2

Parabolic equations in one space variable

2.1 Introduction

In this chapter we shall be concerned with the numerical solution of parabolic equations in one space variable and the time variable t. We begin with the simplest model problem, for heat conduction in a uniform medium. For this model problem an explicit difference method is very straightforward in use, and the analysis of its error is easily accomplished by the use of a maximum principle, or by Fourier analysis. As we shall show, however, the numerical solution becomes unstable unless the time step is severely restricted, so we shall go on to consider other, more elaborate, numerical methods which can avoid such a restriction. The additional complication in the numerical calculation is more than offset by the smaller number of time steps needed. We then extend the methods to problems with more general boundary conditions, then to more general linear parabolic equations. Finally we shall discuss the more difficult problem of the solution of nonlinear equations.

2.2 A model problem

Many problems in science and engineering are modelled by special cases of the linear parabolic equation for the unknown $u(x,t)$

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(b(x,t)\frac{\partial u}{\partial x}\right) + c(x,t)u + d(x,t) \tag{2.1}$$

where b is strictly positive. An *initial condition* will be needed; if this is given at $t = 0$ it will take the form

$$u(x,0) = u^0(x) \tag{2.2}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

where $u^0(x)$ is a given function. The solution of the problem will be required to satisfy (2.1) for $t > 0$ and x in an open region R which will be typically either the whole real line, the half-line $x > 0$, or an interval such as $(0, 1)$. In the two latter cases we require the solution to be defined on the closure of R and to satisfy certain *boundary conditions*; we shall assume that these also are linear, and may involve u or its first space derivative $\partial u/\partial x$, or both. If $x = 0$ is a left-hand boundary, the boundary condition will be of the form

$$\alpha_0(t)u + \alpha_1(t)\frac{\partial u}{\partial x} = \alpha_2(t) \tag{2.3}$$

where

$$\alpha_0 \geq 0,\ \alpha_1 \leq 0 \quad \text{and} \quad \alpha_0 - \alpha_1 > 0. \tag{2.4}$$

If $x = 1$ is a right-hand boundary we shall need a condition of the form

$$\beta_0(t)u + \beta_1(t)\frac{\partial u}{\partial x} = \beta_2(t) \tag{2.5}$$

where

$$\beta_0 \geq 0,\ \beta_1 \geq 0 \quad \text{and} \quad \beta_0 + \beta_1 > 0. \tag{2.6}$$

The reason for the conditions on the coefficients α and β will become apparent later. Note the change of sign between α_1 and β_1, reflecting the fact that at the right-hand boundary $\partial/\partial x$ is an outward normal derivative, while in (2.3) it was an inward derivative.

We shall begin by considering a simple model problem, the equation for which models the flow of heat in a homogeneous unchanging medium, of finite extent, with no heat source. We suppose that we are given homogeneous *Dirichlet boundary conditions*, i.e., the solution is given to be zero at each end of the range, for all values of t. After changing to dimensionless variables this problem becomes: find $u(x, t)$ defined for $x \in [0, 1]$ and $t \geq 0$ such that

$$u_t = u_{xx} \quad \text{for} \quad t > 0,\ 0 < x < 1, \tag{2.7}$$

$$u(0, t) = u(1, t) = 0 \quad \text{for} \quad t > 0, \tag{2.8}$$

$$u(x, 0) = u^0(x), \quad \text{for} \quad 0 \leq x \leq 1. \tag{2.9}$$

Here we have introduced the common subscript notation to denote partial derivatives.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

2.3 Series approximation

This differential equation has special solutions which can be found by the method of *separation of variables*. The method is rather restricted in its application, unlike the finite difference methods which will be our main concern. However, it gives useful solutions for comparison purposes, and leads to a natural analysis of the stability of finite difference methods by the use of Fourier analysis.

We look for a solution of the special form $u(x,t) = f(x)g(t)$; substituting into the differential equation we obtain

$$fg' = f''g,$$

i.e.,

$$g'/g = f''/f. \tag{2.10}$$

In this last equation the left-hand side is independent of x, and the right-hand side is independent of t, so that both sides must be constant. Writing this constant as $-k^2$, we immediately solve two simple equations for the functions f and g, leading to the solution

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

This shows the reason for the choice of $-k^2$ for the constant; if we had chosen a positive value here, the solution would have involved an exponentially increasing function of t, whereas the solution of our model problem is known to be bounded for all positive values of t. For all values of the number k this is a solution of the differential equation; if we now restrict k to take the values $k = m\pi$, where m is a positive integer, the solution vanishes at $x = 1$ as well as at $x = 0$. Hence any linear combination of such solutions will satisfy the differential equation and the two boundary conditions. This linear combination can be written

$$u(x,t) = \sum_{m=1}^{\infty} a_m \mathrm{e}^{-(m\pi)^2 t} \sin m\pi x. \tag{2.11}$$

We must now choose the coefficients a_m in this linear combination in order to satisfy the given initial condition. Writing $t = 0$ we obtain

$$\sum_{m=1}^{\infty} a_m \sin m\pi x = u^0(x). \tag{2.12}$$

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

$$a_m = 2 \int_0^1 u^0(x) \sin m\pi x \, \mathrm{d}x. \tag{2.13}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

This final result may be regarded as an exact analytic solution of the problem, but it is much more like a numerical approximation, for two reasons. If we require the value of $u(x,t)$ for specific values of x and t, we must first determine the Fourier coefficients a_m; these can be found exactly only for specially simple functions $u^0(x)$, and more generally would require some form of numerical integration. And secondly we can only sum a finite number of terms of the infinite series. For the model problem, however, it is a very efficient method; for even quite small values of t a few terms of the series will be quite sufficient, as the series converges extremely rapidly. The real limitation of the method in this form is that it does not easily generalise to even slightly more complicated differential equations.

2.4 An explicit scheme for the model problem

To approximate the model equation (2.7) by finite differences we divide the closed domain $\bar{R}\times[0,t_F]$ by a set of lines parallel to the x- and t-axes to form a grid or mesh. We shall assume, for simplicity only, that the sets of lines are equally spaced, and from now on we shall assume that $\bar{R}$ is the interval $[0,1]$. Note that in practice we have to work in a finite time interval $[0,t_F]$, but t_F can be as large as we like.

We shall write Δx and Δt for the line spacings. The crossing points

$$(x_j = j\Delta x,\ t_n = n\Delta t),\ j = 0,1,\ldots,J,\ n = 0,1,\ldots, \tag{2.14}$$

where

$$\Delta x = 1/J, \tag{2.15}$$

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

$$U_j^n \approx u(x_j,t_n). \tag{2.16}$$

We shall approximate the derivatives in (2.7) by finite differences and then solve the resulting difference equations in an evolutionary manner starting from $n=0$.

We shall often use notation like U_j^n; there should be no confusion with other expressions which may look similar, such as λ^n which, of course, denotes the nth power of λ. If there is likely to be any ambiguity we shall sometimes write such a power in the form $(\lambda_j)^n$.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

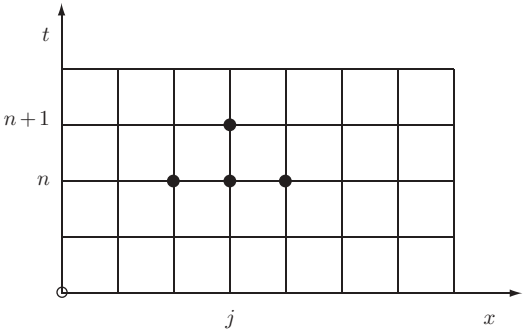

Fig. 2.1. An explicit scheme.

For the model problem the simplest difference scheme based at the mesh point (x_j, t_n) uses a forward difference for the time derivative; this gives

$$\frac{v(x_j, t_{n+1}) - v(x_j, t_n)}{\Delta t} \approx \frac{\partial v}{\partial t}(x_j, t_n) \tag{2.17}$$

for any function v with a continuous t-derivative. The scheme uses a centred second difference for the second order space derivative:

$$\frac{v(x_{j+1}, t_n) - 2v(x_j, t_n) + v(x_{j-1}, t_n)}{(\Delta x)^2} \approx \frac{\partial^2 v}{\partial x^2}(x_j, t_n). \tag{2.18}$$

The approximation generated by equating the left-hand sides of (2.17) and (2.18) thus satisfies

$$U_j^{n+1} = U_j^n + \mu(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \tag{2.19}$$

where

$$\mu := \frac{\Delta t}{(\Delta x)^2}. \tag{2.20}$$

The pattern of grid points involved in (2.19) is shown in Fig. 2.1; 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

$$U_j^0 = u^0(x_j),\ j = 1, 2, \ldots, J-1, \tag{2.21}$$

$$U_0^n = U_J^n = 0,\ n = 0, 1, 2, \ldots, \tag{2.22}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$u^0(0) = u^0(1) = 0 \tag{2.23}$$

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

However, if we carry out a calculation using (2.19), (2.21) and (2.22) 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. 2.2 we show results corresponding to initial data in the form of a 'hat function',

$$u^0(x) = \begin{cases} 2x & \text{if} \quad 0 \le x \le \frac{1}{2}, \\ 2 - 2x & \text{if} \quad \frac{1}{2} \le x \le 1. \end{cases} \tag{2.24}$$

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.

We shall now analyse this behaviour, and obtain bounds on the error, in a more formal way. First we introduce some notation and definitions.

2.5 Difference notation and truncation error

We define finite differences in the same way in the two variables t and x; there are three kinds of finite differences:

forward differences

$$\Delta_{+t} v(x,t) := v(x, t + \Delta t) - v(x,t), \tag{2.25a}$$

$$\Delta_{+x} v(x,t) := v(x + \Delta x, t) - v(x,t); \tag{2.25b}$$

backward differences

$$\Delta_{-t} v(x,t) := v(x,t) - v(x, t - \Delta t), \tag{2.26a}$$

$$\Delta_{-x} v(x,t) := v(x,t) - v(x - \Delta x, t); \tag{2.26b}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

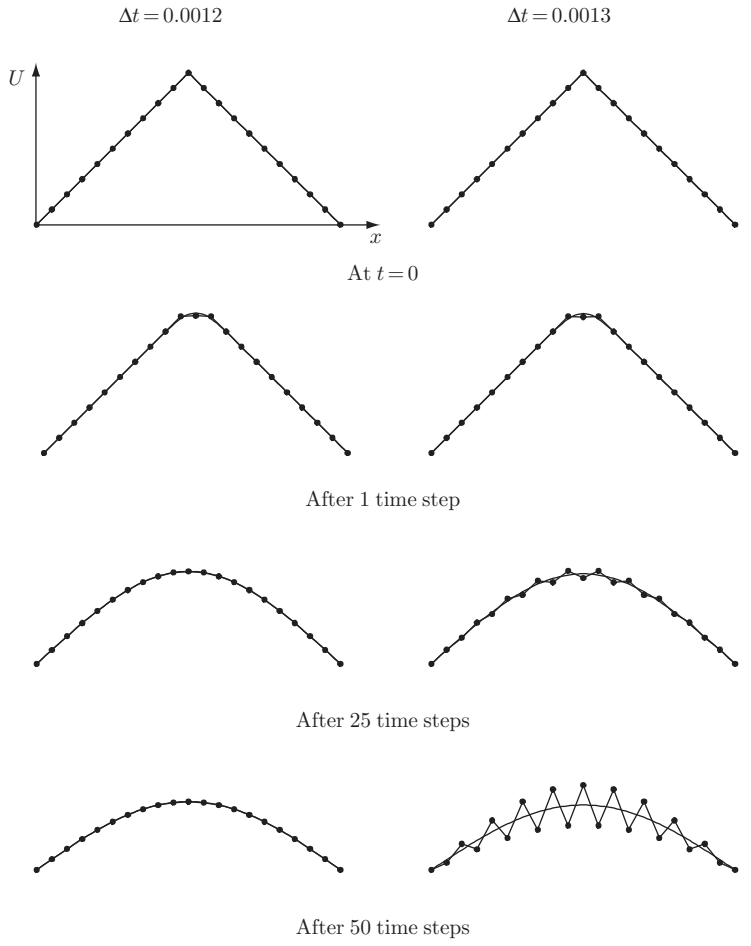

Fig. 2.2. Results obtained for the data of (2.24) with the explicit method; $J = 20$, $\Delta x = 0.05$. The exact solution is shown by the full curved line.

