

$$[1 + 2\theta\mu(1 + \alpha^{n+1}\Delta x)]U^{n+1}_0 = [1 - 2(1-\theta)\mu(1 + \alpha^n\Delta x)]U^n_0 + 2\theta\mu U^{n+1}_1 - 2\mu\Delta x\left[\theta g^{n+1} + (1-\theta)g^n\right] \tag{1.77}$$

we also see that the error analysis based on a maximum principle still holds with only the slight strengthening of condition needed in Maximum Principle Theorem to

$$\mu(1-\theta)(1 + \alpha^n\Delta x) \le \frac{1}{2}. \tag{1.78}$$

See numerical comparison for three schemes in Fig. 1.8.

1.11 More general linear problems

1.11.1 variable-coefficient heat equation

explicit scheme

First of all, consider the problem

$$\frac{\partial u}{\partial t} = b(x,t)\frac{\partial^2 u}{\partial x^2}, \tag{1.79}$$

where the function $b(x,t)$ is, as usual, assumed to be strictly positive. Then the explicit scheme (1.3) is extended in an obvious way to give

$$U^{n+1}_j = U^n_j + \frac{\Delta t}{(\Delta x)^2}b^n_j(U^n_{j+1} - 2U^n_j + U^n_{j-1}), \tag{1.80}$$

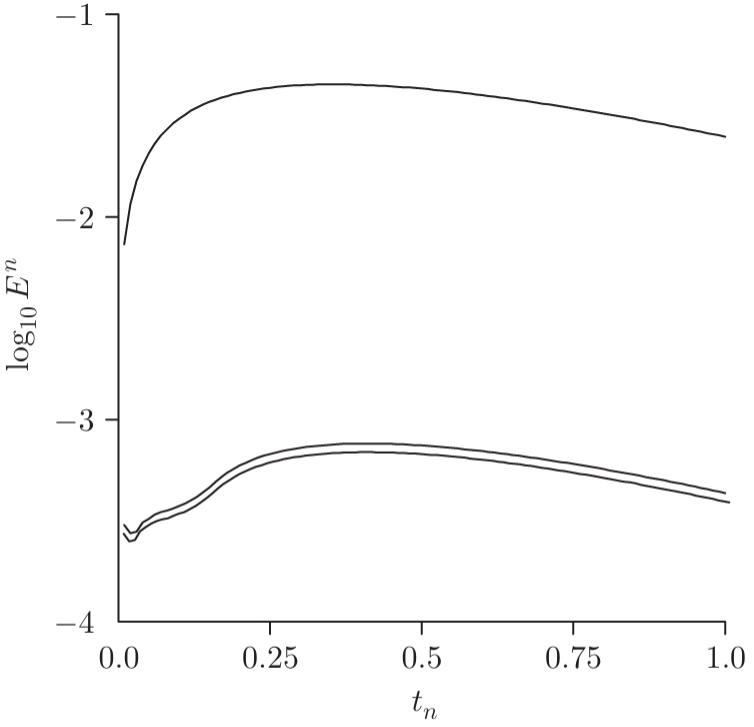

Fig. 2.10. The effect of a Neumann boundary condition approximation on the error for the Crank–Nicolson scheme with $J = 10$, $\Delta x = 0.1$; the top curve is for (2.103) and the lower two for (2.114) and (2.110).

Figure 1.8: Neumann boundary condition.

where $b_j^n = b(x_j, t_n)$. The practical implementation of this scheme is just as easy as before, and the analysis of the error is hardly altered. The same expansion in Taylor series leads, as for (1.3), to the expression

$$T(x,t) = \frac{1}{2}\Delta t u_{tt} - \frac{1}{12}b(x,t)(\Delta x)^2 u_{xxxx} + \cdots . \tag{1.81}$$

The analysis leading to (1.14) still applies, but the stability condition has to be replaced by

$$\frac{\Delta t}{(\Delta x)^2}b(x,t) \le \frac{1}{2} \tag{1.82}$$

for all values of x and t in the region. The final error bound becomes

$$E^n \le \frac{1}{2}\Delta t \left[M_{tt} + \frac{B(\Delta x)^2}{6\Delta t}M_{xxxx}\right] t_F \tag{1.83}$$

where B is a uniform upper bound for $b(x,t)$ in the region $[0,1]\times[0,t_F]$.

