# MATH2103: Lecture Note on Numerical Solution of Partial Differential Equations

Shixiao W. Jiang

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

jiangshx@shanghaitech.edu.cn

2025 年 11 月 18 日

# Contents

# Chapter 1

# Linear second order elliptic equations in two dimensions

## 1.1 A model problem

As in previous chapters we shall begin with the simplest model problem, to solve

$$u_{xx} + u_{yy} + f(x, y) = 0, \quad (x, y) \in \Omega, \tag{6.1a}$$

$$u = 0, \quad (x, y) \in \partial\Omega, \tag{6.1b}$$

where $\Omega$ is the unit square

$$\Omega := (0, 1) \times (0, 1) \tag{6.2}$$

and $\partial\Omega$ is the boundary of the square. If we compare with the parabolic equation, of the type discussed in Chapter 3,

$$\frac{\partial u}{\partial t} = u_{xx} + u_{yy} + f(x, y) \tag{6.3}$$

we see that, if the solution converges to a limit as $t \to \infty$, this limit will be the solution of (6.1a). This connection between the elliptic problem and the time-dependent solution of the parabolic problem has often been exploited in the solution of elliptic problems; in Chapter 7 we shall discuss the relation between iterative methods for the solution of elliptic problems and time stepping finite difference methods for the solution of the corresponding parabolic problems.

We cover the unit square with a uniform square grid with $J$ intervals in each direction, so that

$$\Delta x = \Delta y = 1/J, \tag{6.4}$$

and we approximate (6.1) by the central difference scheme

$$\frac{U_{r+1,s} + U_{r-1,s} + U_{r,s+1} + U_{r,s-1} - 4U_{r,s}}{(\Delta x)^2} + f_{r,s} = 0. \tag{6.5}$$

Writing equation (6.5) with $r = 1, 2, \ldots, J-1$ and $s = 1, 2, \ldots, J-1$ we obtain a system of $(J-1)^2$ equations which have exactly the same structure as the systems which arose in the solution of parabolic equations by implicit methods in Chapter 3. We shall assume for the moment that the equations have been solved in some way, and go on to investigate the accuracy of the result.

## 1.2 Error analysis of the model problem

As usual we begin with the truncation error. Substituting into equation (6.5) the exact solution $u(x_r, y_s)$ of the differential equation (6.1), and expanding in Taylor series, the truncation error is easily found to be

$$T_{r,s} = \frac{1}{12}(\Delta x)^2 \left(u_{xxxx} + u_{yyyy}\right)_{r,s} + o\left((\Delta x)^2\right). \tag{6.6}$$

Indeed this is bounded by $T$, where

$$|T_{r,s}| \leq T := \frac{1}{12}(\Delta x)^2 \left(M_{xxxx} + M_{yyyy}\right) \tag{6.7}$$

in the usual notation for bounds on the partial derivatives of $u(x, y)$.

We now define an operator $\mathbf{L_h}$ on the set of all arrays $U$ of values $U_{r,s}$,

$$(\mathbf{L_h}U)_{r,s} \equiv L_h U_{r,s} := \frac{1}{(\Delta x)^2}\left(U_{r+1,s} + U_{r-1,s} + U_{r,s+1} + U_{r,s-1} - 4U_{r,s}\right), \tag{6.8}$$

at all the interior points $J_\Omega \equiv \{(x_r, y_s); r = 1, 2, \ldots, J-1 \text{ and } s = 1, 2, \ldots, J-1\}$. Then the numerical approximation satisfies the equation

$$\mathbf{L_h}U_{r,s} + f_{r,s} = 0, \tag{6.9}$$

and the exact solution satisfies

$$\mathbf{L_h}u_{r,s} + f_{r,s} = T_{r,s}. \tag{6.10}$$

We define the error in the usual way as

$$e_{r,s} := U_{r,s} - u_{r,s} \tag{6.11}$$

and see that

$$\mathbf{L_h}e_{r,s} = -T_{r,s}. \tag{6.12}$$

Since the values of $u(x, y)$ were given at all points on the boundary, it follows that the boundary values of $e_{r,s}$ are zero.

To obtain a bound on $e_{r,s}$ we first define a comparison function

$$\Phi_{r,s} := \left(x_r - \frac{1}{2}\right)^2 + \left(y_s - \frac{1}{2}\right)^2. \tag{6.13}$$

Then

$$\mathbf{L_h}\Phi_{r,s} = 4; \tag{6.14}$$

this can be shown either by direct calculation, or more easily by noting that $\Phi$ is a quadratic function of $x$ and $y$, and therefore $\mathbf{L_h}\Phi$ must give the exact value of $\Phi_{xx} + \Phi_{yy}$, which is evidently 4. If we now write

$$\psi_{r,s} := e_{r,s} + \frac{1}{4}T\Phi_{r,s} \tag{6.15}$$

we obtain

$$\begin{aligned} \mathbf{L_h}\psi_{r,s} &= \mathbf{L_h}e_{r,s} + \frac{1}{4}T\mathbf{L_h}\Phi_{r,s} \\ &= -T_{r,s} + T \geq 0 \quad \forall (x_r, y_s) \in J_\Omega. \end{aligned} \tag{6.16}$$

We now appeal to **a maximum principle** (极值原理), similar to that used in Theorem 2.2 in Chapter 2, and which we shall prove in a more general context in Section 6.5. Briefly, the operator $\mathbf{L_h}$ is such that, if

at a point $(x_r, y_s)$ we have $\mathbf{L_h}\psi_{r,s} \geq 0$, then $\psi_{r,s}$ cannot be greater than all the neighbouring values. Hence it follows from (6.16) that a positive maximum value of $\psi$ must be attained at a point on the boundary of our square region. But $e_{r,s}$ vanishes on the boundary, and the maximum of $\Phi$ is $\frac{1}{2}$, being attained at the corners. Hence the maximum of $\psi$ on the boundary is $\frac{1}{8}T$, and

$$\psi_{r,s} \leq \frac{1}{8}T \quad \forall (x_r, y_s) \in J_\Omega. \tag{6.17}$$

But since $\Phi$ is nonnegative, from the definition of $\psi$ we obtain

$$\begin{aligned} U_{r,s} - u(x_r, y_s) = e_{r,s} &\leq \psi_{r,s} \leq \frac{1}{8}T \\ &= \frac{1}{96}(\Delta x)^2 (M_{xxxx} + M_{yyyy}). \end{aligned} \tag{6.18}$$

Notice that this is a one-sided bound; it is a simple matter to repeat the analysis, but defining $\psi_{r,s} = \frac{1}{4}T\Phi_{r,s} - e_{r,s}$. The result will be to show that $-e_{r,s} \leq \frac{1}{8}T$, from which we finally obtain the required bound

$$|U_{r,s} - u(x_r, y_s)| \leq \frac{1}{96}(\Delta x)^2 (M_{xxxx} + M_{yyyy}). \tag{6.19}$$

## 1.3 The general diffusion equation

We shall now extend this approach to a more general elliptic problem, the diffusion equation

$$\nabla \cdot (a \nabla u) + f = 0 \quad \text{in } \Omega, \tag{6.20}$$

where

$$a(x, y) \geq a_0 > 0. \tag{6.21}$$

We shall suppose that $\Omega$ is a bounded open region with boundary $\partial\Omega$. The boundary conditions may have the general form

$$\alpha_0 u + \alpha_1 \frac{\partial u}{\partial n} = g \quad \text{on } \partial\Omega, \tag{6.22}$$

where $\partial/\partial n$ represents the derivative in the direction of the outward normal and

$$\alpha_0 \geq 0, \quad \alpha_1 \geq 0, \quad \alpha_0 + \alpha_1 > 0. \tag{6.23}$$

As in the previous chapter we cover the region $\Omega$ with a regular mesh, with size $\Delta x$ in the $x$-direction and $\Delta y$ in the $y$-direction.