central differences

$$\delta_t v(x,t) := v(x, t + \tfrac{1}{2}\Delta t) - v(x, t - \tfrac{1}{2}\Delta t), \qquad (2.27a)$$

$$\delta_x v(x,t) := v(x + \tfrac{1}{2}\Delta x, t) - v(x - \tfrac{1}{2}\Delta x, t). \qquad (2.27b)$$

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

$$\delta_x^2 v(x,t) := v(x + \Delta x, t) - 2v(x,t) + v(x - \Delta x, t). \qquad (2.28)$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

For first differences it is often convenient to use the double interval central difference

$$\begin{aligned}\Delta_{0x}v(x,t) &:= \tfrac{1}{2}(\Delta_{+x}+\Delta_{-x})v(x,t)\\ &= \tfrac{1}{2}[v(x+\Delta x,t)-v(x-\Delta x,t)].\end{aligned}$$

A Taylor series expansion of the forward difference in t gives for the solution of (2.7)

$$\begin{aligned}\Delta_{+t}u(x,t) &= u(x,t+\Delta t)-u(x,t)\\ &= u_t\Delta t+\tfrac{1}{2}u_{tt}(\Delta t)^2+\tfrac{1}{6}u_{ttt}(\Delta t)^3+\cdots.\end{aligned}\tag{2.29}$$

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

$$\delta_x^2\,u(x,t)=u_{xx}(\Delta x)^2+\tfrac{1}{12}u_{xxxx}(\Delta x)^4\ +\cdots.\tag{2.30}$$

We can now define the *truncation error* of the scheme (2.19). We first multiply the difference equation throughout by a factor, if necessary, so that each term is an approximation to the corresponding derivative in the differential equation. Here this step is unnecessary, provided that we use the form

$$\frac{U_j^{n+1}-U_j^n}{\Delta t}=\frac{U_{j+1}^n-2U_j^n+U_{j-1}^n}{(\Delta x)^2}\tag{2.31}$$

rather than (2.19). 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 $T(x,t)$

$$T(x,t):=\frac{\Delta_{+t}u(x,t)}{\Delta t}-\frac{\delta_x^2u(x,t)}{(\Delta x)^2}\tag{2.32}$$

so that

$$\begin{aligned}T(x,t) &= (u_t-u_{xx})+\left(\tfrac{1}{2}u_{tt}\Delta t-\tfrac{1}{12}u_{xxxx}(\Delta x)^2\right)+\cdots\\ &= \tfrac{1}{2}u_{tt}\Delta t-\tfrac{1}{12}u_{xxxx}(\Delta x)^2+\cdots\end{aligned}\tag{2.33}$$

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.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

We have used Taylor series expansions to express the truncation error as an infinite series. It is often convenient to truncate the infinite Taylor series, introducing a remainder term; for example

$$\begin{aligned} u(x,t+\Delta t) &= u(x,t) + u_t\Delta t + \tfrac{1}{2}u_{tt}(\Delta t)^2 + \tfrac{1}{6}u_{ttt}(\Delta t)^3 + \cdots \\ &= u(x,t) + u_t\Delta t + \tfrac{1}{2}u_{tt}(x,\eta)(\Delta t)^2, \end{aligned} \tag{2.34}$$

where η lies somewhere between t and $t+\Delta t$. If we do the same thing for the x expansion the truncation error becomes

$$T(x,t) = \tfrac{1}{2}u_{tt}(x,\eta)\Delta t - \tfrac{1}{12}u_{xxxx}(\xi,t)(\Delta x)^2 \tag{2.35}$$

where $\xi \in (x-\Delta x, x+\Delta x)$, from which it follows that

$$|T(x,t)| \le \tfrac{1}{2}M_{tt}\Delta t + \tfrac{1}{12}M_{xxxx}(\Delta x)^2 \tag{2.36}$$

$$= \tfrac{1}{2}\Delta t\left[M_{tt} + \tfrac{1}{6\mu}M_{xxxx}\right], \tag{2.37}$$

where M_{tt} is a bound for $|u_{tt}|$ and M_{xxxx} is a bound for $|u_{xxxx}|$. 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 problem of Fig. 2.2 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 (2.37) 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.*

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

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

Thus for $\mu = \frac{1}{6}$ 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.

2.6 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 $\mu = \Delta t/(\Delta x)^2$, but with successive refinement of the two meshes, so 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 $(0,1) \times (\tau, t_F)$,

$$x_j \to x^*, \quad t_n \to t^* \quad \text{implies} \quad U_j^n \to u(x^*, t^*). \tag{2.38}$$

We shall prove that the explicit scheme for our problem is convergent if $\mu \le \frac{1}{2}$.

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 $\bar{T} = \bar{T}(\Delta x, \Delta t)$ for the truncation error, which holds for all mesh points on a given mesh, and use the notation T_j^n for $T(x_j, t_n)$:

$$|T_j^n| \le \bar{T}. \tag{2.39}$$

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

$$e_j^n \;:=\; U_j^n - u(x_j, t_n). \tag{2.40}$$

Now U_j^n satisfies the equation (2.19) 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. Hence by subtraction we obtain

$$e_j^{n+1} = e_j^n + \mu\delta_x^2 e_j^n \;-\; T_j^n \Delta t \tag{2.41}$$

which is in detail

$$e_j^{n+1} \;=\; (1-2\mu)e_j^n + \mu e_{j+1}^n + \mu e_{j-1}^n - \; T_j^n \Delta t. \tag{2.42}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$E^n \;:=\; \max\{|e_j^n|, \quad j = 0, 1, \ldots, J\}, \tag{2.43}$$

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

$$\begin{aligned} \left|e_j^{n+1}\right| &\leq (1-2\mu)E^n + \mu E^n + \mu E^n + \; |T_j^n|\Delta t \\ &\leq E^n + \; \bar{T}\Delta t. \end{aligned} \tag{2.44}$$

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

$$E^{n+1} \;\leq\; E^n + \bar{T}\Delta t. \tag{2.45}$$

Suppose for the moment that the bound (2.39) 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 (2.37)

$$\begin{aligned} E^n \;&\leq\; \tfrac{1}{2}\Delta t \left[M_{tt} + \tfrac{1}{6\mu}M_{xxxx}\right] t_F \\ &\to 0 \quad \text{as} \quad \Delta t \to 0. \end{aligned} \tag{2.46}$$

In our model problem, if it is useful we can write $M_{tt} = M_{xxxx}$.

We can now state this convergence property in slightly more general terms. In order to define convergence of a difference scheme which involves two mesh sizes Δt and Δx we need to be clear about what relationship we assume between them as they both tend to zero. We therefore introduce the concept of a refinement path. A *refinement path* is a sequence of pairs of mesh sizes, Δx and Δt, each of which tends to zero:

$$\textit{refinement path} := \{((\Delta x)_i, (\Delta t)_i), i = 0, 1, 2, \ldots; (\Delta x)_i, (\Delta t)_i \;\to\; 0\}\,. \tag{2.47}$$

We can then specify particular refinement paths by requiring, for example, that $(\Delta t)_i$ is proportional to $(\Delta x)_i$, or to $(\Delta x)_i^2$. Here we just define

$$\mu_i = \frac{(\Delta t)_i}{(\Delta x)_i^2} \tag{2.48}$$

and merely require that $\mu_i \leq \frac{1}{2}$. Some examples are shown in Fig. 2.3.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

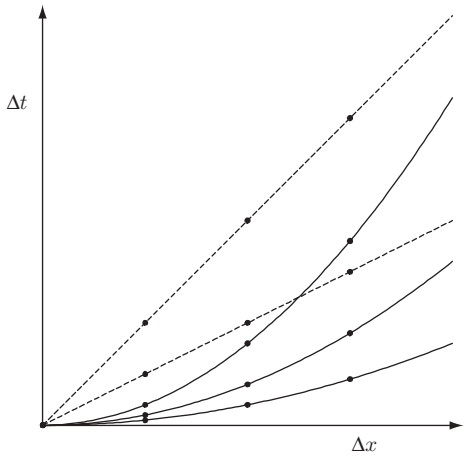

Fig. 2.3. Refinement paths; shown as full lines for constant $\Delta t/(\Delta x)^2$ and as dashed lines for constant $\Delta t/\Delta x$.

Theorem 2.1 *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 (2.19) for $i = 0, 1, 2 \ldots$ converge to the solution $u(x,t)$ of the differential equation, uniformly in the region.

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.

As an example with smoother properties than is given by the data of (2.24), consider the solution of the heat equation with

$$u(x,0) = x(1-x), \tag{2.49a}$$

$$u(0,t) = u(1,t) = 0, \tag{2.49b}$$

on the region $[0,1] \times [0,1]$. Errors obtained with the explicit method are shown in Fig. 2.4. This shows a graph of $\log_{10} E^n$ against t_n, where

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

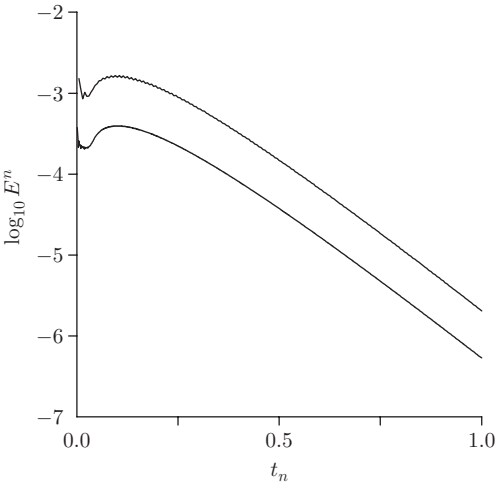

Fig. 2.4. Error decay for the explicit method applied to the heat equation with initial condition $u(x,0) = x(1-x)$. The top curve is for $\Delta x = 0.1$, $\mu = 0.5$, and the bottom curve is for $\Delta x = 0.05, \mu = 0.5$.

E^n is given by (2.43). 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. 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 $\Delta t = \frac{1}{2}(\Delta x)^2$. Notice also that after some early variation the error tends to zero as t increases; our error bound in (2.45) is pessimistic, as it continues to increase with t. 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 Section 2.10.

2.7 Fourier analysis of the error

We have already expressed the exact solution of the differential equation as a Fourier series; 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. Suppose we substitute

$$U_j^n = (\lambda)^n \mathrm{e}^{\mathrm{i}k(j\Delta x)} \tag{2.50}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

into the difference equation (2.19), 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

$$\begin{aligned}\lambda \equiv \lambda(k) &= 1 + \mu\left(\mathrm{e}^{\mathrm{i}k\Delta x} - 2 + \mathrm{e}^{-\mathrm{i}k\Delta x}\right)\\ &= 1 - 2\mu(1 - \cos k\Delta x)\\ &= 1 - 4\mu\,\sin^2 \tfrac{1}{2}k\Delta x; \end{aligned} \tag{2.51}$$

$\lambda(k)$ is called the *amplification factor* for the mode. By taking $k = m\pi$ as in (2.11), we can therefore write our numerical approximation in the form

$$U_j^n = \sum_{-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi(j\Delta x)}\left[\lambda(m\pi)\right]^n. \tag{2.52}$$

The low frequency terms in this expansion give a good approximation to the exact solution of the differential equation, given by (2.11), because the series expansions for $\lambda(k)$ and $\exp(-k^2\Delta t)$ match reasonably well:

$$\begin{aligned}\exp(-k^2\Delta t) &= 1 - k^2\Delta t + \tfrac{1}{2}k^4(\Delta t)^2 - \cdots,\\ \lambda(k) &= 1 - 2\mu\left[\tfrac{1}{2}(k\Delta x)^2 - \tfrac{1}{24}(k\Delta x)^4 + \cdots\right]\\ &= 1 - k^2\Delta t + \tfrac{1}{12}k^4\Delta t(\Delta x)^2 - \cdots. \end{aligned} \tag{2.53}$$

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, but when $(\Delta x)^2 = 6\Delta t$ we shall have second 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

$$|\lambda(k) - \mathrm{e}^{-k^2\Delta t}| \le C(\mu)k^4(\Delta t)^2 \qquad \forall k, \Delta t > 0. \tag{2.54}$$

Theorem 2.1 establishes convergence and an error bound under the restriction $\mu \le \frac{1}{2}$, 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. For large values of k the modes in the exact solution are rapidly damped by the exponential term $\exp(-k^2t)$. But in the numerical solution the damping factor $|\lambda(k)|$ will become greater than unity for large values of k if $\mu > \frac{1}{2}$; in particular this will happen when $k\Delta x = \pi$, for then $\lambda(k) = 1 - 4\mu$. These Fourier modes will then grow unboundedly as n increases. It is possible in principle to choose the initial conditions so that these Fourier modes do not appear in the solution. But this would be a very special problem, and in practice the effect of rounding errors would be to introduce small

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

components of all the modes, some of which would then grow without bound. For the present model problem we shall say that a method is *stable* if there exists a constant K, independent of k, such that

$$|[\lambda(k)]^n| \leq K, \quad \text{for} \quad n\Delta t \leq t_F, \ \forall k. \tag{2.55}$$

Essentially, stability has to do with the bounded growth in a finite time of the difference between two solutions of the difference equations, uniformly in the mesh size; we shall formulate a general definition of stability in a later chapter. Evidently, for stability we require the condition, due to von Neumann,

$$|\lambda(k)| \leq 1 + K'\Delta t \tag{2.56}$$

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 have used a representation for U_j^n as the infinite Fourier series (2.52), 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

$$k = m\pi, \quad m = -(J-1), -(J-2), \ldots, -1, 0, 1, \ldots, J. \tag{2.57}$$

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 (2.51) 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. 2.2.