θ-method with b at $t_{n+1/2}$

The θ-method can be applied to this more general problem in several slightly different ways. Evidently equation (1.35) can be generalised to

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{(\Delta x)^2}b^* \left[\theta\delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n\right], \tag{1.84}$$

but it is not obvious what is the best value to use for b^*. In our previous analysis of the truncation error of this scheme we expanded in Taylor series about the centre point $(x_j, t_{n+1/2})$. This suggests the choice

$$b^* = b_j^{n+1/2}; \tag{1.85}$$

and in fact it is easy to see that with this choice our former expansion of the truncation error is unaltered, except for the inclusion of the extra factor b in (1.46), which becomes

$$\begin{aligned} T_j^{n+1/2} = \quad & [(\frac{1}{2}-\theta)\Delta t u_{xxt} - \frac{b}{12}(\Delta x)^2 u_{xxxx} + \frac{1}{24}(\Delta t)^2 u_{ttt} \\ & -\frac{b}{8}(\Delta t)^2 u_{xxtt} + \frac{1}{12}(\frac{1}{2}-\theta)\Delta t(\Delta x)^2 u_{xxxxt} \\ & -\frac{2b}{6!}(\Delta x)^4 u_{xxxxxx} + \cdots]_j^{n+1/2}. \end{aligned} \tag{1.86}$$

The proof of convergence by means of a maximum principle is also unaltered, except that the stability condition now requires that

$$\frac{\Delta t}{(\Delta x)^2}(1-\theta)b(x,t) \le \frac{1}{2} \tag{1.87}$$

for all points (x,t) in the region considered.

θ-method with $b^* = (b^{n+1} + b^n)/2$

This choice of b^* requires the computation of $b(x,t)$ for values of t half-way between time steps. This may be awkward in some problems, and an obvious alternative is to use

$$b^* = \frac{1}{2}(b_j^{n+1} + b_j^n). \tag{1.88}$$

Now we need another Taylor expansion, about the centre point, giving

$$b^* = [b + \frac{1}{4}(\Delta t)^2 b_{tt} + \cdots]_j^{n+1/2} \tag{1.89}$$

which will lead to an additional higher order term, involving b_{tt}, appearing in the expansion of the truncation error.

1.11.2 the most general form of the linear parabolic equation

central difference scheme for advection

The most general form of the linear parabolic equation is

$$\frac{\partial u}{\partial t} = b(x,t)\frac{\partial^2 u}{\partial x^2} - a(x,t)\frac{\partial u}{\partial x} + c(x,t)u + d(x,t), \tag{1.90}$$

where as before $b(x,t)$ is assumed to be always positive. The notation used here is chosen to match that used in later chapters. In particular the negative sign in front of $a(x,t)$ is convenient but unimportant, since $a(x,t)$ may take either sign; only $b(x,t)$ is required to be positive. We can easily construct an explicit scheme for this equation; only the term in $\partial u/\partial x$ needs any new consideration. As we have used the central difference approximation for the second derivative, it is natural to use the central difference approximation for the first derivative, leading to the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{b_j^n}{(\Delta x)^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) - \frac{a_j^n}{2\Delta x}(U_{j+1}^n - U_{j-1}^n) + c_j^n U_j^n + d_j^n. \tag{1.91}$$