Suppose $a$ is smoothly varying and write $\mathbf{b} = \partial a/\partial x$, $\mathbf{c} = \partial a/\partial y$. Then we could expand (6.20) as

$$a\nabla^2 u + b u_x + c u_y + f = 0. \tag{6.24}$$

At points away from the boundary we can approximate this equation by using central differences, giving an approximation $\mathbf{U} := \{U_{r,s}, (r, s) \in J_\Omega\}$ satisfying

$$a_{r,s}\left[\frac{\delta_x^2 U_{r,s}}{(\Delta x)^2} + \frac{\delta_y^2 U_{r,s}}{(\Delta y)^2}\right] + b_{r,s}\left[\frac{\Delta_{0x} U_{r,s}}{\Delta x}\right] + c_{r,s}\left[\frac{\Delta_{0y} U_{r,s}}{\Delta y}\right] + f_{r,s} = 0. \tag{6.25}$$

The truncation error of this five-point scheme is defined in the usual way and is easily found to be second order in $\Delta x, \Delta y$. The terms in this equation which involve $U_{r+1,s}$ and $U_{r-1,s}$ are

$$\left[\frac{a_{r,s}}{(\Delta x)^2} - \frac{b_{r,s}}{2\Delta x}\right] U_{r-1,s} + \left[\frac{a_{r,s}}{(\Delta x)^2} + \frac{b_{r,s}}{2\Delta x}\right] U_{r+1,s}. \tag{6.26}$$

In order to use a maximum principle to analyse the error of this scheme it is necessary to ensure that all the coefficients at the points which are neighbours to $(r, s)$ have the same sign. This would evidently require that

$$|b_{r,s}|\Delta x \leq 2a_{r,s} \quad \forall r, s \tag{6.27}$$

with a similar restriction for $c_{r,s}$. This would imply the use of a fine mesh where the diffusion coefficient $a(x, y)$ is small but changing rapidly.

**Remark 1.3.1** *This scheme works for time independent problem but might not work for time dependent problem. I do not suggest it for parabolic problem. The constraint in (6.27) is not necessary.*

A more natural scheme however is based more directly on an integral form of (6.20), as we saw when considering polar co-ordinates in Section 2.8. Consider the control volume $V$ around a mesh point, which was introduced in the last chapter and is indicated by dotted lines in Fig. 6.1. Integrating (6.20) over this volume and using Gauss' theorem we obtain

$$\int_{\partial V} a(\partial u/\partial n)dl + \int_V f dxdy = 0. \tag{6.28}$$

We can now construct a difference scheme by approximating the terms in (6.28). The boundary $\partial V$ of $V$ has normals in the co-ordinate directions and the normal derivatives can be approximated by divided differences using the same five points as in (6.25). We approximate each of the line integrals along the four sides of the rectangle by the product of the length of the side and the value of the normal derivative at the mid-point of the side. In the same way, we approximate the integral of $f(x, y)$ over the element by the product of the area of the element and the value of $f(x, y)$ at the centre. As a result, we obtain the scheme

$$\begin{aligned}&\frac{\Delta y}{\Delta x}\left[a_{r+1/2,s}\left(U_{r+1,s} - U_{r,s}\right) - a_{r-1/2,s}\left(U_{r,s} - U_{r-1,s}\right)\right] \\ &+ \frac{\Delta x}{\Delta y}\left[a_{r,s+1/2}\left(U_{r,s+1} - U_{r,s}\right) - a_{r,s-1/2}\left(U_{r,s} - U_{r,s-1}\right)\right] + \Delta x \Delta y f_{r,s} = 0\end{aligned} \tag{6.29a}$$

or

$$\left[\frac{\delta_x(a\delta_x U)}{(\Delta x)^2} + \frac{\delta_y(a\delta_y U)}{(\Delta y)^2}\right]_{r,s} + f_{r,s} = 0. \tag{6.29b}$$

It is often convenient to use the 'compass point' notation indicated in Fig. 6.1 and to write (6.29a), (6.29b) as

$$a_e(U_E - U_P) - a_w(U_P - U_W) + a_n(U_N - U_P) - a_s(U_P - U_S) + f_P = 0. \tag{6.30}$$

Since we have assumed that the function $a(x, y)$ is positive it is now easy to see that the coefficients in the scheme (6.30) always have the correct sign, without any restriction on the size of the mesh.

Another advantage of the form (6.29), comes in problems where there is an interface at which material properties (represented by $a$) change abruptly, yet the normal flux $a\partial u/\partial n$ is continuous. Suppose we arrange the grid so that the material interface passes vertically through the point $e$, with the constant value $a_E$ holding on the right and the value $a_P$ on the left. Then in terms of an intermediate value $U_e$ on the interface, the common value of the flux can be approximated by

$$\begin{aligned}\frac{a_E(U_E - U_e)}{\frac{1}{2}\Delta x} &= \frac{a_P(U_e - U_P)}{\frac{1}{2}\Delta x} \\ &= \frac{a_e(U_E - U_P)}{\Delta x}, \quad \text{say,}\end{aligned} \tag{6.31}$$

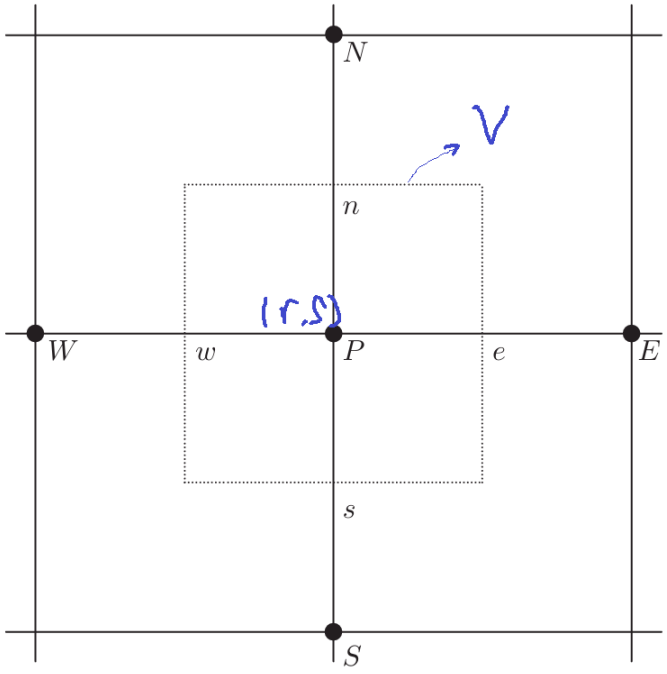


Fig. 6.1. The control volume about the point $P$.

Figure 1.1: The control volume about the point $P$.

where the elimination of $U_e$ to give the last expression, ready for substitution into (6.30), is accomplished by defining $a_e$ through

$$\frac{2}{a_e} = \frac{1}{a_E} + \frac{1}{a_P}. \tag{6.32}$$

**Remark 1.3.2** *This finite volume scheme has advantage for interface with discontinuous heat conductivity problem.* 界面问题有有优势。

## 1.4 Boundary conditions on a curved boundary

Either form (6.25) or (6.30) has to be modified in the neighbourhood of a boundary which does not lie along one of the mesh lines. We saw in Section 3.4 how the second derivatives can be approximated at such points, and now have to extend the method to the more general diffusion equation.