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,

$$\begin{aligned} e_j^n &= U_j^n - u(x_j, t_n) \\ &= \sum_{-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}m\pi j\Delta x} \left\{ [\lambda(m\pi)]^n - \mathrm{e}^{-m^2\pi^2 n\Delta t} \right\}, \end{aligned} \tag{2.58}$$

where we have also used the full Fourier series for $u(x,t)$ instead of the sine series as in the particular case of (2.11); this will allow treatment other than of the simple boundary conditions of (2.8). We now split this infinite sum into two parts. Given an arbitrary positive ϵ, we choose m_0 such that

$$\sum_{|m|>m_0} |A_m| \leq \tfrac{1}{4}\epsilon. \tag{2.59}$$

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

$$|(\lambda_1)^n - (\lambda_2)^n| \leq n|\lambda_1 - \lambda_2|; \tag{2.60}$$

so from (2.54) we have

$$\begin{aligned} |e_j^n| &\leq \tfrac{1}{2}\epsilon + \sum_{|m|\leq m_0} |A_m| \left| [\lambda(m\pi)]^n - \mathrm{e}^{-m^2\pi^2 n\Delta t} \right| \\ &\leq \tfrac{1}{2}\epsilon + \sum_{|m|\leq m_0} |A_m| nC(\mu) \left(m^2\pi^2\Delta t\right)^2. \end{aligned} \tag{2.61}$$

We can thus deduce that

$$|e_j^n| \leq \tfrac{1}{2}\epsilon + t_F C(\mu)\pi^4 \left[\sum_{|m|\leq m_0} |A_m| m^4 \right] \Delta t \tag{2.62}$$

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.

2.8 An implicit method

The stability limit $\Delta t \leq \frac{1}{2}(\Delta x)^2$ 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

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

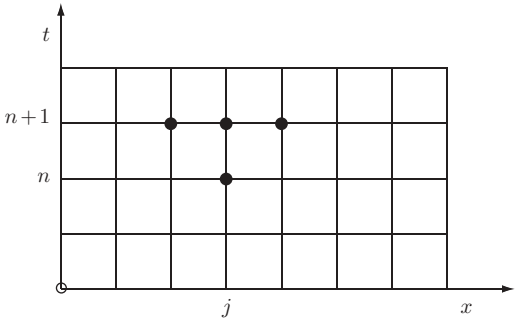

Fig. 2.5. The fully implicit scheme.

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

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{(\Delta x)^2} \tag{2.63}$$

instead of (2.31). This may be written using the difference notation given in Section 2.5 as

$$\Delta_{-t} U_j^{n+1} = \mu \delta_x^2 U_j^{n+1},$$

where $\mu = \Delta t/(\Delta x)^2$, 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 (2.63) 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

$$-\mu U_{j-1}^{n+1} + (1+2\mu) U_j^{n+1} - \mu U_{j+1}^{n+1} = U_j^n. \tag{2.64}$$

Giving j the values $1, 2, \ldots, (J-1)$ we thus obtain a system of $J-1$ linear equations in the $J-1$ unknowns $U_j^{n+1}, j = 1, 2, \ldots, J-1$. Instead of calculating each of these unknowns by a separate trivial formula, we

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

must now solve this system of equations to give the values simultaneously. 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.

2.9 The Thomas algorithm

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

$$-a_j U_{j-1} + b_j U_j - c_j U_{j+1} = d_j, \quad j = 1, 2, \ldots, J-1, \tag{2.65}$$

with

$$U_0 = 0, \quad U_J = 0. \tag{2.66}$$

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

$$a_j > 0, \quad b_j > 0, \quad c_j > 0, \tag{2.67}$$

$$b_j > a_j + c_j. \tag{2.68}$$

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.

The Thomas algorithm operates by reducing the system of equations to upper triangular form, by eliminating the term in U_{j-1} in each of the equations. This is done by treating each equation in turn. Suppose that the first k of equations (2.65) have been reduced to

$$U_j - e_j U_{j+1} = f_j, \quad j = 1, 2, \ldots, k. \tag{2.69}$$

The last of these equations is therefore $U_k - e_k U_{k+1} = f_k$, and the next equation, which is still in its original form, is

$$-a_{k+1} U_k + b_{k+1} U_{k+1} - c_{k+1} U_{k+2} = d_{k+1}.$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

It is easy to eliminate U_k from these two equations, giving a new equation involving U_{k+1} and U_{k+2},

$$U_{k+1} - \frac{c_{k+1}}{b_{k+1} - a_{k+1}e_k} U_{k+2} = \frac{d_{k+1} + a_{k+1}f_k}{b_{k+1} - a_{k+1}e_k}.$$

Comparing this with (2.69) shows that the coefficients e_j and f_j can be obtained from the recurrence relations

$$e_j = \frac{c_j}{b_j - a_j e_{j-1}}, \quad f_j = \frac{d_j + a_j f_{j-1}}{b_j - a_j e_{j-1}}, \quad j = 1, 2, \ldots, J-1; \tag{2.70}$$

while identifying the boundary condition $U_0 = 0$ with (2.69) for $j = 0$ gives the initial values

$$e_0 = f_0 = 0. \tag{2.71}$$

Having used these recurrence relations to find the coefficients, the values of U_j are easily obtained from (2.69): beginning from the known value of U_J, this equation gives the values of U_{J-1}, U_{J-2}, ..., in order, finishing with U_1.

The use of a recurrence relation like (2.69) to calculate the values of U_j in succession may in general be numerically unstable, and lead to increasing errors. However, this will not happen if, for each j, $|e_j| < 1$ in (2.69), and we leave it as an exercise to show that the conditions (2.67) and (2.68) are sufficient to guarantee this (see Exercise 4).

The algorithm is very efficient (on a serial computer) so that (2.64) is solved with 3(add) + 3(multiply) + 2(divide) operations per mesh point, as compared with 2(add) + 2(multiply) operations per mesh point for the explicit algorithm (2.19). Thus it takes about twice as long for each time step. 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 shall give a proof of the convergence of this implicit scheme in the next section, as a particular case of a more general method. First we can examine its stability by the Fourier method of Section 2.7.

We construct a solution of the difference equations for Fourier modes of the same form as before,

$$U_j^n = (\lambda)^n \mathrm{e}^{\mathrm{i}k(j\Delta x)}. \tag{2.72}$$

This will satisfy (2.64) provided that

$$\begin{aligned} \lambda - 1 &= \mu\lambda(\mathrm{e}^{\mathrm{i}k\Delta x} - 2 + \mathrm{e}^{-\mathrm{i}k\Delta x}) \\ &= -4\mu\lambda \sin^2 \tfrac{1}{2}k\Delta x, \end{aligned} \tag{2.73}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

which shows that

$$\lambda = \frac{1}{1 + 4\mu \sin^2 \frac{1}{2} k\Delta x}. \tag{2.74}$$

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.

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.

2.10 The weighted average or θ-method

We have now considered two finite difference methods, which differ only in that one approximates the second space derivative by three points on the old time level, t_n, and the other uses the three points on the new time level, t_{n+1}. 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. 2.6)

$$U_j^{n+1} - U_j^n = \mu \left[\theta \delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n\right], \quad j = 1, 2, \ldots, J-1. \tag{2.75}$$

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 $\left\{U_j^{n+1}\right\}$, namely,

$$-\theta\mu U_{j-1}^{n+1} + (1 + 2\theta\mu) U_j^{n+1} - \theta\mu U_{j+1}^{n+1} = \left[1 + (1-\theta)\mu\delta_x^2\right] U_j^n. \tag{2.76}$$

Clearly the coefficients satisfy (2.67) and (2.68), so the system can be solved stably by the Thomas algorithm given above for the fully implicit scheme.

Let us consider the stability of this one-parameter family of schemes by using Fourier analysis as in Section 2.7 and above. Substituting the mode (2.72) into equation (2.75), we obtain

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

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

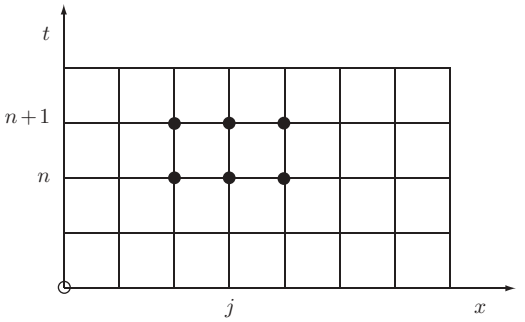

Fig. 2.6. The θ-method.

i.e.,

$$\lambda = \frac{1 - 4(1-\theta)\mu \sin^2 \frac{1}{2}k\Delta x}{1 + 4\theta\mu \sin^2 \frac{1}{2}k\Delta x}. \tag{2.77}$$

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

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

i.e.,

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

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

$$\mu(1-2\theta) > \tfrac{1}{2}. \tag{2.78}$$

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 $\theta \geq \frac{1}{2}$ is unstable for any μ. If condition (2.78) 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 (2.78) is not satisfied, we have $|\lambda(k)| \leq 1$ for every mode k, so that no mode grows at all and the scheme

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

is stable. Thus we can summarise the necessary and sufficient conditions for the stability of (2.75) as

$$\left.\begin{array}{l} \text{when } 0 \le \theta < \tfrac{1}{2}, \quad \text{stable if and only if } \ \mu \le \tfrac{1}{2}(1-2\theta)^{-1} \\ \\ \text{when } \tfrac{1}{2} \le \theta \le 1, \quad \text{stable for all } \ \mu. \end{array}\right\} \tag{2.79}$$

The two cases are often referred to as conditional and unconditional stability respectively. As soon as θ is non-zero a tridiagonal system has to be solved, so there would seem to be no advantage in using schemes with $0 < \theta < \frac{1}{2}$ which are only conditionally stable – unless they were more accurate. Thus we should next look at the truncation error for (2.75).