The calculation of the leading terms in the truncation error is straightforward, and is left as an exercise. However, a new difficulty arises in the analysis of the behaviour of the error e_j^n. Just as in the analysis of the simpler problem, which led to (1.12), we find that

$$e_j^{n+1} = e_j^n + \mu_j^n(e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \frac{1}{2}\nu_j^n(e_{j+1}^n - e_{j-1}^n) + \Delta t c_j^n e_j^n - \Delta t T_j^n \tag{1.92}$$

$$= (1 - 2\mu_j^n + \Delta t c_j^n)e_j^n + (\mu_j^n - \frac{1}{2}\nu_j^n)e_{j+1}^n + (\mu_j^n + \frac{1}{2}\nu_j^n)e_{j-1}^n - \Delta t T_j^n, \tag{1.93}$$

where we have written

$$\mu_j^n = \frac{\Delta t}{(\Delta x)^2}b_j^n, \quad \nu_j^n = \frac{\Delta t}{\Delta x}a_j^n. \tag{1.94}$$

In order to go on to obtain similar bounds for e_j^n as before, we need to ensure that the coefficients of the three terms in e^n on the right of this equation are all nonnegative and have a sum no greater than unity. We always assume that the function $b(x,t)$ is strictly positive, but we cannot in general assume anything about the sign of $a(x,t)$. We are therefore led to the conditions:

$$\frac{1}{2}|\nu_j^n| \le \mu_j^n, \tag{1.95}$$

$$2\mu_j^n - \Delta t c_j^n \le 1, \tag{1.96}$$

as well as $c_j^n \le 0$. The second of these conditions is only slightly more restrictive than in the simpler case, because of the condition $c_j^n \le 0$; indeed, if we had $0 \le c(x,t) \le C$ condition (1.96) would represent a slight

relaxation of the condition on μ, but then one can only establish $E^{n+1} \leq (1 + C\Delta t)E^n + T\Delta t$. However, the first condition is much more serious. If we replace ν and μ by their expressions in terms of Δt and Δx this becomes

$$\Delta x \leq \frac{2b_j^n}{|a_j^n|}, \quad \text{or} \quad \frac{|a_j^n|\Delta x}{b_j^n} \leq 2, \tag{1.97}$$

and this condition must hold for all values of n and j. We therefore have a restriction on the size of Δx, which also implies a restriction on the size of Δt.

upwind difference scheme for advection

In many practical problems the function $b(x,t)$ may be very small compared with $a(x,t)$. This will happen, for example, in the flow of most fluids, which have a very small viscosity. In such situations a key dimensionless parameter is the Péclet number UL/ν, where U is a velocity, L a length scale and ν the viscosity. These are close to what are known as singular perturbation problems, and cannot easily be solved by this explicit, central difference method: for (1.97) imposes a limit of 2 on a mesh Péclet number in which the length scale is the mesh length. Suppose, for example, that $b = 0.001, a = 1, c = 0$. Then our conditions require that $\Delta x \leq 0.002$, and therefore that $\Delta t \leq 0.000002$. We thus need at least 500 mesh points in the x-direction, and an enormous number of time steps to reach any sensible time t_F.

A simple way of avoiding this problem is to use forward or backward differences for the first derivative term, instead of the central difference. Suppose, for example, it is known that $a(x,t) \geq 0$ and $c(x,t) = 0$. We then use the backward difference, and our difference formula becomes

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{b_j^n}{(\Delta x)^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) - \frac{a_j^n}{\Delta x}(U_j^n - U_{j-1}^n) + c_j^n U_j^n + d_j^n, \tag{1.98}$$

which leads to

$$e_j^{n+1} = e_j^n + \mu_j^n(e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \nu_j^n(e_j^n - e_{j-1}^n) - \Delta t T_j^n \tag{1.99}$$

$$= (1 - 2\mu_j^n - \nu_j^n)e_j^n + \mu_j^n e_{j+1}^n + (\mu_j^n + \nu_j^n)e_{j-1}^n - \Delta t T_j^n. \tag{1.100}$$

In order to ensure that all the coefficients on the right of this equation are nonnegative, we now need only

$$2\mu_j^n + \nu_j^n \leq 1. \tag{1.101}$$

This requires a more severe restriction on the size of the time step when $a \neq 0$, but no restriction on the size of Δx.