Let us first consider a situation such as that in Fig. 6.2 where Dirichlet data are given on the curved boundary. If we use Taylor series expansions to obtain an approximation to (6.24) at the point $P$ in the form (6.25), we need to use the given values at $A$ and $B$ instead of the values at $E$ and $N$. It is sufficient to consider $A$ only, and we write $PA = \theta\Delta x$. Then Taylor expansions for $u_A$ and $u_W$ give

$$\begin{aligned} u_A &= \left[u + \theta\Delta x\, u_x + \frac{1}{2}(\theta\Delta x)^2 u_{xx} + \cdots\right]_P, \\ u_W &= \left[u - \Delta x\, u_x + \frac{1}{2}(\Delta x)^2 u_{xx} - \cdots\right]_P. \end{aligned}$$

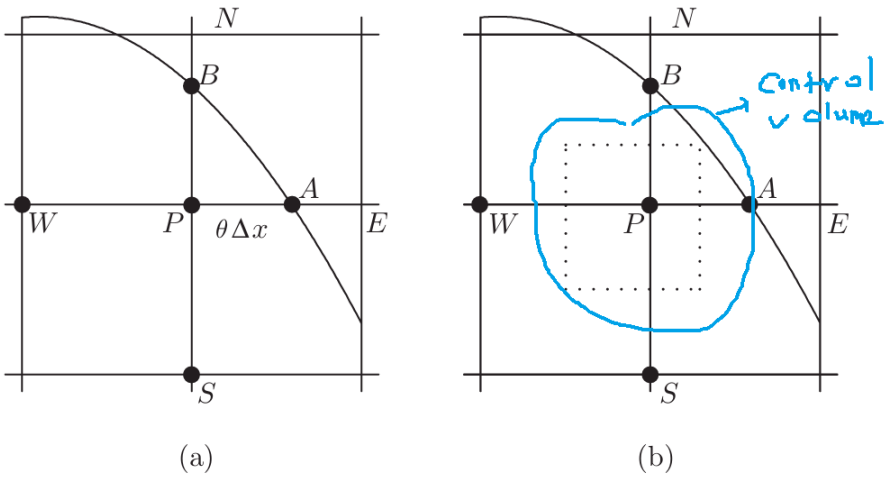


Fig. 6.2. Dirichlet condition on a curved boundary; (a) shows the points used in the modified difference scheme and (b) shows the modified control volume.

Figure 1.2: Dirichlet condition on a curved boundary; (a) shows the points used in the modified difference scheme and (b) shows the modified control volume.

From these, by eliminating first $u_{xx}$ and then $u_x$, we obtain approximations

$$[u_x]_P \approx \frac{u_A - \theta^2 u_W - (1-\theta^2)u_P}{\theta(1+\theta)\Delta x}, \tag{6.33a}$$

$$[u_{xx}]_P \approx \frac{u_A + \theta u_W - (1+\theta)u_P}{\frac{1}{2}\theta(1+\theta)(\Delta x)^2}. \tag{6.33b}$$

Carrying out the same construction for $U_P$, $U_S$ and $U_B$, we readily obtain an appropriately modified form of (6.25). For a maximum principle to hold we shall again require a restriction on the size of the mesh, just as at an ordinary interior point.

The alternative integral form of the difference scheme (6.30) is modified near a Dirichlet boundary in a similar way. Thus for each mesh point we draw a rectangular control volume about it as in Fig. 6.1 but, if the mesh line joining $P$ to one of its neighbours crosses the boundary, the corresponding side of the rectangle is drawn perpendicular to the mesh line, crossing it half-way to the boundary, as in Fig. 6.2(b). For example, the distance $PA$ in this figure is a fraction $\theta$ of the mesh size $PE = \Delta x$, so the width of the volume $V$ is $\frac{1}{2}(1+\theta)\Delta x$.

Hence the line integral along the bottom side of the element is approximated by

$$\frac{1}{2}(1+\theta)\Delta x \left[\frac{-a_s(U_P - U_S)}{\Delta y}\right]. \tag{6.34}$$

We must also make an adjustment to the approximation to the normal derivative in the line integral up the right-hand side of the element; this derivative is approximated by

$$\frac{a_a(U_A - U_P)}{\theta \Delta x}, \tag{6.35}$$

where $a_a$ is the value of $a(x,y)$ at the point midway between $A$ and $P$ and $a_b$ will have a similar meaning. Noting that in Fig. 6.2(b) the boundary cuts the mesh lines at two points, $A$ and $B$ with $PA = \theta\Delta x$ and

$PB = \phi\Delta y$, we obtain the difference approximation at the point $P$ as

$$\frac{1}{2}(1+\phi)\Delta y\left[\frac{a_a(U_A-U_P)}{\theta\Delta x}-\frac{a_w(U_P-U_W)}{\Delta x}\right] + \frac{1}{2}(1+\theta)\Delta x\left[\frac{a_b(U_B-U_P)}{\phi\Delta y}-\frac{a_s(U_P-U_S)}{\Delta y}\right] + \frac{1}{4}(1+\theta)(1+\phi)\Delta x\Delta y f_P = 0, \quad (6.36)$$

**where the value of** $a_a = a((A+P)/2)$. It is clear that in the case where $a(x,y)$ is constant this scheme will be identical to that given by (6.33a, b). More generally, **it has the advantage that the coefficients still satisfy the conditions required for a maximum principle.**

**Derivative boundary conditions are more difficult to deal with.** As we saw in Chapter 3, it is difficult to construct accurate difference approximations to the normal derivative, and **it is necessary to take account of a number of different possible geometrical configurations.** Moreover, we shall see in Section 6.7 that derivative boundary conditions are dealt with much more straightforwardly in a finite element approximation. However we will show how in a simple case the integral form of the equation can be adapted.

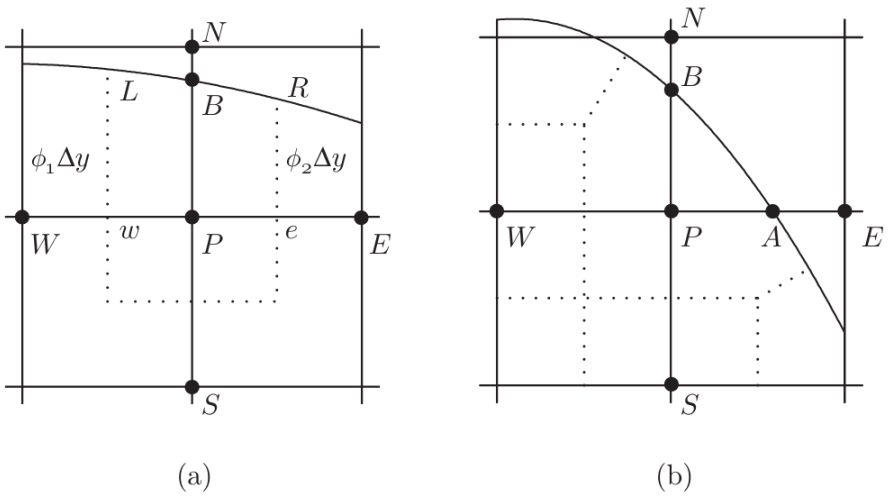


Fig. 6.3. Integral form of Neumann condition; the boundary intersects one mesh line in (a) and two in (b).