To calculate the truncation error for such a six-point scheme it is important to make a careful choice of the point about which the Taylor series are to be expanded. It is clear that the leading terms in the truncation error will be the same for any choice of this expansion point: but the convenience and simplicity of the calculation can be very materially affected. Thus for the explicit scheme (2.31) the natural and convenient point was (x_j, t_n): by the same argument the natural point of expansion for the purely implicit scheme (2.63) would be (x_j, t_{n+1}). However, for any intermediate value of θ we shall use the centre of the six mesh points, namely $(x_j, t_{n+1/2})$, and often write the truncation error as $T_j^{n+1/2}$. It is also helpful to group the terms in the scheme in a symmetric manner so as to take maximum advantage of cancellations in the Taylor expansions. Working from (2.75) we therefore have, using the superscript/subscript notation for u as well as U,

$$u_j^{n+1} = \left[u + \tfrac{1}{2}\Delta t\, u_t + \tfrac{1}{2}\left(\tfrac{1}{2}\Delta t\right)^2 u_{tt} + \tfrac{1}{6}\left(\tfrac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots\right]_j^{n+1/2},$$

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

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

$$\delta_t u_j^{n+1/2} = u_j^{n+1} - u_j^n = \left[\Delta t\, u_t + \tfrac{1}{24}(\Delta t)^3 u_{ttt} + \cdots\right]_j^{n+1/2}. \tag{2.80}$$

Also from (2.30) we have

$$\delta_x^2 u_j^{n+1} = \left[(\Delta x)^2 u_{xx} + \tfrac{1}{12}(\Delta x)^4 u_{xxxx} + \tfrac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right]_j^{n+1}. \tag{2.81}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

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

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

$$\begin{aligned}\theta\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n = &\\ &\left[(\Delta x)^2 u_{xx} + \tfrac{1}{12}(\Delta x)^4 u_{xxxx} + \tfrac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right] \\ &+ (\theta - \tfrac{1}{2})\Delta t\left[(\Delta x)^2 u_{xxt} + \tfrac{1}{12}(\Delta x)^4 u_{xxxxt} + \cdots\right] \\ &+ \tfrac{1}{8}(\Delta t)^2(\Delta x)^2\left[u_{xxtt}\right] + \cdots . \qquad (2.82)\end{aligned}$$

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

$$T_j^{n+1/2} := \frac{\delta_t u_j^{n+1/2}}{\Delta t} - \frac{\theta\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n}{(\Delta x)^2} \qquad (2.83)$$

$$\begin{aligned}= &\left[u_t - u_{xx}\right] + \left[\left(\tfrac{1}{2} - \theta\right)\Delta t\, u_{xxt} - \tfrac{1}{12}(\Delta x)^2 u_{xxxx}\right] \\ &+ \left[\tfrac{1}{24}(\Delta t)^2 u_{ttt} - \tfrac{1}{8}(\Delta t)^2 u_{xxtt}\right] \\ &+ \left[\tfrac{1}{12}\left(\tfrac{1}{2} - \theta\right)\Delta t\,(\Delta x)^2 u_{xxxxt} - \tfrac{2}{6!}(\Delta x)^4 u_{xxxxxx}\right] \qquad (2.84)\end{aligned}$$

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

The first term in (2.84) 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

[1] Crank, J. and Nicolson, P. (1947) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. Philos. Soc.* **43**, 50–67.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

term in (2.84) does not cancel even when we exploit the differential equation to obtain

$$T_j^{n+1/2} = -\tfrac{1}{12}\left[(\Delta x)^2 u_{xxxx} + (\Delta t)^2 u_{ttt}\right]_j^{n+1/2} + \cdots \quad (2.85)$$
$$(\text{when } \theta = \tfrac{1}{2}),$$

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 $O\big((\Delta t)^2\big)$ we can achieve good accuracy economically.

Another choice which is sometimes advocated is a generalisation of that discussed in Section 2.5. It involves eliminating the second term in (2.84) completely by relating the choice of θ to that of Δt and Δx so that

$$\theta = \tfrac{1}{2} - (\Delta x)^2/12\Delta t, \quad (2.86)$$

i.e.,

$$\mu = \frac{1}{6(1-2\theta)}, \quad (2.87)$$

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 (2.79) 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

$$T_j^{n+1/2} = -\tfrac{1}{12}\left[(\Delta t)^2 u_{ttt} + \tfrac{1}{20}(\Delta x)^4 u_{xxxxxx}\right]_j^{n+1/2} + \cdots \quad (2.88)$$
$$(\text{when } \theta = \tfrac{1}{2} - \tfrac{1}{12\mu}),$$

which is $O\left((\Delta t)^2 + (\Delta x)^4\right)$. 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. However the two-time-level, three-space-point schemes of (2.75) are by far the most widely used in practice, although the best choice of the parameter θ varies from problem to problem. Even for a given problem there may not be general agreement as to which scheme is the best. In the next section we consider the convergence analysis of these more general methods: but first we give

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

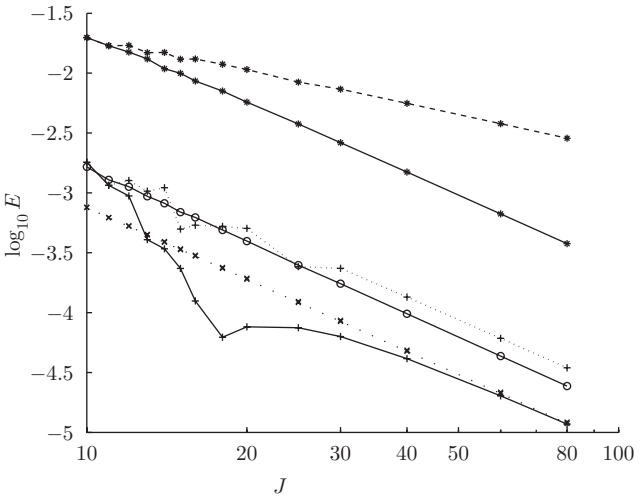

Fig. 2.7. Maximum error on $[0,1] \times [0.1,1]$ plotted against J, for various schemes.

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 :	$\theta = 1$, $\mu = 5$	—∗—∗—∗—
	$\theta = 1$, $\nu = \frac{1}{2}$	- - - ∗ - - - ∗ - - - ∗ - - -

results for the problem of (2.49) obtained with implicit methods, which show similar behaviour to those of Fig. 2.4 obtained with the explicit method, with the Crank–Nicolson method being particularly accurate. In the set of graphs in Fig. 2.7 the maximum error E is plotted against the number of mesh points J for various schemes: to eliminate transient behaviour for small t we have used

$$E := \max\left\{\left|e_j^n\right|, (x_j, t_n) \in [0,1] \times [0.1,1]\right\}.$$

We start with $J = 10$; for each implicit scheme we show a graph with fixed $\mu = \Delta t/(\Delta x)^2$ as a solid line, and also a graph with fixed $\nu = \Delta t/\Delta x$ as a dotted line; note that in the latter case the number of time steps that are needed increases much more slowly. The values of μ

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

and ν are chosen so that they give the same value of Δt when $J = 10$; this requires that $\mu = 10\nu$. For the explicit scheme there is just one graph, for $\mu = \frac{1}{2}$, the largest possible value for a stable result.

Plot A, for the explicit scheme, shows the expected $O(\Delta t) = O\big(J^{-2}\big)$ behaviour; so too does Plot B, for the Crank–Nicolson scheme with $\mu = \frac{1}{2}$. In the expression (2.85) for the truncation error of the Crank–Nicolson scheme we see that when μ is kept constant the second term is negligible compared with the first, so the two plots just differ by the fixed ratio of $\frac{1}{2}$ between their truncation errors. Also, as we shall see in the next section a maximum principle applies to both schemes so their overall behaviour is very similar. For Plot B, with $\nu = \frac{1}{20}$ kept constant, the two terms are both of order $O\big((\Delta x)^2\big)$, but the second term is numerically much smaller than the first; this accounts for the fact that the two lines in Plot B are indistinguishable.

Plot C also shows the Crank–Nicolson scheme, but here $\mu = 5$ is much larger. The numerical solution has oscillatory behaviour for small t, and the two graphs in Plot C are therefore much more erratic, not settling to their expected behaviour until J is about 40. For J larger than this the two graphs with $\mu = \frac{1}{2}$ and $\mu = 5$ are close together, illustrating the fact that the leading term in the truncation error in (2.85) is independent of μ. However, when $\nu = \frac{1}{2}$ is constant, the second term in (2.85) is a good deal larger than the first, and when $J > 40$ this graph lies well above the corresponding line in Plot B. Further analysis in Section 5.8 will help to explain this behaviour.

For the fully implicit method (plot D), where the maximum principle will apply again, the results are poor but as expected: with $\mu = 5$ we have $O(\Delta t) = O\big(J^{-2}\big)$ behaviour; and with $\Delta t/\Delta x = \frac{1}{2}$ we get only $O(\Delta t) = O(J^{-1})$ error reduction.

These graphs do not give a true picture of the relative effectiveness of the various θ-methods because they do not take account of the work involved in each calculation. So in the graphs in Fig. 2.8 the same results are plotted against a measure of the computational effort involved in each calculation: for each method this should be roughly proportional to the total number of mesh points $(\Delta x\, \Delta t)^{-1}$, with the explicit method requiring approximately half the effort of the implicit methods. The two lines in Plot B are no longer the same: when J increases with fixed ν the time step Δt decreases more slowly than when μ is fixed, so less computational work is required. These graphs show that, for this problem, the Crank–Nicolson method with $\nu = \frac{1}{2}$ is the most efficient of those tested, provided that J is taken large enough to remove the initial

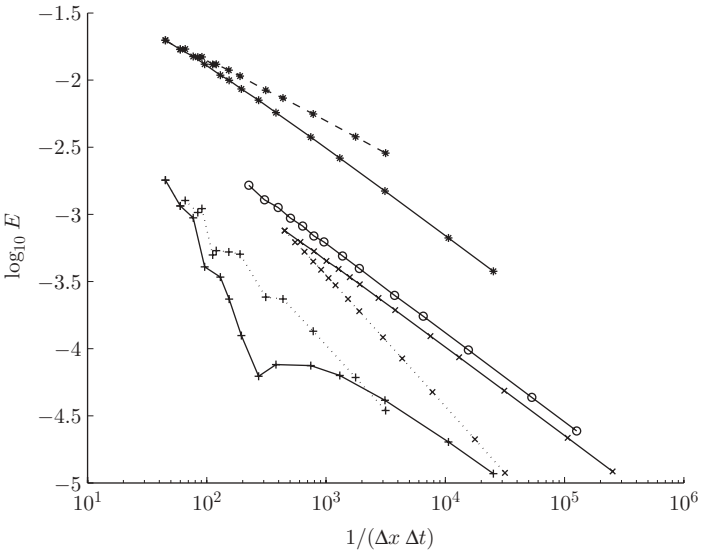

Fig. 2.8. Maximum error on $[0,1] \times [0.1,1]$ plotted against the total number of mesh points for various schemes.

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 :	$\theta = 1$, $\mu = 5$	—∗—∗—∗—
	$\theta = 1$, $\nu = \frac{1}{2}$	- - - ∗ - - - ∗ - - - ∗ - - -

oscillations; but a comparison with the $\nu = \frac{1}{20}$ plot suggests alternative choices of ν might be better.

2.11 A maximum principle and convergence for $\mu(1-\theta) \le \frac{1}{2}$

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 initial data and the values on the boundary up to time t. Such a principle also lay

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

behind the proof of convergence for the explicit scheme in Section 2.6: 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 2.2 *The θ-method of (2.75) with $0 \le \theta \le 1$ and $\mu(1-\theta) \le \frac{1}{2}$ yields $\{U_j^n\}$ satisfying*

$$U_{min} \le U_j^n \le U_{max} \tag{2.89}$$

where

$$U_{min} := \min\left\{U_0^m,\ 0 \le m \le n;\ U_j^0,\ 0 \le j \le J;\ U_J^m,\ 0 \le m \le n\right\}, \tag{2.90}$$

and

$$U_{max} := \max\left\{U_0^m,\ 0 \le m \le n;\ U_j^0,\ 0 \le j \le J;\ U_J^m,\ 0 \le m \le n\right\}. \tag{2.91}$$

For any refinement path which eventually satisfies this stability condition, the approximations given by (2.75) 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 (2.75) in the form

$$\begin{aligned}(1+2\theta\mu)U_j^{n+1} = \theta\mu\left(U_{j-1}^{n+1} + U_{j+1}^{n+1}\right) + (1-\theta)\mu\left(U_{j-1}^n + U_{j+1}^n\right) \\ + \left[1 - 2(1-\theta)\mu\right]U_j^n.\end{aligned} \tag{2.92}$$

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 (2.92). 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 (2.92). 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.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

By the definition of truncation error (see (2.84)), the solution of the differential equation satisfies the same relation as (2.92) 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

$$(1+2\theta\mu)e_j^{n+1} = \theta\mu\left(e_{j-1}^{n+1} + e_{j+1}^{n+1}\right) + (1-\theta)\mu\left(e_{j-1}^{n} + e_{j+1}^{n}\right) + \left[1 - 2(1-\theta)\mu\right] e_j^n - \Delta t T_j^{n+1/2} \tag{2.93}$$

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, as in Section 2.6,

$$E^n := \max_{0\le j\le J} \left|e_j^n\right|, \quad T^{n+1/2} := \max_{1\le j\le J-1} \left|T_j^{n+1/2}\right|. \tag{2.94}$$

Because of the nonnegative coefficients, it follows that

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

and hence that

$$E^{n+1} \le E^n + \Delta t\, T^{n+1/2} \tag{2.95}$$

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

$$\begin{aligned} E^n &\le \Delta t \sum_0^{n-1} T^{m+1/2}, \\ &\le n\Delta t \max_m T^{m+1/2} \end{aligned} \tag{2.96}$$

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 ϵ_j^0, ϵ_0^m and ϵ_J^m with $0 \le j \le J$ and $0 \le m \le N$, say. Then the errors e_j^n satisfy the recurrence relation (2.93) with initial and boundary values

$$\begin{aligned} e_j^0 &= \epsilon_j^0, \quad j = 0, 1, \ldots, n, \\ e_0^m = \epsilon_0^m,\ e_J^m &= \epsilon_J^m, \quad 0 \le m \le N. \end{aligned}$$

Then (by Duhamel's principle) e_j^N can be written as the sum of two terms. The first term satisfies (2.93) with zero initial and boundary values; this term is bounded by (2.96). The second term satisfies the homogeneous form of (2.93), with the term in T omitted, and with the

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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. □

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. In view of this large gap the reader may wonder about the sharpness of this theorem. In fact, the maximum principle condition is sharp, but a little severe: for with $J = 2$ and $U_0^0 = U_2^0 = 0$, $U_1^0 = 1$, one obtains $U_1^1 = 1 - 2(1-\theta)\mu$ which is nonnegative only if the given condition is satisfied; but, of course, one would use larger values of J in practice and this would relax the condition a little (see Exercise 11). Moreover, if with $U_0^n = U_J^n = 0$ one wants to have

$$\left|U_j^n\right| \leq K \max_{0 \leq i \leq J} \left|U_i^0\right| \qquad \forall j, n \tag{2.97}$$

satisfied with $K = 1$, which is the property needed to deduce the error bound (2.96), it has recently been shown[1] that it is necessary and sufficient that $\mu(1-\theta) \leq \frac{1}{4}(2-\theta)/(1-\theta)$, giving $\mu \leq \frac{3}{2}$ for the Crank–Nicolson scheme. It is only when any value of K is accepted in this growth bound, which is all that is required in the stability definition of (2.55), that the weaker condition $\mu(1-2\theta) \leq \frac{1}{2}$ is adequate. Then for Crank–Nicolson one can actually show that $K \leq 23$ holds!