If the function $a(x,t)$ changes sign, we can use the backward difference where a is positive, and the forward difference where it is negative; this idea is known as **upwind differencing**. Unfortunately we have to pay a price for this lifting of the restriction needed to ensure a maximum principle. The truncation error is now of lower order: the forward difference introduces an error of order Δx, instead of the order $(\Delta x)^2$ given by the central difference. However, we shall discuss this issue in the chapter on hyperbolic equations.

1.11.3 conservation form

A general parabolic equation may also often appear in the self-adjoint form

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(p(x,t)\frac{\partial u}{\partial x}\right) \tag{1.102}$$

where, as usual, we assume that the function $p(x,t)$ is strictly positive. It is possible to write this equation in the form just considered, as

$$\frac{\partial u}{\partial t} = p\frac{\partial^2 u}{\partial x^2} + \frac{\partial p}{\partial x}\frac{\partial u}{\partial x}, \tag{1.103}$$

but it is usually better to construct a difference approximation to the equation in its original form. We can write

$$\left[p\frac{\partial u}{\partial x}\right]^n_{j+1/2} \approx p^n_{j+1/2}\left(\frac{u^n_{j+1} - u^n_j}{\Delta x}\right), \tag{1.104}$$

and a similar approximation with j replaced by $j-1$ throughout. If we subtract these two, and divide by Δx, we obtain an approximation to the right-hand side of the equation, giving the explicit difference scheme

$$\frac{U^{n+1}_j - U^n_j}{\Delta t} = \frac{1}{(\Delta x)^2}\left[p^n_{j+1/2}(U^n_{j+1} - U^n_j) - p^n_{j-1/2}(U^n_j - U^n_{j-1})\right]. \tag{1.105}$$

We will write

$$\mu' = \frac{\Delta t}{(\Delta x)^2} \tag{1.106}$$

which gives in explicit form

$$U^{n+1}_j = \left(1 - \mu'(p^n_{j+1/2} + p^n_{j-1/2})\right)U^n_j + \mu' p^n_{j+1/2}U^n_{j+1} + \mu' p^n_{j-1/2}U^n_{j-1}. \tag{1.107}$$

This shows that the form of error analysis which we have used before will again apply here, with each of the coefficients on the right-hand side being nonnegative provided that

$$\mu' P \leq \frac{1}{2}, \tag{1.108}$$

where P is an upper bound for the function $p(x,t)$ in the region. So this scheme gives just the sort of time step restriction which we should expect, without any restriction on the size of Δx.

The same type of difference approximation can be applied to give an obvious generalisation of the θ-method. The details are left as an exercise, as is the calculation of the leading terms of the truncation error (see Exercises 7 and 8).

1.12 Nonlinear problems

We shall just consider one example, the equation

$$u_t = b(u)u_{xx} \tag{1.109}$$

where the coefficient $b(u)$ depends on the solution u only and must be assumed strictly positive for all u. This simplification is really only for ease of notation; it is not much more difficult to treat the case in which b is a function of x and t as well as of u.

The explicit method is little affected; it becomes, in the same notation as before,

$$U^{n+1}_j = U^n_j + \mu' b(U^n_j)\left(U^n_{j+1} - 2U^n_j + U^n_{j-1}\right). \tag{1.110}$$

The actual calculation is no more difficult than before, the only extra work being the computation of the function $b(U_j^n)$. The truncation error also has exactly the same form as before and the conditions for the values U_j^n to satisfy a maximum principle are unchanged. However, the analysis of the behaviour of the global error e_j^n is more difficult, as it propagates in a nonlinear way as n increases.