Figure 1.3: Integral form of Neumann condition; the boundary intersects one mesh line in (a) and two in(b).

Consider the situation shown in Fig. 6.3(a) where the boundary cuts the mesh line $PN$ at the point $B$, but does not cut the other lines joining $P$ to its nearest neighbours. We also assume that a Neumann condition,

$$\frac{\partial u}{\partial n} = g, \quad (6.37)$$

is given on this part of the boundary; **the general condition (6.22) is more difficult to deal with.** We construct an element $V$ round $P$ by drawing three sides of the rectangle as we have done before, but extending the two vertical sides to meet the boundary at the points $L$ and $R$ and using the part of the

boundary between them. Writing

$$wL = \phi_1 \Delta y, \quad eR = \phi_2 \Delta y, \tag{6.38}$$

we can approximate the line integrals along the horizontal and vertical sides of the element just as we did before, but we notice that the results will be less accurate, since we are not evaluating the normal derivatives at the mid-points of the vertical sides. In the same way the double integral is approximated by the product of the area of the element and the value of $f$ at $P$. Again there is a loss of accuracy, since $P$ is not at the centre of the element.

Finally we must approximate the line integral along the boundary $RBL$; here we can write

$$\int_{RBL} a \frac{\partial u}{\partial n} dl = \int_{RBL} agdl \tag{6.39}$$

which is approximated by

$$a_B g_B \psi \Delta x \tag{6.40}$$

where $\psi \Delta x$ is the length of the line $LR$. This leads finally to the difference equation

$$\begin{aligned}(\phi_2 + \frac{1}{2})\Delta y \left[a_e \frac{U_E - U_P}{\Delta x}\right] + (\phi_1 + \frac{1}{2})\Delta y \left[a_w \frac{U_W - U_P}{\Delta x}\right] \\ + \Delta x \left[a_s \frac{U_S - U_P}{\Delta y}\right] + \psi \Delta x a_B g_B + \frac{1}{2}(\phi_1 + \phi_2 + 1)\Delta x \Delta y f_P = 0.\end{aligned} \tag{6.41}$$

In the fairly common situation where the boundary cuts two of the mesh lines from $P$ to its neighbours, the procedure is much the same, but it will be necessary to use some diagonal lines in the construction of the element, as **in Fig. 6.3(b). The geometrical details are tedious, and we shall not discuss them further.**

## 1.5 Error analysis using a maximum principle

We shall assume that in approximating (6.20), (6.22), or any linear elliptic equation, we have been successful in constructing, at each interior point $P \in J_\Omega$, an approximation of the form

$$L_h U_P + f_P + g_P = 0, \tag{6.42}$$

where $g_P$ represents any boundary data that are picked up from other than Dirichlet boundary conditions. We assume that the following conditions hold:-

(i) For each $P \in J_\Omega$, $L_h$ has the form

$$L_h U_P = \sum_k c_k U_k - c_P U_P, \tag{6.43}$$

where the coefficients are **all** positive $(c_k > 0, c_p > 0)$ and the sum over $k$ is taken over mesh points which are neighbours of $P$. In the difference scheme above these have only included the four nearest neighbours, but the analysis can be applied to more general schemes which involve neighbours such as $U_{NE}$ in a diagonal direction. Also, when $P$ is near the boundary, some of the neighbours may be in a set $J_{\partial\Omega}$ of points on the boundary such as $A$ and $B$ in Fig. 6.2, and equation (6.36), and then the corresponding values $U_k$ are given by Dirichlet boundary conditions. **The key requirement is that all the coefficients which occur in (6.43) must be positive.**

(ii) For each $P \in J_\Omega$

$$c_P \geq \sum_k c_k. \tag{6.44}$$

(iii) The set $J_\Omega$ is connected. We say that a point $P$ is connected to each of its neighbours that occurs in (6.43) with a non-zero coefficient. The set is then connected if, given any two points $P$ and $Q$ in $\Omega_J$, there is a sequence of points $P = P_0, P_1, P_2, \ldots, P_m = Q$ such that each point $P_r$ is connected to $P_{r-1}$ and $P_{r+1}$, for $r = 1, 2, \ldots, m-1$.

(iv) At least one of the equations (6.43) must involve a boundary value $U_A$ which is given by a Dirichlet boundary condition. In other words, Dirichlet boundary conditions must be given on at least part of the boundary. (Such a condition is needed, or at least (6.22) with $\alpha_0 \neq 0$, to ensure **uniqueness** of the solution of (6.1).)

消去边界$U$

Note that in these conditions we refer to **interior points** $J_\Omega$ and **boundary points** $J_{\partial\Omega}$ in a specific sense. An interior point is one at which an equation of the form (6.42) is satisfied, with the coefficients satisfying the appropriate conditions. A boundary point is a point which appears in at least one equation but is one at which no equation is given and the value of $U$ is prescribed, by a Dirichlet condition. **At a point on the boundary of the region at which a Neumann or mixed condition is given we will normally eliminate the value of $U$, as in the situation illustrated in Fig. 6.3 where the unknown value $U_B$ does not appear in the equation (6.41) for $U_P$,** or in any of the equations; **or we treat it as an interior point with its own equation, as in the treatment of the symmetry boundary condition in the example of Section 6.9 below.** The **boundary** points are therefore those at which values are prescribed by Dirichlet conditions. Note that these definitions and conventions, which are convenient for our present purposes, are slightly different from those used in Chapter 5 (where difference operators as in (5.8) involved only points in $J_\Omega$) and from those to be used in Chapter 7.

**Lemma 1.5.1 (Maximum Principle)** *Suppose that $L_h, J_\Omega$ and $J_{\partial\Omega}$ satisfy all the above assumptions and that a mesh function $U$ satisfies*

$$L_h U_P \geq 0 \quad \forall P \in J_\Omega. \tag{6.45}$$

*Then $U$ cannot attain a nonnegative maximum at an interior point, i.e.*

$$\max_{P \in J_\Omega} U_P \leq \max\{\max_{A \in J_{\partial\Omega}} U_A, 0\}. \tag{6.46}$$

**Proof.** We prove this by contradiction. Suppose the interior maximum $M_\Omega \geq 0$ occurs at a point P and that $M_\Omega > M_{\partial\Omega}$, the maximum over the boundary points. Then from (6.45), (6.43) and (6.44)

$$\begin{aligned} M_\Omega = U_P &\leq \frac{1}{c_P} \sum_k c_k U_k \\ &\leq \frac{1}{c_P} \sum_k c_k M_\Omega \leq M_\Omega. \end{aligned} \tag{6.47}$$

This means that equality holds throughout (6.47), which implies that all the values $U_k$ involved are also equal to $M_\Omega$. Hence the maximum value is attained also at all the points which are connected neighbours of P. The same argument can then be used at each of these points, and so on. Since we have assumed that $J_\Omega$ is connected this shows that $U$ takes the same value $M_\Omega$ at all the interior points, and at least one of these

points must have a connected neighbour which is on the boundary. This contradicts the assumption that $M_\Omega > M_{\partial\Omega}$. ■

**Corollary 1.5.2** *The inequality*

$$L_h U_P \le 0 \quad \forall P \in J_\Omega$$

*implies*

$$\min_{P\in J_\Omega} U_P \ge \min\{\min_{A\in J_{\partial\Omega}} U_A, 0\}. \tag{6.48}$$

**Proof.** The proof follows in the same manner as above, or alternatively we may simply apply the lemma to the mesh function $-U$. ■

We next define the truncation error at each interior point in the usual way, noting that (6.42) has been properly scaled,

$$T_P := L_h u_P + f_P + g_P. \tag{6.49}$$

Hence for the error $e_P = U_P - u_P$ at each interior point we have

$$L_h e_P = -T_P, \quad P \in J_\Omega. \tag{6.50}$$

We shall assume as usual that $e_A = 0$ at all boundary points at which a Dirichlet condition is prescribed, i.e. Dirichlet conditions are imposed exactly. This will enable us to bound $e_P$ by means of the maximum principle in Lemma 6.1.