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 (see below in Section 2.15); but, as we see above, it is easy to derive only sufficient stability conditions.

These points are illustrated in Fig. 2.9. 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

[1] See Kraaijevanger, J.F.B.M. (1992) Maximum norm contractivity of discretization schemes for the heat equation. *Appl. Numer. Math.* **99**, 475–92.

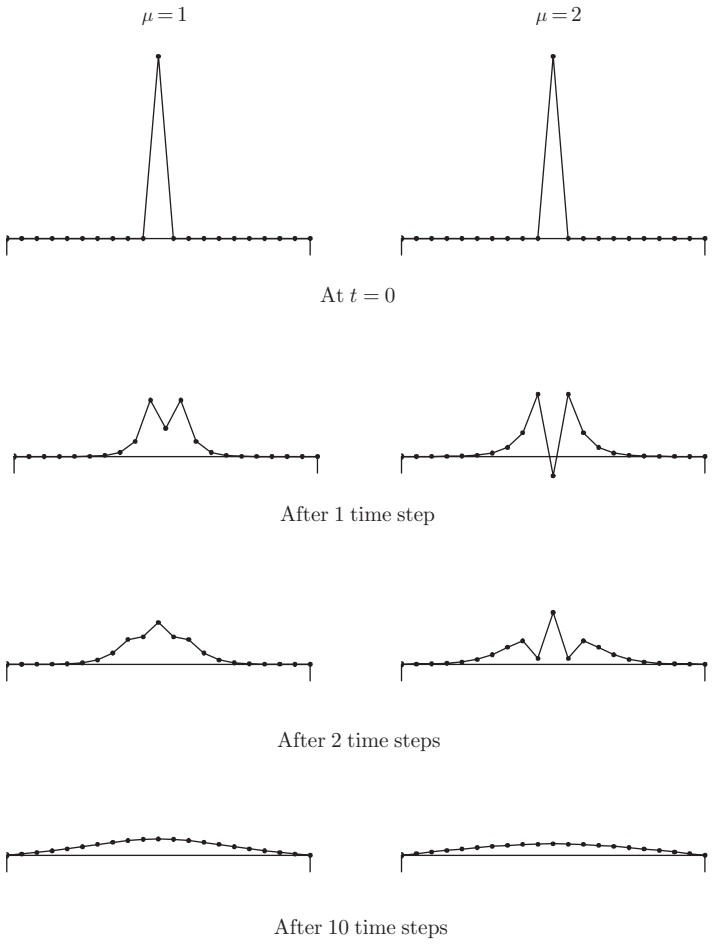

Fig. 2.9. The Crank–Nicolson method applied to the heat equation where the initial distribution has a sharp spike at the mid-point; $J = 20$, $\Delta x = 0.05$.

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. 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

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

numerical solution shows two peaks, one each side of the mid-point; the exact solution of the problem 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.

2.12 A three-time-level scheme

We have seen how the Crank–Nicolson scheme improves on the accuracy of the explicit scheme by the use of symmetry in the time direction to remove the even time derivative terms in the truncation error. This improvement has to be balanced against the extra complication involved in the use of an implicit method. This suggests investigation of the possibility of using more than two time levels to improve accuracy, while retaining the efficiency of an explicit method.

Consider, for example, the use of a symmetric central difference for the time derivative, leading to the explicit three-level scheme

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{(\Delta x)^2}. \tag{2.98}$$

It is easy to see that the truncation error of this approximation involves only even powers of both Δx and Δt, and hence has order $O\big((\Delta x)^2 + (\Delta t)^2\big)$. However, if we investigate the stability of the scheme we find a solution of the usual form (2.72) provided that

$$\frac{\lambda - 1/\lambda}{2\Delta t} = \frac{-4\,\sin^2 \tfrac{1}{2}k\Delta x}{(\Delta x)^2} \tag{2.99}$$

or

$$\lambda^2 + 8\lambda\mu\,\sin^2 \tfrac{1}{2}k\Delta x - 1 = 0. \tag{2.100}$$

This quadratic equation for λ has two roots, giving two solution modes for each value of k. The roots are both real with a negative sum, and their product is -1. Hence one of them has magnitude greater than unity, giving an unstable mode. The scheme is therefore useless in practice since it is always unstable, for every value of μ.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

This result does not, of course, mean that every three-level explicit scheme is always unstable. We leave as an exercise the proof that the scheme

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} = \frac{\theta\delta_x^2 U_j^n + (1-\theta)\delta_x^2 U_j^{n-1}}{(\Delta x)^2} \tag{2.101}$$

has both solution modes satisfying $|\lambda| \leq 1$ if $\theta \leq \frac{1}{2}$ and $4(1-\theta)\mu \leq 1$ (see also Exercise 6); but then this stability restriction is as bad as that for our first simple scheme.

2.13 More general boundary conditions

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

$$\frac{\partial u}{\partial x} = \alpha(t)u + g(t), \quad \alpha(t) \geq 0. \tag{2.102}$$

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

$$\frac{U_1^n - U_0^n}{\Delta x} = \alpha^n U_0^n + g^n \tag{2.103}$$

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

$$U_0^n = \beta^n U_1^n - \beta^n g^n \Delta x, \tag{2.104a}$$

where

$$\beta^n = \frac{1}{1 + \alpha^n \Delta x}. \tag{2.104b}$$

Then we can apply the θ-method (2.75) in just the same way as for Dirichlet boundary conditions. We need to solve the usual tridiagonal system of linear equations, but now this is a system of J equations in the J unknowns, namely the interior values at the new time step and the value at the left-hand boundary. Equation (2.104a) is then the first equation of the system, and it is clear that the augmented system is still tridiagonal and, because we have assumed $\alpha(t) \geq 0$ and therefore $0 < \beta^n \leq 1$, the coefficients still satisfy the conditions of (2.67) and (2.68), except that when $\alpha(t) = 0$ we have $b_0 = a_0$ and $e_0 = 1$.

To consider the accuracy and stability of the resulting scheme, we need to concentrate attention on the first interior point. We can use

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

(2.104a) to eliminate U_0^n; the second difference at the first interior point then has the form

$$\delta_x^2 U_1^n = U_2^n - (2-\beta^n)U_1^n - \beta^n g^n \Delta x. \tag{2.105}$$

Thus with the usual definition of truncation error, after some manipulation, the global error can be shown to satisfy, instead of (2.93), the new relation

$$\begin{aligned}\left[1+\theta\mu(2-\beta^{n+1})\right] e_1^{n+1} = {} & \left[1-(1-\theta)\mu(2-\beta^n)\right] e_1^n \\ & + \theta\mu e_2^{n+1} + (1-\theta)\mu e_2^n \\ & - \Delta t T_1^{n+1/2}.\end{aligned} \tag{2.106}$$

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 (2.106) 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

$$\theta(1-\beta^{n+1}) \ge -(1-\theta)(1-\beta^n), \tag{2.107}$$

which again is always satisfied if $\alpha(t) \ge 0$. Hence we can deduce the bound (2.96) 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: to assume otherwise would correspond to having heat inflow rather than outflow in proportion to surface temperature, which would lead to an exponentially increasing solution. This is very unlikely to occur in any real problem, and would in any case lead to a problem which is not well-posed.

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 (2.103) as applying the boundary condition at $(0, t_n)$ and expand about this point for the exact solution to obtain

$$\frac{u_1^n - u_0^n}{\Delta x} - \alpha^n u_0^n - g^n = \left[\tfrac{1}{2}\Delta x\, u_{xx} + \tfrac{1}{6}(\Delta x)^2 u_{xxx} + \cdots\right]_0^n. \tag{2.108}$$

Then we write the truncation error in the following form, in which an appropriate multiple of the approximation (2.103) to the boundary

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

condition is added to the difference equation in order to cancel the terms in u_0^n,

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

to obtain

$$T_1^{n+1/2} \approx -\tfrac{1}{2}\beta^n u_{xx}. \tag{2.109}$$

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.

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

$$\frac{U_1^n - U_0^n}{\Delta x} = \tfrac{1}{2}\alpha^n\left(U_0^n + U_1^n\right) + g^n, \tag{2.110}$$

$$U_0^n = \frac{1-\frac{1}{2}\alpha^n\Delta x}{1+\frac{1}{2}\alpha^n\Delta x}U_1^n - \frac{\Delta x}{1+\frac{1}{2}\alpha^n\Delta x}g^n. \tag{2.111}$$

Then (2.108) is replaced by an expansion about $j = \frac{1}{2}$, giving

$$\begin{aligned} &\frac{u_1^n - u_0^n}{\Delta x} - \tfrac{1}{2}\alpha^n\left(u_0^n + u_1^n\right) - g^n = \\ &\qquad \left[\tfrac{1}{24}(\Delta x)^2 u_{xxx} - \tfrac{1}{8}\alpha^n(\Delta x)^2 u_{xx} + \cdots\right]_{1/2}^n \end{aligned} \tag{2.112}$$

and hence

$$\begin{aligned} T^{n+1/2} &= \left[\tfrac{1}{2}\Delta t u_{tt} - \tfrac{1}{12}(\Delta x)^2 u_{xxxx} + \ldots\right]_1^n \\ &\quad - \frac{1}{1+\frac{1}{2}\alpha^n\Delta x}\left[\tfrac{1}{24}\Delta x\left(u_{xxx} - 3\alpha^n u_{xx}\right) + \cdots\right]_0^n \\ &= O(\Delta x). \end{aligned} \tag{2.113}$$

Only minor modifications are necessary to (2.106) 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 $O\big((\Delta x)^2\big)$ despite this $O(\Delta x)$ truncation error near the boundary.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$\frac{U^n_1 - U^n_{-1}}{2\Delta x} = \alpha^n U^n_0 + g^n. \tag{2.114}$$

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

$$\begin{aligned}\frac{U^{n+1}_0 - U^n_0}{\Delta t} &- \frac{\delta_x^2}{(\Delta x)^2}\left[\theta U^{n+1}_0 + (1-\theta)U^n_0\right] \\ &- \frac{2\theta}{\Delta x}\left[\frac{U^{n+1}_1 - U^{n+1}_{-1}}{2\Delta x} - \alpha^{n+1}U^{n+1}_0 - g^{n+1}\right] \\ &- \frac{2(1-\theta)}{\Delta x}\left[\frac{U^n_1 - U^n_{-1}}{2\Delta x} - \alpha^n U^n_0 - g^n\right] = 0. \end{aligned} \tag{2.115}$$