Writing u_j^n for the value of the exact solution $u(x_j, t_n)$ we know that U_j^n and u_j^n satisfy the respective equations

$$U_j^{n+1} = U_j^n + \mu' b(U_j^n)(U_{j+1}^n - 2U_j^n + U_{j-1}^n), \tag{1.111}$$

$$u_j^{n+1} = u_j^n + \mu' b(u_j^n)(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t T_j^n, \tag{1.112}$$

where T_j^n is the truncation error. But we cannot simply subtract these equations to obtain a relation for e_j^n, since the two coefficients $b(\cdot)$ are different. However we can first write

$$\begin{aligned} b(u_j^n) &= b(U_j^n) + (u_j^n - U_j^n)\frac{\partial b}{\partial u}(\eta) &(1.113)\\ &= b(U_j^n) - e_j^n q_j^n &(1.114)\end{aligned}$$

where

$$q_j^n = \frac{\partial b}{\partial u}(\eta) \tag{1.115}$$

and η is some number between U_j^n and u_j^n.

We can now subtract (1.112) from (1.111), and obtain

$$e_j^{n+1} = e_j^n + \mu' b(U_j^n)(e_{j+1}^n - 2e_j^n + e_{j-1}^n) + \mu' e_j^n q_j^n (u_{j+1}^n - 2u_j^n + u_{j-1}^n) - \Delta t T_j^n. \tag{1.116}$$

The coefficients of $e_{j-1}^n, e_j^n, e_{j+1}^n$ arising from the first two terms on the right are now nonnegative provided that

$$\Delta t[\max b(U_j^n)] \le \frac{1}{2}(\Delta x)^2. \tag{1.117}$$

This is our new stability condition, and the condition for the approximation to satisfy a maximum principle; in general it will need to be checked (and Δt adjusted) at each time step. However, assuming that we can use a constant step Δt which satisfies (1.117) for all j and n, and that we have bounds

$$|u_{j+1} - 2u_j^n + u_{j-1}^n| \le M_{xx}(\Delta x)^2, \quad |q_j^n| \le K, \tag{1.118}$$

we can write

$$E^{n+1} \le [1 + KM_{xx}\Delta t]E^n + \Delta t T \tag{1.119}$$

in our previous notation. Moreover,

$$(1 + KM_{xx}\Delta t)^n \le e^{KM_{xx}n\Delta t} \le e^{KM_{xx}tF} \tag{1.120}$$

and this allows a global error bound to be obtained in terms of T.

However, although the stability condition (1.117) is not much stronger than that for the linear problem, the error bound is much worse unless the a priori bounds on $|\partial b/\partial u|$ and $|u_{xx}|$ are very small. Furthermore, our example (1.109) is rather special; equally common would be the case

$$u_t = (b(u)u_x)_x, \tag{1.121}$$

and that gives an extra term $(\partial b/\partial u)(u_x)^2$ which can make very great changes to the problem and its analysis.

To summarise, then, **the actual application of our explicit scheme to nonlinear problems gives little difficulty**. Indeed the main practical use of numerical methods for partial differential equations is for nonlinear problems, where alternative methods break down. **Even our implicit methods are not very much more difficult to use**; there is just a system of nonlinear equations to solve with a good first approximation given from the previous time level. **However, the analysis of the convergence and stability behaviour of these schemes is very much more difficult than for the linear case**.

1.13 The explicit method in a rectilinear box for 2D problems

1.13.1 scheme

The natural generalisation of the one-dimensional model problem in two dimensions is the equation

$$\begin{aligned} u_t &= b\nabla^2 u \quad (b>0) \\ &= b[u_{xx}+u_{yy}], \end{aligned} \tag{1.122}$$

where b is a positive constant. We shall consider the rectangular domain in the (x,y)-plane

$$0<x<X, \quad 0<y<Y,$$

and assume Dirichlet boundary conditions, so that $u(x,y,t)$ is given at all points on the rectangular boundary, for all positive values of t. In addition, of course, an initial condition is given, so that $u(x,y,0)$ is given on the rectangular region. The region is covered with a uniform rectangular grid of points, with a spacing Δx in the x-direction and Δy in the y-direction, where

$$\Delta x = \frac{X}{J_x}, \quad \Delta y = \frac{Y}{J_y}, \quad J_x, J_y \in \mathbb{Z}.$$