**Theorem 1.5.3** *Suppose a* ***nonnegative mesh function*** $\Phi$ *is defined on* $J_\Omega \cup J_{\partial\Omega}$ *such that*

$$L_h \Phi_P \ge 1 \quad \forall P \in J_\Omega, \tag{6.51}$$

*and that all the above four conditions are satisfied. Then the error in the approximation (6.42) is bounded by*

$$|e_P| \le \left[\max_{A\in J_{\partial\Omega}} \Phi_A\right] \left[\max_{P\in J_\Omega} |T_P|\right]. \tag{6.52}$$

**Proof.** Let us denote by $T$ the maximum of the absolute value of the truncation error, $|T_P|$. Then we can apply Lemma 6.1 to the function $T\Phi_P + e_P$ because

$$L_h(T\Phi_P + e_P) \ge T - T_P \ge 0. \tag{6.53}$$

We therefore deduce, using the fact that $\Phi$ is nonnegative, that

$$\begin{aligned} \max_{P\in J_\Omega}(e_P) &\le \max_{P\in J_\Omega}(T\Phi_P + e_P) \\ &\le \max_{A\in J_{\partial\Omega}}(T\Phi_A + e_A), \end{aligned} \tag{1.1}$$

i.e.,

$$\max_{P\in J_\Omega} e_P \le \left[\max_{A\in J_{\partial\Omega}} \Phi_A\right] T, \tag{6.54}$$

where we have used the fact that $e_A = 0$ at all boundary points, because of the assumption that we use the exact values of the given Dirichlet boundary conditions.

This gives a one-sided bound. If we apply the lemma to the function $T\Phi_P - e_P$ we obtain a similar bound for $-e_P$, thus giving the required result. ■

The presentation and analysis here have been given in rather general terms. The analysis of the model problem in Section 6.2 followed the same method, the only difference being that we used a comparison function $\Phi$ for which $L_h\Phi_P = 4$; a more general theorem that includes both cases will be given below. When applying the technique to a particular problem we must find a bound on the truncation error, giving $T$, and then **construct a function** $\Phi$. Evidently the determination of $T$ is quite straightforward, requiring simply a Taylor series expansion, whereas it may be more difficult to construct a suitable $\Phi$. This function is not, of course, unique. In the model problem we could, for example, have defined

$$\Phi_{r,s} = \frac{1}{4}\left[(x_r - p)^2 + (y_s - q)^2\right]; \tag{6.55}$$

the required conditions would be satisfied for any values of the constants $p$ and $q$. The particular values that were chosen, $p = q = \frac{1}{2}$, give the smallest value for $\max(\Phi_A)$.

We shall now apply this general analysis to some more complicated cases. Consider first the solution of Poisson's equation (6.1) in a region with a curved boundary, as discussed in Section 6.4. The finite difference approximation will satisfy a maximum principle, and for all points at which all the four neighbours are also interior mesh points, the truncation error has the form (6.6) and satisfies (6.7). However, at mesh points next to the boundary, where one or more of the neighbours lie on the boundary, we must use a more general difference approximation, such as (6.33b). A Taylor series expansion easily shows that

$$\frac{u_A + \theta u_W - (1+\theta)u_P}{\frac{1}{2}\theta(1+\theta)(\Delta x)^2} = (u_{xx})_P - \frac{1}{3}(1-\theta)\Delta x(u_{xxx})_P + O((\Delta x)^2). \tag{6.56}$$

It is possible for an interior mesh point to have more than one neighbour on the boundary, but since $0 < \theta < 1$ in (6.56) it is easy to see that in all cases we can choose positive constants $K_1$ and $K_2$ such that

$$|T_{r,s}| \le K_1(\Delta x)^2 \quad \textbf{at ordinary points}, \tag{6.57a}$$

$$|T_{r,s}| \le K_2\Delta x \quad \textbf{next to the boundary}, \tag{6.57b}$$

provided that $\Delta x$ is sufficiently small. Hence

$$|T_{r,s}| \le K_1(\Delta x)^2 + K_2\Delta x \tag{6.58}$$

at all interior mesh points.

Now suppose that the region is contained in a circle with centre $(p, q)$ and radius $R$, and define the comparison function $\Phi_{r,s}$ by (6.55). Then $L_h\Phi_P = 1$ at all ordinary interior mesh points, as before; this result also holds at points next to the boundary, since the truncation error in (6.56) involves the third and higher derivatives, and vanishes for a quadratic polynomial. We can therefore apply Theorem 6.1 and deduce that

$$|U_{r,s} - u(x_r, y_s)| \le \frac{1}{4}R^2[K_1(\Delta x)^2 + K_2\Delta x], \tag{1.2}$$

since $0 \le \Phi \le \frac{1}{4}R^2$ throughout the region, and on the boundary. This shows that the error is $O(\Delta x)$ as the mesh size tends to zero, rather than the $O((\Delta x)^2)$ obtained for a simple region.

### 1.5.1 First Example for a sharper bound

**A slightly modified comparison function can however be used to produce a sharper error bound.** In analysis of problems of this kind it is quite common for the truncation error to have one form at

the ordinary points of the mesh, but to be different near the boundary. It is therefore convenient to have a generalised form of Theorem 6.1.

**Theorem 1.5.4 (6.2)** *Suppose that, in the notation of Theorem 6.1, the set $J_\Omega$ is partitioned into two disjoint sets*

$$J_\Omega = J_1 \cup J_2, \quad J_1 \cap J_2 = \emptyset; \tag{6.59}$$

*the nonnegative mesh function $\Phi$ is defined on $J_\Omega \cup J_{\partial\Omega}$ and satisfies*

$$L_h\Phi_P \geq C_1 > 0 \quad \forall\, P \in J_1, \tag{6.60a}$$

$$L_h\Phi_P \geq C_2 > 0 \quad \forall\, P \in J_2, \tag{6.60b}$$

*and the truncation error of the approximation (6.42) satisfies*

$$|T_P| \leq T_1 \quad \forall\, P \in J_1, \tag{6.61a}$$

$$|T_P| \leq T_2 \quad \forall\, P \in J_2, \tag{6.61b}$$

*Then the error in the approximation is bounded by*

$$|e_P| \leq \left[\max_{A \in J_{\partial\Omega}} \Phi_A\right] \max\left\{\frac{T_1}{C_1}, \frac{T_2}{C_2}\right\}. \tag{6.62}$$

**Proof.** The proof is an easy extension of that of Theorem 6.1; it is only necessary to apply Lemma 6.1 to the function $K\Phi + e$, where the constant $K$ is chosen to ensure that the maximum principle applies. The details of the proof are left as an exercise. ■