Clearly for the truncation error we pick up terms like

$$\frac{2\theta}{\Delta x}\left[\frac{u^{n+1}_1 - u^{n+1}_{-1}}{2\Delta x} - \alpha^{n+1}u^{n+1}_0 - g^{n+1}\right] = \theta\left[\tfrac{1}{3}\Delta x u_{xxx}\right]_0^{n+1} + \cdots \tag{2.116}$$

to add to the usual truncation error terms. If we rewrite (2.115) in the form

$$\begin{aligned}\left[1 + 2\theta\mu\left(1 + \alpha^{n+1}\Delta x\right)\right]U^{n+1}_0 &= \left[1 - 2(1-\theta)\mu\left(1 + \alpha^n\Delta x\right)\right]U^n_0 \\ &\quad + 2\theta\mu U^{n+1}_1 \\ &\quad - 2\mu\Delta x\left[\theta g^{n+1} + (1-\theta)g^n\right] \end{aligned} \tag{2.117}$$

we also see that the error analysis based on a maximum principle still holds with only the slight strengthening of condition needed in Theorem 2.2 to

$$\mu(1-\theta)\left(1 + \alpha^n\Delta x\right) \le \tfrac{1}{2}. \tag{2.118}$$

In this section we have considered the solution of the heat equation with a derivative boundary condition at the left-hand end, and a Dirichlet condition at the other. The same idea can be applied to a problem with a derivative condition at the right-hand end, or at both ends.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

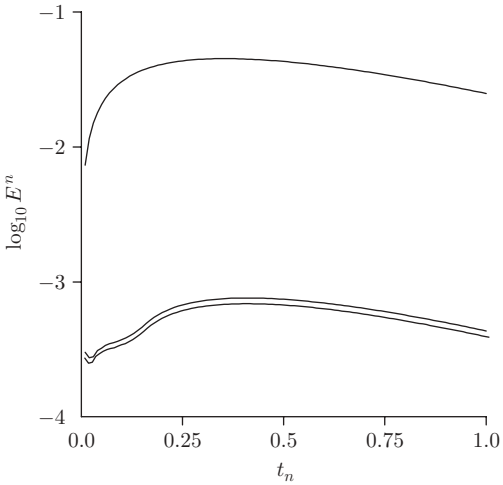

Fig. 2.10. The effect of a Neumann boundary condition approximation on the error for the Crank–Nicolson scheme with $J = 10$, $\Delta x = 0.1$; the top curve is for (2.103) and the lower two for (2.114) and (2.110).

On carrying through the similar analysis, we soon discover why the condition at $x = 1$ must be of the form

$$\frac{\partial u}{\partial x} = \beta(t)u + g(t), \quad \beta(t) \leq 0. \tag{2.119}$$

As an illustration of different methods of treating the boundary condition, we compute solutions to the problem $u_t = u_{xx}$ on $0 < x < 1$, with initial condition $u(x, 0) = 1 - x^2$, and boundary conditions $u_x(0, t) = 0, u(1, t) = 0$, giving a Neumann condition at the left-hand end, and a Dirichlet condition at the right-hand end. We use the Crank–Nicolson method, with $J = 10$ and $\mu = 1$, so that $\Delta t = 0.01$. The maximum error in the numerical solution, as a function of t_n, is shown in Fig. 2.10 for the three methods described above: namely, the use of the forward difference approximation to $u_x(0, t)$, the use of the central difference incorporating the fictitious value U^n_{-1}, and the placing of the boundary half-way between the first two mesh points. The numerical results from the second and third of these methods are very similar, but show a quite dramatic difference from those for the first method; the error in this case is some 50 times larger.

Notice also that the maximum error in each method increases with n for part of the range, before beginning to decrease again, rather slowly; compare the behaviour with that in Fig. 2.4 where Dirichlet conditions

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

were applied at each boundary. In the next section we consider Neumann conditions at each boundary.

2.14 Heat conservation properties

Suppose that in our model heat flow problem $u_t = u_{xx}$ we define the total heat in the system at time t by

$$h(t) = \int_0^1 u(x,t)\,\mathrm{d}x. \tag{2.120}$$

Then from the differential equation we have

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \int_0^1 u_t\,\mathrm{d}x = \int_0^1 u_{xx}\,\mathrm{d}x = \big[u_x\big]_0^1. \tag{2.121}$$

This is not very helpful if we have Dirichlet boundary conditions: but suppose we are given Neumann boundary conditions at each end; say, $u_x(0,t) = g_0(t)$ and $u_x(1,t) = g_1(t)$. Then we have

$$\frac{\mathrm{d}h}{\mathrm{d}t} = g_1(t) - g_0(t), \tag{2.122}$$

so that h is given by integrating an ordinary differential equation.

Now suppose we carry out a similar manipulation for the θ-method equations (2.75), introducing the total heat by means of a summation over the points for which (2.75) holds:

$$H^n = \sum_1^{J-1} \Delta x U_j^n. \tag{2.123}$$

Then, recalling from the definitions of the finite difference notation that $\delta_x^2 U_j = \Delta_{+x} U_j - \Delta_{+x} U_{j-1}$, we have

$$\begin{aligned} H^{n+1} - H^n &= \frac{\Delta t}{\Delta x} \sum_1^{J-1} \delta_x^2 \left[\theta U_j^{n+1} + (1-\theta) U_j^n\right] \\ &= \frac{\Delta t}{\Delta x} \Big\{ \Delta_{+x} \left[\theta U_{J-1}^{n+1} + (1-\theta) U_{J-1}^n\right] \\ &\qquad - \Delta_{+x} \left[\theta U_0^{n+1} + (1-\theta) U_0^n\right] \Big\}. \end{aligned} \tag{2.124}$$

The rest of the analysis will depend on how the boundary condition is approximated. Consider the simplest case as in (2.103): namely we set $U_1^n - U_0^n = \Delta x\, g_0^n$, $U_J^n - U_{J-1}^n = \Delta x\, g_1^n$. Then we obtain

$$H^{n+1} - H^n = \Delta t \left[\theta \left(g_1^{n+1} - g_0^{n+1}\right) + (1-\theta)\left(g_1^n - g_0^n\right)\right] \tag{2.125}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

as an approximation to (2.122); this approximation may be very accurate, even though we have seen that U^n may not give a good pointwise approximation to u^n. In particular, if g_0 and g_1 are independent of t the change in H in one time step exactly equals that in $h(t)$. How should we interpret this?

Clearly to make the most of this matching we should relate (2.123) as closely as possible to (2.120). If u and U were constants that would suggest we take $(J-1)\Delta x = 1$, rather than $J\Delta x = 1$ as we have been assuming; and we should compare U_j^n with

$$u_j^n := \frac{1}{\Delta x}\int_{(j-1)\Delta x}^{j\Delta x} u(x,t_n)\,\mathrm{d}x, \quad j = 1,2,\ldots,J-1, \tag{2.126}$$

so that it is centred at $(j-\frac{1}{2})\Delta x$ and we have

$$h(t_n) = \sum_1^{J-1} \Delta x\, u_j^n. \tag{2.127}$$

Note that this interpretation matches very closely the scheme that we were led to in (2.110)–(2.113) by analysing the truncation error. It would also mean that for initial condition we should take $U_j^0 = u_j^0$ as defined by (2.126). Then for time-independent boundary conditions we have $H^n = h(t_n)$ for all n. Moreover, it is easy to see that the function

$$\hat{u}(x,t) := (g_1 - g_0)\,t + \tfrac{1}{2}(g_1 - g_0)\,x^2 + g_0 x + C \tag{2.128}$$

with any constant C satisfies the differential equation, and the two boundary conditions. It can also be shown that the exact solution of our problem, with any given initial condition, will tend to such a solution as $t \to \infty$. Since the function (2.128) is linear in t and quadratic in x it will also satisfy the finite difference equations exactly; hence the error in a numerical solution produced and interpreted in this way will decrease to zero as t increases. As we have seen above, the usual finite difference approximations, with Neumann boundary conditions and the usual interpretation of the errors, may be expected to give errors which initially increase with n, and then damp only very slowly.

These observations are illustrated by a solution of the heat equation with homogeneous Neumann boundary conditions $u_x = 0$ at $x = 0$ and $x = 1$, and with initial value $u(x,0) = 1 - x^2$. Fig. 2.11 shows the maximum error as a function of t_n for three cases, each using $J = 10$ for the Crank–Nicolson method with $\mu = \frac{1}{2}$. The top curve corresponds to using the Neumann conditions (2.103) as above and interpreting the

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

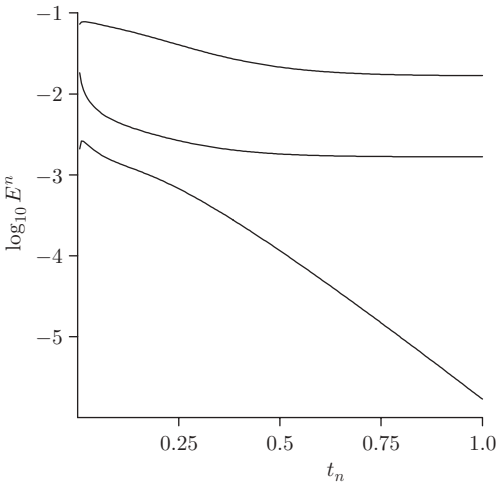

Fig. 2.11. Effect of error interpretation for a pure Neumann problem: the top curve corresponds to boundary conditions (2.103) and the second to (2.114), both with the usual initial data and definition of E^n; the bottom curve is computed as for the top curve but with initial data from (2.126) and the error reinterpreted through the heat conservation principle.

error E^n in the usual way, with $\Delta x = 0.1,\ \Delta t = 0.005$. The second curve is the same apart from using the approximaton (2.114) for the boundary conditions. Clearly both give a substantial residual error. However, the bottom curve corresponds to the same method as the top but with the initial data obtained from the u_j^n of (2.126), and the error reinterpreted to reflect the heat conservation principle as described above: namely, $E^n := \max\{|U_j^n - u_j^n|,\quad j = 0, 1, 2, \ldots, J\}$; and $t_n = n\Delta t = n\mu(\Delta x)^2 = \frac{1}{2}n(J-1)^{-2}$. We see clearly the decay of the error as predicted by the argument given above.

2.15 More general linear problems

The form of the heat equation which we have considered so far corresponds to a physical situation where all the physical properties of the material are constant in time, and independent of x. More generally these properties may be functions of x, or t, or both. In particular, a dependence on x is often used to model the nearly one-dimensional flow of heat in a thin bar whose cross-sectional area depends on x. We shall therefore examine briefly how the methods so far discussed may be adapted to problems of increasing generality.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

First of all, consider the problem

$$\frac{\partial u}{\partial t} = b(x,t)\frac{\partial^2 u}{\partial x^2}, \tag{2.129}$$

where the function $b(x,t)$ is, as usual, assumed to be strictly positive. Then the explicit scheme (2.19) is extended in an obvious way to give

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{(\Delta x)^2} b_j^n (U_{j+1}^n - 2U_j^n + U_{j-1}^n), \tag{2.130}$$

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 (2.19), to the expression

$$T(x,t) = \tfrac{1}{2}\Delta t\, u_{tt} - \tfrac{1}{12} b(x,t)(\Delta x)^2 u_{xxxx} + \cdots. \tag{2.131}$$

The analysis leading to (2.44) still applies, but the stability condition has to be replaced by

$$\frac{\Delta t}{(\Delta x)^2} b(x,t) \le \tfrac{1}{2} \tag{2.132}$$

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

$$E^n \le \tfrac{1}{2}\Delta t \left[M_{tt} + \tfrac{B(\Delta x)^2}{6\Delta t}\, M_{xxxx} \right] t_F \tag{2.133}$$

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

The θ-method can be applied to this more general problem in several slightly different ways. Evidently equation (2.75) can be generalised to

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{(\Delta x)^2} b^* \left[\theta \delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n \right], \tag{2.134}$$

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

$$b^* = b_j^{n+1/2}; \tag{2.135}$$

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 (2.84), which becomes

$$\begin{aligned} T_j^{n+1/2} = [&(\tfrac{1}{2} - \theta)\Delta t\, u_{xxt} - \tfrac{b}{12}(\Delta x)^2 u_{xxxx} + \tfrac{1}{24}(\Delta t)^2 u_{ttt} \\ &- \tfrac{b}{8}(\Delta t)^2 u_{xxtt} + \tfrac{1}{12}(\tfrac{1}{2} - \theta)\Delta t (\Delta x)^2 u_{xxxxt} \\ &- \tfrac{2b}{6!}(\Delta x)^4 u_{xxxxxx} + \cdots]_j^{n+1/2}. \end{aligned} \tag{2.136}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$\frac{\Delta t}{(\Delta x)^2}(1-\theta)b(x,t) \leq \tfrac{1}{2} \tag{2.137}$$