The approximate solution is then denoted by

$$U^n_{r,s} \approx u(x_r, y_s, t_n), \quad r=0,1,\ldots,J_x, \quad s=0,1,\ldots,J_y. \tag{1.123}$$

The simplest explicit difference scheme is the natural extension of the explicit scheme in one dimension, and is given by

$$\frac{U^{n+1}-U^n}{\Delta t} = b\left[\frac{\delta_x^2 U^n}{(\Delta x)^2} + \frac{\delta_y^2 U^n}{(\Delta y)^2}\right]. \tag{1.124}$$

Here we have omitted the subscripts (r,s) throughout, and used the notation of (2.28) for the second order central differences in the x- and y-directions. This is an explicit scheme, since there is only one unknown value $U^{n+1}_{r,s}$ on the new time level. This unknown value is calculated from five neighbouring values on the previous time level,

$$U^n_{r,s}, \quad U^n_{r+1,s}, \quad U^n_{r-1,s}, \quad U^n_{r,s+1} \text{ and } U^n_{r,s-1}. \tag{1.125}$$

Most of the analysis of this scheme in one dimension is easily extended to the two-dimensional case; the details are left as an exercise. The truncation error is

$$T(x,t) = \frac{1}{2}\Delta t u_{tt} - \frac{1}{12} b\left[(\Delta x)^2 u_{xxxx} + (\Delta y)^2 u_{yyyy}\right] + \cdots, \tag{1.126}$$

from which a bound on the truncation error can be obtained in terms of bounds on the derivatives of u, written in the same notation as before as M_{tt}, M_{xxxx} and M_{yyyy}. The proof of convergence follows in a similar way, leading to

$$E^n \leq \left[\frac{1}{2}\Delta t M_{tt} + \frac{1}{12}b\left((\Delta x)^2 M_{xxxx} + (\Delta y)^2 M_{yyyy}\right)\right] t_F, \tag{1.127}$$

provided that the mesh sizes satisfy the condition

$$\mu_x + \mu_y \leq \frac{1}{2}, \tag{1.128}$$

where

$$\mu_x = \frac{b\Delta t}{(\Delta x)^2}, \quad \mu_y = \frac{b\Delta t}{(\Delta y)^2}. \tag{1.129}$$

1.13.2 stability

The stability of the scheme can also be analysed by Fourier series, assuming that b is constant, and ignoring the effect of boundary conditions, for which a justification will be given in Chapter 5. We construct solutions of the difference equation of the form

$$U^n \sim (\lambda)^n \exp\left(i[k_x x + k_y y]\right) \tag{1.130}$$

to obtain the amplification factor λ as

$$\lambda \equiv \lambda(\mathbf{k}) = 1 - 4\left[\mu_x \sin^2 \frac{1}{2}k_x\Delta x + \mu_y \sin^2 \frac{1}{2}k_y\Delta y\right] \tag{1.131}$$

where $\mathbf{k} = (k_x, k_y)$. Just as in the one-dimensional problem it is clear that the stability condition is

$$\mu_x + \mu_y \leq \frac{1}{2}. \tag{1.132}$$

The sufficiency of this condition in establishing $|\lambda(\mathbf{k})| \leq 1$ follows just as in the one-dimensional case: its necessity follows from the fact that all the components of $\mathbf{k}$ can be chosen independently to give the worst mode, for which $k_x\Delta x = k_y\Delta y = \pi$.

Calculating the approximation from (1.124) is clearly just as easy as in the one-dimensional case. **However the stability condition is more restrictive**; and when b is variable we need to apply the condition (1.132) at each point so that any local peak in b will cut down the time step that can be used. Thus this simple explicit scheme is generally impractical and we must introduce some implicitness to avoid, or relax, the stability restriction. This is even more true in three dimensions, to which all of the above is readily extended.