We now apply this theorem to the problem with the curved boundary, taking the set $J_1$ to contain all the ordinary internal mesh points, and $J_2$ to contain all those mesh points which have one or more neighbours on the boundary. We then define the mesh function $\Phi$ by

$$\begin{aligned} \Phi_P &= E_1\left\{(x_r - p)^2 + (y_s - q)^2\right\} \quad \forall\, P \in J_\Omega, \\ \Phi_P &= E_1\left\{(x_r - p)^2 + (y_s - q)^2\right\} + E_2 \quad \forall\, P \in J_{\partial\Omega}, \end{aligned}$$

where $E_1$ and $E_2$ are positive constants to be chosen later. Then

$$L_h\Phi_P = 4E_1 \quad \forall\, P \in J_1, \tag{6.63a}$$

(1.3)

but for points in $J_2$ there is an additional term, or terms, arising from the boundary points. In the approximation (6.33b) the coefficients of $u_A$ is

$$\frac{2}{\theta(1+\theta)(\Delta x)^2} \tag{1.4}$$

and since $0 < \theta < 1$ this coefficient is bounded away from zero and cannot be less than $1/(\Delta x)^2$. Hence

$$\begin{aligned} L_h\Phi_P &\geq 4E_1 + E_2/(\Delta x)^2 \\ &\geq E_2/(\Delta x)^2 \quad \forall\, P \in J_2. \end{aligned} \tag{6.63b}$$

Applying Theorem 6.2, with the truncation error bounds of (6.61) given by (6.57) as $T_1 = K_1(\Delta x)^2$ and $T_2 = K_2\Delta x$, we obtain

$$|e_P| \le (E_1R^2 + E_2)\max\left\{\frac{K_1(\Delta x)^2}{4E_1}, \frac{K_2(\Delta x)^3}{E_2}\right\}. \tag{6.64}$$

This bound depends only on the ratio $E_2/E_1$, and is optimised when the two quantities in $\max\{\cdot,\cdot\}$ are equal; so we get the result

$$|e_P| \le \frac{1}{4}K_1R^2(\Delta x)^2 + K_2(\Delta x)^3, \tag{6.65}$$

showing that the error is in fact second order in the mesh size. Notice that the leading term in this error bound is unaffected by the lower order terms in the truncation error near the boundary.

### 1.5.2 Second Example

As a second example, consider the solution of Poisson's equation in the unit square, with **the Neumann boundary condition** $u_x(1, y) = g(y)$ on the right-hand boundary $x = 1$, and Dirichlet conditions on the other three sides. As in a similar problem in Chapter 2 we introduce an extra line of points outside the boundary $x = 1$, with $r = J + 1$. The boundary condition is approximated by

$$\frac{U_{J+1,s} - U_{J-1,s}}{2\Delta x} = g_s. \tag{6.66}$$

We then eliminate the extra unknown $U_{J+1,s}$ from the standard difference equation at $r = J$, giving

$$\frac{U_{J,s+1} + U_{J,s-1} + 2U_{J-1,s} - 4U_{J,s} + 2g_s\Delta x}{(\Delta x)^2} + f_{J,s} = 0. \tag{6.67}$$

This is now an equation of the general form (6.42), satisfying the required conditions for the maximum principle; so in the application of the maximum principle these points with $r = J$ are to be regarded as **internal points**.

**only ghost points are boundary and all other original points are internal points**

The truncation error at the ordinary points is as before, and an expansion in Taylor series gives the truncation error of (6.67). The result is

$$T_{r,s} = \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy}) + O((\Delta x)^4), \quad r < J, \tag{6.68a}$$

$$T_{J,s} = \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy}) - \frac{1}{3}\Delta x u_{xxx} + O((\Delta x)^3). \tag{6.68b}$$

The same argument as before, using Theorem 6.1 and the comparison function $\Phi$ given in (6.13), shows that the error is bounded by

$$|e_{r,s}| \le \frac{1}{8}\left\{\frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy}) + \frac{1}{3}\Delta x M_{xxx}\right\}. \tag{6.69}$$

This error bound is of order $O(\Delta x)$, but as in the previous example a sharper bound can be obtained by choice of a different comparison function and application of Theorem 6.2. We define

$$\Phi = (x - p)^2 + (y - q)^2 \tag{6.70}$$

where $p$ and $q$ are constants to be determined. We partition the internal points of the region into $J_1$ and $J_2$, where $J_2$ consists of those points with $r = J$. Then in region $J_1$ the standard difference equation is used, and

$$L_h\Phi_P = 4, \quad P \in J_1. \tag{6.71}$$

At the points where $r = J$, a different operator is used, as in (6.67), and we have

$$L_h\Phi_P = 4 - 4(1-p)/\Delta x, \quad P \in J_2. \tag{6.72}$$

In the notation of Theorem 6.2 we then find that

$$\frac{T_1}{C_1} = \frac{\frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy})}{4}, \tag{6.73a}$$

$$\frac{T_2}{C_2} = \frac{\frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy}) + \frac{1}{3}\Delta x M_{xxx}}{4 - (1-p)/\Delta x}. \tag{6.73b}$$

If we now choose, for example, $p = 2, q = \frac{1}{2}$, we obtain

$$\frac{T_2}{C_2} \leq \frac{1}{3}(\Delta x)^2 M_{xxx} + \frac{1}{12}(\Delta x)^3(M_{xxxx} + M_{yyyy}), \tag{6.74}$$

and at all points of the square we see that $(x-2)^2 + (y-\frac{1}{2})^2 \leq \frac{17}{4}$; so by adding $T_1/C_1$ and $T_2/C_2$ we obtain the error bound

$$|e_{r,s}| \leq \frac{17}{4}(\Delta x)^2 \left[\frac{1}{3}M_{xxx} + \frac{1}{12}\left(\frac{1}{4} + \Delta x\right)(M_{xxxx} + M_{yyyy})\right]. \tag{6.75}$$

This shows that, in this example also, the error is second order in the mesh size.

### 1.5.3 Third Example

**What about Robin?**

The same technique can be used to show that the error in our approximation for the solution of Poisson's equation in a fairly general region, given either **Dirichlet or Neumann conditions** on a curved boundary, is still second order; the technique is complicated only by the need to take account of a number of different geometrical possibilities. Indeed, the same ideas can be applied quite generally to both elliptic and parabolic problems where maximum principles hold. Thus we end this section by sharpening some of the results that we gave in Chapter 2 for one-dimensional heat flow.

In Section 2.11 we gave a proof of convergence of the $\theta$-method for the heat equation. We now define the operator $L_h$ by

$$L_h\psi = \frac{[\theta\delta_x^2\psi_j^{n+1} + (1-\theta)\delta_x^2\psi_j^n]}{(\Delta x)^2} - \frac{\psi_j^{n+1} - \psi_j^n}{\Delta t}, \tag{6.76}$$

and take $P$ as the point $(x_j, t_{n+1})$. It is easy to see that, provided $0 \leq \theta \leq 1$ and $\mu(1-\theta) \leq \frac{1}{2}$, the conditions of Theorem 6.1 are satisfied for an appropriately defined set of points $J_\Omega$. We shall suppose that Dirichlet conditions are given on $x = 0$ and $x = 1$; then the interior points $J_\Omega$ are those for which $1 \leq j \leq J-1$ and $1 \leq n \leq N$, and the boundary points $J_{\partial\Omega}$ have $j = 0$ or $j = J$ or $n = 0$. In the notation of Section 2.10, the exact solution $u$ and the numerical solution $U$ satisfy $L_h u_j^n = -T_j^{n+1/2}$ and $L_h U_j^n = 0$, so that

$$L_h e_j^n \quad = \quad T_j^{n+1/2}$$

and we have

$$T_j^{n+1/2} = O(\Delta t) + O((\Delta x)^2).$$

Now define the comparison function

$$\Phi_j^n = At_n + Bx_j(1-x_j) \tag{6.77}$$

where $A$ and $B$ are nonnegative constants; it is easy to show that $L_h\Phi_j^n = -(A+2B)$. Hence

$$L_h(e_j^n - \Phi_j^n) = A + 2B + T_j^{n+1/2} \geq 0 \tag{6.78}$$

if the constants are so chosen that $A + 2B \geq T := \max |T_j^{n+1/2}|$.