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

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

$$b^* = \tfrac{1}{2}(b_j^{n+1} + b_j^n). \tag{2.138}$$

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

$$b^* = \left[b + \tfrac{1}{4}(\Delta t)^2 b_{tt} + \cdots\right]_j^{n+1/2} \tag{2.139}$$

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

The most general form of the linear parabolic equation is

$$\frac{\partial u}{\partial t} = b(x,t)\frac{\partial^2 u}{\partial x^2} - a(x,t)\frac{\partial u}{\partial x} + c(x,t)u + d(x,t), \tag{2.140}$$

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, and specifically in (5.48) of Section 5.7. 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

$$\begin{aligned}\frac{U_j^{n+1} - U_j^n}{\Delta t} = {} & \frac{b_j^n}{(\Delta x)^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \\ & - \frac{a_j^n}{2\Delta x}(U_{j+1}^n - U_{j-1}^n) + c_j^n U_j^n + d_j^n. \end{aligned} \tag{2.141}$$

The calculation of the leading terms in the truncation error is straightforward, and is left as an exercise. However, a new difficulty arises in

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

the analysis of the behaviour of the error e_j^n. Just as in the analysis of the simpler problem, which led to (2.42), we find that

$$\begin{aligned} e_j^{n+1} &= e_j^n + \mu_j^n\,(e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \tfrac{1}{2}\nu_j^n(e_{j+1}^n - e_{j-1}^n) \\ &\quad + \Delta t c_j^n e_j^n - \Delta t T_j^n \\ &= (1 - 2\mu_j^n + \Delta t c_j^n)e_j^n \\ &\quad + (\mu_j^n - \tfrac{1}{2}\nu_j^n)e_{j+1}^n + (\mu_j^n + \tfrac{1}{2}\nu_j^n)e_{j-1}^n - \Delta t T_j^n, \end{aligned} \tag{2.142}$$

where we have written

$$\mu_j^n = \frac{\Delta t}{(\Delta x)^2} b_j^n, \qquad \nu_j^n = \frac{\Delta t}{\Delta x} a_j^n. \tag{2.143}$$

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:

$$\tfrac{1}{2}|\nu_j^n| \le \mu_j^n, \tag{2.144}$$

$$2\mu_j^n - \Delta t c_j^n \le 1, \tag{2.145}$$

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 (2.145) would represent a slight relaxation of the condition on μ, but then one can only establish $E^{n+1} \le (1 + C\Delta t)E^n + T\Delta t$. 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

$$\Delta x \le \frac{2b_j^n}{|a_j^n|}, \quad \text{or} \quad \frac{|a_j^n|\,\Delta x}{b_j^n} \le 2, \tag{2.146}$$

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.

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 (2.146) imposes a limit of 2

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{b_j^n}{(\Delta x)^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) - \frac{a_j^n}{\Delta x}(U_j^n - U_{j-1}^n) + c_j^n U_j^n + d_j^n, \tag{2.147}$$

which leads to

$$\begin{aligned} e_j^{n+1} &= e_j^n + \mu_j^n(e_{j+1}^n - 2e_j^n + e_{j-1}^n) - \nu_j^n(e_j^n - e_{j-1}^n) - \Delta t T_j^n \\ &= (1 - 2\mu_j^n - \nu_j^n)e_j^n \\ &\quad + \mu_j^n e_{j+1}^n + (\mu_j^n + \nu_j^n)e_{j-1}^n - \Delta t T_j^n. \end{aligned} \tag{2.148}$$

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

$$2\mu_j^n + \nu_j^n \leq 1. \tag{2.149}$$

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.

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

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(p(x,t)\frac{\partial u}{\partial x}\right) \tag{2.150}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$\frac{\partial u}{\partial t} = p\frac{\partial^2 u}{\partial x^2} + \frac{\partial p}{\partial x}\frac{\partial u}{\partial x}, \tag{2.151}$$

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

$$\left[p\frac{\partial u}{\partial x}\right]^n_{j+1/2} \approx p^n_{j+1/2}\left(\frac{u^n_{j+1} - u^n_j}{\Delta x}\right), \tag{2.152}$$

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

$$\frac{U^{n+1}_j - U^n_j}{\Delta t} = \frac{1}{(\Delta x)^2}\left[p^n_{j+1/2}(U^n_{j+1} - U^n_j) - p^n_{j-1/2}(U^n_j - U^n_{j-1})\right]. \tag{2.153}$$

We will write

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

which gives in explicit form

$$U^{n+1}_j = \left(1 - \mu'(p^n_{j+1/2} + p^n_{j-1/2})\right)U^n_j + \mu' p^n_{j+1/2}U^n_{j+1} + \mu' p^n_{j-1/2}U^n_{j-1}. \tag{2.154}$$

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

$$\mu' P \leq \tfrac{1}{2}, \tag{2.155}$$

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).

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

2.16 Polar co-ordinates

One-dimensional problems often result from physical systems in three dimensions which have cylindrical or spherical symmetry. In polar co-ordinates the simple heat equation becomes

$$\frac{\partial u}{\partial t} = \frac{1}{r^\alpha}\frac{\partial}{\partial r}\left(r^\alpha \frac{\partial u}{\partial r}\right) \tag{2.156}$$

or

$$u_t = u_{rr} + \frac{\alpha}{r}u_r, \tag{2.157}$$

where $\alpha = 0$ reduces to the case of plane symmetry which we have considered so far, while $\alpha = 1$ corresponds to cylindrical symmetry and $\alpha = 2$ to spherical symmetry. The methods just described can easily be applied to this equation, either in the form (2.156), or in the form (2.157). Examination of the stability restrictions in the two cases shows that there is not much to choose between them in this particular situation. However, in each case there is clearly a problem at the origin $r = 0$.

A consideration of the symmetry of the solution, in either two or three dimensions, shows that $\partial u/\partial r = 0$ at the origin; alternatively, (2.157) shows that either u_{rr} or u_t, or both, would be infinite at $r = 0$, were u_r non-zero. Now keep t constant, treating u as a function of r only, and expand in a Taylor series around $r = 0$, giving

$$\begin{aligned} u(r) &= u(0) + ru_r(0) + \tfrac{1}{2}r^2u_{rr}(0) + \cdots \\ &= u(0) + \tfrac{1}{2}r^2u_{rr}(0) + \cdots \end{aligned} \tag{2.158}$$

and

$$\begin{aligned} \frac{1}{r^\alpha}\frac{\partial}{\partial r}\left(r^\alpha\frac{\partial u}{\partial r}\right) &= \frac{1}{r^\alpha}\frac{\partial}{\partial r}\left[r^\alpha u_r(0) + r^{\alpha+1}u_{rr}(0) + \cdots\right] \\ &= \frac{1}{r^\alpha}\left[(\alpha+1)r^\alpha u_{rr}(0) + \cdots\right] \\ &= (\alpha+1)u_{rr}(0) + \cdots . \end{aligned} \tag{2.159}$$

Writing Δr for r in (2.158) we get

$$u(\Delta r) - u(0) = \tfrac{1}{2}(\Delta r)^2 u_{rr}(0) + \cdots \tag{2.160}$$

and we thus obtain a difference approximation to be used at the left end of the domain,

$$\frac{U_0^{n+1} - U_0^n}{\Delta t} = \frac{2(\alpha+1)}{(\Delta r)^2}(U_1^n - U_0^n). \tag{2.161}$$

This would also allow any of the θ-methods to be applied.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

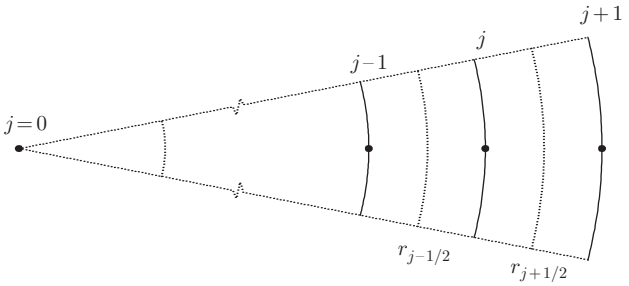

Fig. 2.12. Polar co-ordinates.

An alternative, more physical, viewpoint springs directly from the form (2.156). Consider the heat balance for an annular region between two surfaces at $r = r_{j-1/2}$ and $r = r_{j+1/2}$ as in Fig. 2.12: the term $r^\alpha \partial u/\partial r$ on the right-hand side of (2.156) is proportional to a heat flux times a surface area; and the difference between the fluxes at surfaces with radii $r_{j-1/2}$ and $r_{j+1/2}$ is applied to raising the temperature in a volume which is proportional to $(r_{j+1/2}^{\alpha+1} - r_{j-1/2}^{\alpha+1})/(\alpha+1)$. Thus on a uniform mesh of spacing Δr, a direct differencing of the right-hand side of (2.156) gives

$$\begin{aligned}\frac{\partial U_j}{\partial t} &\approx \frac{\alpha+1}{r_{j+1/2}^{\alpha+1} - r_{j-1/2}^{\alpha+1}} \delta_r \left(r_j^\alpha \frac{\delta_r U_j}{\Delta r} \right) \\ &= \frac{(\alpha+1)\left[r_{j+1/2}^\alpha U_{j+1} - \left(r_{j+1/2}^\alpha + r_{j-1/2}^\alpha \right) U_j + r_{j-1/2}^\alpha U_{j-1} \right]}{\left[r_{j+1/2}^\alpha + r_{j+1/2}^{\alpha-1} r_{j-1/2} + \cdots + r_{j-1/2}^\alpha \right] (\Delta r)^2} \\ &\qquad\qquad \text{for } j = 1, 2, \ldots . \end{aligned} \tag{2.162a}$$

At the origin where there is only one surface (a cylinder of radius $r_{1/2} = \frac{1}{2}\Delta r$ when $\alpha = 1$, a sphere of radius $r_{1/2}$ when $\alpha = 2$) one has immediately

$$\frac{\partial U_0}{\partial t} \approx \frac{\alpha+1}{r_{1/2}^{\alpha+1}} r_{1/2}^\alpha \frac{U_1 - U_0}{\Delta r} = 2(\alpha+1)\frac{U_1 - U_0}{(\Delta r)^2}, \tag{2.162b}$$

which is in agreement with (2.161). Note also that (2.162a) is identical with difference schemes obtained from either (2.156) or (2.157) in the case of cylindrical symmetry ($\alpha = 1$); but there is a difference in the spherical case because $r_{j+1/2}^2 + r_{j+1/2} r_{j-1/2} + r_{j-1/2}^2$ is not the same as $3r_j^2$.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

The form (2.162a) and (2.162b) is simplest for considering the condition that a maximum principle should hold. From calculating the coefficient of U_j^n in the θ-method, one readily deduces that the worst case occurs at the origin and leads to the condition

$$2(\alpha+1)(1-\theta)\Delta t \leq (\Delta r)^2. \tag{2.162c}$$

This becomes more restrictive as the number of space dimensions increases in a way that is consistent with what we shall see in Chapter 3.

2.17 Nonlinear problems

In the general linear equation we considered in Section 2.15 the physical properties depended on x and t. It is also very common for these properties to depend on the unknown function $u(x,t)$. This leads to the consideration of nonlinear problems.

We shall just consider one example, the equation

$$u_t = b(u)u_{xx} \tag{2.163}$$

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,

$$U_j^{n+1} = U_j^n + \mu' b(U_j^n)\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right). \tag{2.164}$$

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

$$U_j^{n+1} = U_j^n + \mu' b(U_j^n)(U_{j+1}^n - 2U_j^n + U_{j-1}^n), \tag{2.165}$$

$$u_j^{n+1} = u_j^n + \mu' b(u_j^n)(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t\, T_j^n, \tag{2.166}$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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

$$b(u_j^n) = b(U_j^n) + (u_j^n - U_j^n)\frac{\partial b}{\partial u}(\eta) \tag{2.167}$$

$$= b(U_j^n) - e_j^n q_j^n \tag{2.168}$$

where

$$q_j^n = \frac{\partial b}{\partial u}(\eta) \tag{2.169}$$

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

We can now subtract (2.166) from (2.165), and obtain

$$\begin{aligned} e_j^{n+1} = e_j^n &+ \mu' b(U_j^n)(e_{j+1}^n - 2e_j^n + e_{j-1}^n) \\ &+ \mu' e_j^n q_j^n (u_{j+1}^n - 2u_j^n + u_{j-1}^n) - \Delta t\, T_j^n. \end{aligned} \tag{2.170}$$

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

$$\Delta t\, [\max b(U_j^n)] \le \tfrac{1}{2}(\Delta x)^2. \tag{2.171}$$

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 (2.171) for all j and n, and that we have bounds

$$\left|u_{j+1} - 2u_j^n + u_{j-1}^n\right| \le M_{xx}(\Delta x)^2, \quad \left|q_j^n\right| \le K, \tag{2.172}$$

we can write

$$E^{n+1} \le [1 + KM_{xx}\Delta t]\, E^n + \Delta t\, T \tag{2.173}$$

in our previous notation. Moreover,

$$(1 + KM_{xx}\Delta t)^n \le \mathrm{e}^{KM_{xx}n\Delta t} \le \mathrm{e}^{KM_{xx}t_F} \tag{2.174}$$

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

However, although the stability condition (2.171) 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 (2.163) is rather special; equally common would be the case $u_t = (b(u)u_x)_x$, 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.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

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.