1.14 An ADI method in two dimensions

1.14.1 Crank–Nicolson method

The natural extension of our study of the one-dimensional problem would now be to suggest an extension of the θ-method. In particular, the Crank-Nicolson method becomes

$$(1 - \frac{1}{2}\mu_x\delta_x^2 - \frac{1}{2}\mu_y\delta_y^2)U^{n+1} = (1 + \frac{1}{2}\mu_x\delta_x^2 + \frac{1}{2}\mu_y\delta_y^2)U^n. \tag{1.133}$$

In one dimension the great advantage of this type of method was the lifting of the stability restriction with little extra computational labour. **In two or more dimensions this is no longer true; the method is still stable without restriction on the time step, but the extra labour involved is now very considerable.** We have to solve a system of $(J_x-1)(J_y-1)$ linear equations for the unknown values $U_{r,s}^{n+1}$. The equations have a regular structure, each equation involving at most five unknowns; the matrix of the system consists very largely of zeros, but it does not have tridiagonal form; moreover there is no way of permuting the rows and columns so that the non-zero elements form a narrow band. The solution of such a system of equations is by no means out of the question, as we shall see when we come to elliptic equations in Chapter 6, but it requires so much extra sophistication that it suggests we should look for other numerical schemes for parabolic equations in two dimensions.

1.14.2 implicit method in one dimension

Since an implicit method in one dimension can be very efficient, it is natural to look for methods which are implicit in one dimension, but not both. Consider, for example, the scheme

$$(1-\frac{1}{2}\mu_x\delta_x^2)U^{n+1}=(1+\frac{1}{2}\mu_x\delta_x^2+\mu_y\delta_y^2)U^n. \tag{1.134}$$

If we examine the equations of the system corresponding to a particular row of mesh points or value of s, we see that they form a tridiagonal system of order J_x-1, as they do not involve any unknowns with different values of s. The complete system thus involves a set of J_y-1 tridiagonal systems, each of which can be solved very efficiently by the Thomas algorithm of Section 2.9. This scheme will require roughly three times as much computational labour as the explicit scheme. Unfortunately, although the stability of the scheme has been improved, there is still a restriction. We easily find, still assuming b is constant, that the amplification factor is

$$\lambda(\mathbf{k})=\frac{1-2\mu_x\sin^2\frac{1}{2}k_x\Delta x-4\mu_y\sin^2\frac{1}{2}k_y\Delta y}{1+2\mu_x\sin^2\frac{1}{2}k_x\Delta x} \tag{1.135}$$

and the scheme will be unstable if $\mu_y>\frac{1}{2}$. As might be expected, there is no restriction on μ_x.

1.14.3 alternative-direction implicit method

Successful methods can be obtained by combining two such schemes, each of which is implicit in one direction. The first such scheme was proposed by Peaceman and Rachford in 1955[1] and used by them in oil reservoir modelling. We begin by writing a modification of the Crank-Nicolson scheme in the form

$$(1-\frac{1}{2}\mu_x\delta_x^2)(1-\frac{1}{2}\mu_y\delta_y^2)U^{n+1}=(1+\frac{1}{2}\mu_x\delta_x^2)(1+\frac{1}{2}\mu_y\delta_y^2)U^n. \tag{1.136}$$

Noticing that we can expand the product of the difference operators as

$$(1+\frac{1}{2}\mu_x\delta_x^2)(1+\frac{1}{2}\mu_y\delta_y^2)=(1+\frac{1}{2}\mu_x\delta_x^2+\frac{1}{2}\mu_y\delta_y^2+\frac{1}{4}\mu_x\mu_y\delta_x^2\delta_y^2), \tag{1.137}$$

we see that (1.136) is not exactly the same as the Crank-Nicolson scheme, but introduces extra terms which are of similar order to some in the truncation error - see (1.143) below. Introducing an intermediate level $U^{n+1/2}$, (1.136) can be written in the equivalent form

$$(1-\frac{1}{2}\mu_x\delta_x^2)U^{n+1/2} = (1+\frac{1}{2}\mu_y\delta_y^2)U^n, \tag{1.138}$$

$$(1-\frac{1}{2}\mu_y\delta_y^2)U^{n+1} = (1+\frac{1}{2}\mu_x\delta_x^2)U^{n+1/2}, \tag{1.139}$$

the equivalence being seen by operating on (1.138) with $(1+\frac{1}{2}\mu_x\delta_x^2)$ and on (1.139) with $(1-\frac{1}{2}\mu_x\delta_x^2)$.