At points on the boundary $J_{\partial\Omega}$ we have $e_j^n = 0$, $\Phi_0^n = \Phi_j^n = At_n$, and $\Phi_j^0 = Bx_j(1-x_j)$, so that $e_j^n - \Phi_j^n \leq 0$ on the boundary. Hence by Lemma 6.1, $e_j^n - \Phi_j^n \leq 0$ in $J_\Omega$. We now consider two choices of the constants $A$ and $B$.

(i) Take $A = T$, $B = 0$; we have therefore shown that

$$e_j^n \leq \Phi_j^n = t_n T,$$

and the same argument applied to the mesh function $(-e_j^n - \Phi_j^n)$ shows that

$$|e_j^n| \leq t_n T$$

which agrees with (2.96).

(ii) Take $A = 0$, $B = \frac{1}{2}T$ and we obtain in the same way

$$|e_j^n| \leq \frac{1}{2}x_j(1-x_j)T. \tag{6.81}$$

From combining these results it is in fact clear that

$$|e_j^n| \leq \max_{m<n, 0<i<J} \left\{ |T_i^{m+1/2}| \right\} \min\{t_n, \frac{1}{2}x_j(1-x_j)\} \tag{6.82}$$

which properly **reflects the fact that $e = 0$ round the boundary of the domain, and grows away from it.**

We saw that for the model problem in Section 2.6 **the error in the solution tends to zero as $t$ increases, while the bound in (6.82) does not.** If we have some additional information about the solution, and we can show that $|T_j^{n+1/2}| \leq \tau_n$, where $\tau_n$ is a known decreasing function of $n$, **we may be able to construct a comparison function which leads to an error bound which decreases with $n$. An example is given in Exercise 7.**

**How to do to let the error decrease wrt time? very important**

### 1.5.4 Fourth Example

In a similar way we can obtain an error bound in the case of a Neumann boundary condition, the problem considered in Section 2.13. With a homogeneous Neumann condition at $x = 0$, i.e. at $x_0$, the operator at $j = 1$ is replaced by

$$(L_h\psi)_1^n = \frac{\theta(\psi_2^{n+1} - \psi_1^{n+1}) + (1-\theta)(\psi_2^n - \psi_1^n)}{\Delta x^2} - \frac{\psi_1^{n+1} - \psi_1^n}{\Delta t} \tag{6.83}$$

obtained from (2.103) with $\alpha = 0$ and $g = 0$. The truncation error, from (2.109), and with $\theta = 0$, at this point is

$$T_1^{n+1/2} = \frac{1}{2}\Delta t u_{tt} - \frac{1}{12}(\Delta x)^2 u_{xxxx} - \frac{1}{2}u_{xx}. \tag{6.84}$$

We now apply the argument of Theorem 6.2, with $J_1$ including all the interior mesh points with $j > 1$, and $J_2$ all the interior mesh points with $j = 1$. In (6.61) we can take $T_1 = T = \frac{1}{2}\Delta t M_{tt} + \frac{1}{12}(\Delta x)^2 M_{xxxx}$ as before, but in $J_2$ we must use $T_2 = T + \frac{1}{2}M_{xx}$. We then construct the comparison function such that $\Phi_0^n = \Phi_1^n$, namely

$$\Phi_j^n = \begin{cases} At_n + B(1-x_j)(1-\Delta x + x_j) & \text{in } J_1, \\ At_n + B(1-x_j)(1-\Delta x + x_j) + K & \text{in } J_2, \end{cases} \tag{6.85}$$

so that we obtain

$$L_h\Phi_j^n = \begin{cases} -(A+2B) & \text{in } J_1, \\ -(A+2B) - K/(\Delta x)^2 & \text{in } J_2. \end{cases} \tag{6.86}$$

This shows that $L_h(e_j^n - \Phi_j^n) \geq 0$ in $J_1$ and in $J_2$ if we choose the constants so that

$$A + 2B \geq T, \tag{6.87a}$$

$$A + 2B + K/(\Delta x)^2 \geq T + \frac{1}{2}M_{xx}. \tag{6.87b}$$

This is clearly satisfied if we take the same $A$ and $B$ as before and

$$K = \frac{1}{2}(\Delta x)^2 M_{xx}.$$

The boundary of the region only involves the points where $j = J$ or $n = 0$, and at these points it is easy to see as before that $e_j^n - \Phi_j^n \leq 0$. Hence this holds at the interior points also, so that

$$e_j^n \leq At_n + B(1-x_j)(1-\Delta x + x_j) + \frac{1}{2}(\Delta x)^2 M_{xx}. \tag{6.88}$$

With the same choices of $A$ and $B$ as before we get the final result

$$\begin{aligned} |e_j^n| \leq & \max_{\substack{m<n \\ 0<i<J}} \{|T_i^{m+1/2}|\} \times \\ & \min\left\{ t_n, (1-x_j)(1-\Delta x + x_j) + \frac{1}{2}(\Delta x)^2 M_{xx} \right\}, \end{aligned} \tag{6.89}$$

showing that the error is $O(\Delta t) + O((\Delta x)^2)$, just as in the case of Dirichlet boundary conditions.

## 1.6 Asymptotic error estimates and deferred correction

We had examples in the previous section where a straightforward error analysis gave a bound of first order in the mesh size, while a more sophisticated analysis produced a bound of second order. **This must raise the question whether a still more careful analysis might show that the error is in fact of third order.** In those examples it is fairly clear that no improvement in order is possible, but in more complicated problems it may not be at all easy to see what the actual order of error really is. For example, while in those examples the difficulties stemmed from lower order truncation errors at exceptional points near the boundary, the consideration of more general elliptic operators in Section 2.15, Section 3.5 and the next section can lead to lower order truncation errors at all points. It is therefore useful to have estimates which show more precisely how the error behaves in the limit as the mesh size tends to zero.

Such estimates often exploit more fully the maximum principle of Lemma 6.1, and the corresponding result which holds for an elliptic operator $L$. As a preliminary, suppose we denote by $\Phi_{\partial\Omega}$ the maximum of

$\Phi_A$ over all the boundary nodes, as it appears in the bound of (6.52) with $\Phi$ satisfying (6.51). Now let us apply Lemma 6.1 to $\Psi := e - T(\Phi_{\partial\Omega} - \Phi)$: it is clear that

$$L_h \Psi_P = -T_P + T L_h \Phi_P \geq 0,$$

and also that the maximum of $\Psi$ on the boundary is zero; so we can conclude that $\Psi_P \leq 0, \forall P \in J$. We can also repeat this argument with $-e$, and hence deduce that

$$|e_P| \leq T(\Phi_{\partial\Omega} - \Phi_P), \tag{6.90}$$

which can be a much stronger result than that of (6.52) in Theorem 6.1; in particular, it gives an error bound that decreases to zero at some point of the boundary.