Bibliographic notes and recommended reading

The general reference for all the material on initial value problems in this book is the classic text by Richtmyer and Morton (1967). Another classic text covering the whole range of partial differential equation problems is that by Collatz (1966). In both of them many more difference schemes to approximate the problems treated in this chapter will be found.

For a wide-ranging exposition of diffusion problems and applications in which they arise, the reader is referred to the book by Crank (1975), where more discussion on the use of the Crank–Nicolson scheme may be found.

The earliest reference that we have to the important Thomas algorithm is to a report from Columbia University, New York in 1949; but since it corresponds to direct Gaussian elimination without pivoting, many researchers were undoubtedly aware of it at about this time. For a more general discussion of Gaussian elimination for banded matrices the reader is referred to standard texts on numerical analysis, several of which are listed in the Bibliography at the end of the book, or to more specialised texts on matrix computations such as that by Golub and Van Loan (1996).

A fuller discussion of nonlinear problems is given in the book by Ames (1992), where reference is made to the many examples of physical problems which are modelled by nonlinear parabolic equations contained in Ames (1965) and Ames (1972).

Exercises

2.1 (i) The function $u^0(x)$ is defined on [0,1] by

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

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

Show that

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

where $a_m = (8/m^2\pi^2) \sin \frac{1}{2}m\pi$.
(ii) Show that

$$\int_{2p}^{2p+2} \frac{1}{x^2} \mathrm{d}x > \frac{2}{(2p+1)^2}$$

and hence that

$$\sum_{p=p_0}^{\infty} \frac{1}{(2p+1)^2} < \frac{1}{4p_0}.$$

(iii) Deduce that $u^0(x)$ is approximated on the interval $[0, 1]$ to within 0.001 by the sine series in part (i) truncated after $m = 405$.

2.2 (i) Show that for every positive value of $\mu = \Delta t/(\Delta x)^2$ there exists a constant $C(\mu)$ such that, for all positive values of k and Δx,

$$\left|1 - 4\mu \sin^2 \tfrac{1}{2}k\Delta x - \mathrm{e}^{-k^2\Delta t}\right| \leq C(\mu)k^4(\Delta t)^2.$$

Verify that when $\mu = \frac{1}{4}$ this inequality is satisfied by $C = \frac{1}{2}$.
(ii) The explicit central difference method is used to construct a solution of the equation $u_t = u_{xx}$ on the region $0 \leq x \leq 1, \ t \geq 0$. The boundary conditions specify that $u(0, t) = u(1, t) = 0$, and $u(x, 0) = u^0(x)$, the same function as given in Exercise 1. Take $\epsilon = 0.01$, and show that in the sine series given there

$$\sum_{m=2p_0+1}^{\infty} |a_m| \leq \frac{\epsilon}{4} \text{ if } p_0 = 82,$$

and that then

$$\sum_{m=1}^{2p_0-1} |a_m| m^4 \leq 8p_0(2p_0+1)(2p_0-1)/3\pi^2.$$

Deduce that over the range $0 \leq t \leq 1$ the numerical solution will have an error less than 0.01 when $\mu = \frac{1}{4}$ provided that $\Delta t \leq 1.7 \times 10^{-10}$.
(iii) Verify by calculation that the numerical solution has error less than 0.01 over this range when $\mu = \frac{1}{4}$ and $\Delta t = 0.0025$.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

[Observe that for this model problem the largest error is always in the first time step.]

2.3 Suppose that the mesh points x_j are chosen to satisfy

$$0 = x_0 < x_1 < x_2 < \cdots < x_{J-1} < x_J = 1$$

but are otherwise arbitrary. The equation $u_t = u_{xx}$ is approximated over the interval $0 \le t \le t_F$ by

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

where $\Delta x_j = x_{j+1} - x_j$. Show that the leading terms of the truncation error of this approximation are

$$T_j^n = \tfrac{1}{2}\Delta t\, u_{tt} - \tfrac{1}{3}(\Delta x_j - \Delta x_{j-1})u_{xxx} - \tfrac{1}{12}[(\Delta x_j)^2 + (\Delta x_{j-1})^2 - \Delta x_j \Delta x_{j-1}]u_{xxxx}.$$

Now suppose that the boundary conditions prescribe the values of $u(0,t)$, $u(1,t)$ and $u(x,0)$. Write $\Delta x = \max \Delta x_j$, and suppose that the mesh is sufficiently smooth so that $|\Delta x_j - \Delta x_{j-1}| \le \alpha(\Delta x)^2$, for $j = 1, 2, \ldots, J-1$, where α is a constant. Show that

$$|U_j^n - u(x_j, t_n)| \le \left(\tfrac{1}{2}\Delta t\, M_{tt} + (\Delta x)^2 \left\{ \tfrac{1}{3}\alpha M_{xxx} + \tfrac{1}{12}[1 + \alpha \Delta x\,] M_{xxxx} \right\} \right) t_F$$

in the usual notation, provided that the stability condition

$$\Delta t \le \tfrac{1}{2}\Delta x_{j-1}\Delta x_j, \quad j = 1, 2, \ldots, J-1,$$

is satisfied.

2.4 The numbers a_j, b_j, c_j satisfy

$$a_j > 0,\ c_j > 0,\ b_j > a_j + c_j, \quad j = 1, 2, \ldots, J-1,$$

and

$$e_j = \frac{c_j}{b_j - a_j e_{j-1}}, \quad j = 1, 2, \ldots, J-1,$$

with $e_0 = 0$. Show by induction that $0 < e_j < 1$ for $j = 1, 2, \ldots, J-1$.

Show, further, that the conditions

$$b_j > 0, \quad b_j \ge |a_j| + |c_j|, \quad j = 1, 2, \ldots, J-1,$$

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

are sufficient for $|e_0| \leq 1$ to imply that $|e_j| \leq 1$ for $j = 1, 2, \ldots, J-1$.

2.5 Consider the equation $u_t = u_{xx}$, with the boundary condition $u(1,t) = 0$ for all $t \geq 0$, and

$$\frac{\partial u}{\partial x} = \alpha(t)u + g(t) \quad \text{at } x = 0, \text{ for all } t \geq 0,$$

with $\alpha(t) \geq 0$. Show in detail how the Thomas algorithm is used when solving the equation by the θ-method. In particular, derive the starting conditions which replace equation (2.71).

2.6 (i) By considering separately the cases of real roots and complex roots, or otherwise, show that both the roots of the quadratic equation $z^2 + bz + c = 0$ with real coefficients lie in or on the unit circle if and only if $|c| \leq 1$ and $|b| \leq 1 + c$.
(ii) Show that the scheme

$$U_j^{n+1} - U_j^{n-1} = \tfrac{1}{3}\mu\left\{\delta_x^2 U_j^{n+1} + \delta_x^2 U_j^n + \delta_x^2 U_j^{n-1}\right\}$$

is stable for all values of μ.
(iii) Show that the scheme

$$U_j^{n+1} - U_j^{n-1} = \tfrac{1}{6}\mu\left\{\delta_x^2 U_j^{n+1} + 4\delta_x^2 U_j^n + \delta_x^2 U_j^{n-1}\right\}$$

is unstable for all values of μ.

2.7 Find the leading terms in the truncation error of the explicit scheme

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

for the differential equation

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

on the region $0 < x < 1, \ t > 0$, with boundary conditions specifying the values of u at $x = 0$ and $x = 1$. Deduce a bound on the global error of the result in terms of bounds on the derivatives of u and p, under the condition $0 < p(x)\Delta t \leq \frac{1}{2}(\Delta x)^2$.

2.8 Apply the θ-method to the problem of the previous exercise, showing that the conditions required for the stable use of the Thomas algorithm will hold if $p(x) > 0$. Show also that a maximum principle will apply provided that $2\Delta t(1-\theta)p(x) \leq (\Delta x)^2$ for all x.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

2.9 Consider application of the θ-method to approximate the equation $u_t = u_{xx}$ with the choice

$$\theta = \tfrac{1}{2} + \frac{(\Delta x)^2}{12\Delta t}.$$

Show that the resulting scheme is unconditionally stable, has a truncation error which is $O((\Delta t)^2 + (\Delta x)^2)$ and provides rather more damping for all Fourier modes that oscillate from time step to time step than does the Crank–Nicolson scheme. However, show that the mesh ratio $\Delta t/(\Delta x)^2$ must lie in the interval $[\frac{1}{6}, \frac{7}{6}]$ for the maximum principle to apply.

2.10 To solve the equation $u_t = u_{xx}$ suppose that we use a non-uniform mesh in the x-direction, the mesh points being given by

$$x_j = \frac{j^2}{J^2}, \qquad j = 0, 1, 2, \ldots, J.$$

By the change of variable $x = s^2$, write the equation

$$u_t = \frac{1}{2s}\frac{\partial}{\partial s}\left(\frac{1}{2s}\frac{\partial u}{\partial s}\right);$$

use a uniform mesh with $\Delta s = 1/J$, and apply the difference scheme of Exercise 7, with the additional factor $1/2s_j$ on the right-hand side. Show that the leading terms of the truncation error are

$$T_j^n = \tfrac{1}{2}\Delta t\, u_{tt} - \tfrac{1}{24}(\Delta s)^2 \frac{1}{2s}\left[\left(\frac{1}{2s}u_{sss}\right)_s + \left(\frac{1}{2s}u_s\right)_{sss}\right]$$

and that this may be transformed into

$$T_j^n = \tfrac{1}{2}\Delta t\, u_{tt} - (\Delta s)^2(\tfrac{2}{3}u_{xxx} + \tfrac{1}{3}x u_{xxxx}).$$

Compare this with the leading terms of the truncation error obtained in Exercise 3.

2.11 Suppose that the Crank–Nicolson scheme is used for the solution of the equation $u_t = u_{xx}$, with boundary conditions $U_0^n = U_J^n = 0$ for all $n \geq 0$, and the initial condition $U_k^0 = 1$ for a fixed k with $0 < k < J$, $U_j^0 = 0, j \neq k$. Write $w_j = U_j^1$, and verify that w_j satisfies the recurrence relation

$$-\tfrac{1}{2}\mu w_{j-1} + (1+\mu)w_j - \tfrac{1}{2}\mu w_{j+1} = q_j, \quad w_0 = w_J = 0,$$

where $q_k = 1 - \mu$, $q_{k+1} = q_{k-1} = \frac{1}{2}\mu$, and $q_j = 0$ otherwise.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press

Suppose that the mesh is sufficiently fine, and that the point x_k is sufficiently far from the boundary that both k and $J-k$ are large. Explain why a good approximation to w_j may be written

$$w_j = Ap^{|j-k|}, \quad j \neq k,$$
$$w_k = A + B,$$

where $p = (1+\mu-\sqrt{(1+2\mu)})/\mu$. Write down and solve two equations for the constants A and B, and show that $w_k = 2/\sqrt{(1+2\mu)} - 1$. Deduce that (i) $w_k < 1$ for all $\mu > 0$; (ii) $w_k > 0$ if and only if $\mu < \frac{3}{2}$; and (iii) $w_k \geq w_{k+1}$ if and only if $\mu \leq (7-\sqrt{17})/4$.

2.12 Show that the (Hermitian) difference scheme

$$(1+\tfrac{1}{12}\delta_x^2)(U^{n+1}-U^n) = \tfrac{1}{2}\mu\delta_x^2(U^{n+1}+U^n) + \tfrac{1}{2}\Delta t[f^{n+1}+(1+\tfrac{1}{6}\delta_x^2)f^n]$$

for approximating $u_t = u_{xx} + f$, for a given function f and with fixed $\mu = \Delta t/(\Delta x)^2$, has a truncation error which is $O((\Delta t)^2)$.

https://doi.org/10.1017/CBO9780511812248.004 Published online by Cambridge University Press