In (1.138) the terms on the right-hand side are known from the previous step; having computed $U^{n+1/2}$, the terms on the right-hand side of (1.139) are then also known. Just as for the singly implicit scheme (1.135) the solution of each of the systems involves sets of tridiagonal equations. The total work involved in one time step amounts to solving J_y-1 tridiagonal systems (for points such as those marked by crosses in Fig. 3.1), each of order J_x-1, followed by solving J_x-1 similar systems (for points such as those marked by dots in Fig. 3.1), each of order (J_y-1). This whole process can be carried out very much faster than the solution of the full system of order $(J_x-1)(J_y-1)$ in the Crank-Nicolson method; we need approximately 10(add) + 8(mult) + 6(div) operations per mesh point as compared with 4(add) + 3(mult) for the explicit scheme (1.124): that is about three times as much computation. Boundary conditions for $U^{n+1/2}$ are needed at all the points marked with $\square$ in Fig. 3.1 and for U^n at the points marked with $\odot$.

When b is constant we can again analyse the stability of (1.136) by substituting the Fourier mode (1.130). From either form we obtain

$$\lambda(\mathbf{k}) = \frac{(1-2\mu_x\sin^2\frac{1}{2}k_x\Delta x)(1-2\mu_y\sin^2\frac{1}{2}k_y\Delta y)}{(1+2\mu_x\sin^2\frac{1}{2}k_x\Delta x)(1+2\mu_y\sin^2\frac{1}{2}k_y\Delta y)} \tag{1.140}$$

from which the scheme's unconditional stability follows immediately.

We can also apply maximum principle arguments to (1.138)-(1.139). In the first half-step an individual equation takes the form

$$(1+\mu_x)U_{r,s}^{n+1/2} = (1-\mu_y)U_{r,s}^n + \frac{1}{2}\mu_y(U_{r,s-1}^n + U_{r,s+1}^n) + \frac{1}{2}\mu_x(U_{r+1,s}^{n+1/2} + U_{r-1,s}^{n+1/2}). \tag{1.141}$$

Thus, provided that $\mu_y \le 1$, the value of $U_{r,s}^{n+1/2}$ is expressed as a linear combination, with nonnegative coefficients summing to unity, of neighbouring values of U^n and $U^{n+1/2}$. The same is evidently also true of the equation in the second half-step. Thus the maximum principle shows that the numerical values are bounded by the maximum and minimum values on the boundaries, provided that

$$\max\{\mu_x, \mu_y\} \le 1 \tag{1.142}$$

which is a natural generalisation of the condition for the one-dimensional Crank-Nicolson method.

The truncation error is most readily calculated from the unsplit form (1.136). Taking all terms to the left and dividing by Δt we fairly readily deduce that the leading terms are

$$\begin{aligned} T^{n+1/2} &\approx \frac{1}{24}(\Delta t)^2 u_{ttt} - \frac{1}{12}(\Delta x)^2 u_{xxxx} - \frac{1}{12}(\Delta y)^2 u_{yyyy} \\ &\quad -\frac{1}{8}(\Delta t)^2 u_{xxtt} - \frac{1}{8}(\Delta t)^2 u_{yytt} + \frac{1}{4}(\Delta t)^2 u_{xxyyt} \\ &= O\left((\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2\right), \end{aligned} \tag{1.143}$$

the first five terms being as for the Crank-Nicolson scheme in two dimensions (cf. (1.47) for the one-dimensional case) and the last coming from the product term $\delta_x^2\delta_y^2\left(U^{n+1}-U^n\right)$.

See the example for the M-shaped initial data in the textbook.

Three-dimensions and curved boundaries are also very important topics in parabolic equation problems.