To illustrate how to estimate the asymptotic behaviour of the error, we consider first the solution of Poisson's equation in the unit square, with Dirichlet conditions on the boundary. Using the standard five-point difference scheme we can easily write down an expression for the truncation error, and take more terms in the expansion because the underlying solution will be smooth for reasonable data:

$$T_{r,s} = \frac{1}{12}(\Delta x)^2 (u_{xxxx} + u_{yyyy})_{r,s} + \frac{1}{360}(\Delta x)^4 (u_{xxxxxx} + u_{yyyyyy})_{r,s} + \cdots . \tag{6.91}$$

Then the error $e_{r,s}$ satisfies the equation

$$L_h e_{r,s} = -T_{r,s} = -\frac{1}{12}(\Delta x)^2 (u_{xxxx} + u_{yyyy})_{r,s} + O\left((\Delta x)^4\right). \tag{6.92}$$

Suppose now that we write $\psi(x, y)$ for the solution of the equation

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}) \tag{6.93}$$

which vanishes on the boundary of the unit square; and let $\Psi$ be the result if we were to approximate this problem by our numerical scheme. Then an application of the error bound in Theorem 6.1 shows that $\Psi - \psi = O\left((\Delta x)^2\right)$. Moreover, we can combine this result with (6.93) so that another application of the theorem shows that

$$\frac{e_{r,s}}{(\Delta x)^2} = \psi(x_r, y_s) + O\left((\Delta x)^2\right). \tag{6.94}$$

That is, the numerical solution to the original Poisson's equation has an error which is given by

$$U_{r,s} = u(x_r, y_s) + (\Delta x)^2 \psi(x_r, y_s) + O\left((\Delta x)^4\right). \tag{6.95}$$

This shows that the error is exactly second order, **and not of any higher order, except in a very special case where the function $\psi$ is identically zero;** but of course the expression is only valid in the limit as the mesh size goes to zero.

**Important!(deferred correction)**

We can estimate the size of this error from the difference between two numerical solutions using different mesh sizes, comparing the results at the mesh points they have in common. Alternatively, we could use divided differences of the approximate solution $U$ to estimate the right-hand side of (6.93) and then actually solve the discrete equations again to obtain an approximation to what we called $\psi$ above; then substitution into (6.95) leads to a fourth order approximation. This procedure is called **deferred correction**. Again we should emphasise that the extra orders of accuracy will only be obtained if the mesh size is sufficiently small for the asymptotic expansion (6.95) to be valid.

An asymptotic estimate can be obtained in a similar way for the last example in the previous section, with a Neumann boundary condition on one side of the square, where the error bound was obtained by a rather arbitrary choice of the comparison function $\Phi$; it is clear that the error bound (6.75) is **most unlikely to be the best possible**. In this case the error satisfies $L_h(e_{rs}) = -T_{rs}$ where $T_{rs}$ is given by (6.68). If we now define $\psi$ to be the solution of the problem

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}) \tag{6.97}$$

with

$$\psi(0,y) = \psi(x,0) = \psi(x,1) = 0, \quad \psi_x(1,y) = -\frac{1}{6}u_{xxx}(1,y),$$

we find that the same numerical method applied to this problem will lead to the equations satisfied by $e_{r,s}$, with additional truncation terms of higher order; the details are left as an exercise – see Exercise 6. We thus see, as in the previous example, that

$$U_{r,s} = u(x_r, y_s) + (\Delta x)^2\psi(x_r, y_s) + o((\Delta x)^2), \tag{6.98}$$

and $\psi(.,.)$ **could be estimated by the deferred correction approach.**

The extension to problems involving a curved boundary is straightforward. As in the application immediately following Theorem 6.2 we divide the set of mesh points into $J_1$ and $J_2$, where $J_2$ comprises those mesh points that have one or more neighbours on the boundary. At points in $J_1$ we have, as in (6.92),

$$\left| L_h(e_{rs}) + \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy}) \right| \le K_3(\Delta x)^4 \quad \text{in } J_1, \tag{6.99}$$

where $K_3 = \frac{1}{360}(M_{xxxxxx} + M_{yyyyyy})$. We now suppose that Dirichlet conditions are given on the curved boundary. Then at points in $J_2$ we can use (6.56) and (6.57b) to give

$$|L_h(e_{rs})| \le K_2\Delta x \quad \text{in } J_2. \tag{6.100}$$

We define $\psi(x,y)$ as in (6.93) to be the solution of

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}) \tag{6.101}$$

which vanishes on the boundary. Then, provided the functions $u$ and $\psi$ have sufficient bounded derivatives, there exist constants $K_4$ and $K_5$ such that

$$|L_h(\psi_{rs}) + \frac{1}{12}(u_{xxxx} + u_{yyyy})| \le K_4(\Delta x)^2 \quad \text{in } J_1 \tag{6.102a}$$

$$|L_h(\psi_{rs})| \le K_5\Delta x \quad \text{in } J_2. \tag{6.102b}$$

By subtracting (6.102) from (6.99) and (6.100) we then have

$$\left| L_h\left(\frac{e_{rs}}{(\Delta x)^2} - \psi_{rs}\right) \right| \le (K_3 + K_4)(\Delta x)^2 \quad \text{in } J_1, \tag{6.103a}$$

$$\left| L_h\left(\frac{e_{rs}}{(\Delta x)^2} - \psi_{rs}\right) \right| \le \frac{K_2}{\Delta x} + K_5\Delta x \quad \text{in } J_2. \tag{6.103b}$$

We now apply Theorem 6.2; the function $\Phi_P$ and the values of $C_1$ and $C_2$ are the same as those used to obtain (6.64), i.e., $C_1 = 4E_1$, $C_2 = E_2/(\Delta x)^2$, but $T_1 = (K_3 + K_4)(\Delta x)^2$ and $T_2 = K_2/\Delta x + K_5\Delta x$. We thus obtain

$$\left| \frac{e_{rs}}{(\Delta x)^2} - \psi_{rs} \right| \leq (E_1 R^2 + E_2) \max \left\{ \frac{(K_3 + K_4)(\Delta x)^2}{4E_1}, \frac{K_2 + K_5(\Delta x)^2}{E_2/\Delta x} \right\}. \tag{6.104}$$

Choosing $E_1 = \frac{1}{4}(K_3 + K_4)(\Delta x)^2$, $E_2 = K_2\Delta x + K_5(\Delta x)^3$ we then obtain

$$\left| \frac{e_{rs}}{(\Delta x)^2} - \psi_{rs} \right| \leq C(\Delta x), \tag{6.105}$$

showing that

$$e_{rs} = (\Delta x)^2 \psi(x_r, y_s) + O((\Delta x)^3). \tag{6.106}$$

**Once again this shows how the lower order truncation error at points near the boundary does not affect the leading term in the asymptotic expansion of the error.**

A combination of the asymptotic error estimates obtained in this section and the rigorous error bounds of previous sections will give a useful description of the behaviour of the error in the linear problems with smooth solutions that we have been discussing. We should note, however, that for the more general equation involving a variable $a(x, y)$ **the construction of a function $\Phi$ with the required properties is likely to be more difficult; just how difficult, and how sharp will be the resulting bound, will depend on the form of** $a(x, y)$**,** a lower bound for which may well have to be used in the construction.

**However, an attempt to apply the analysis given above to a problem with Neumann conditions on part of a curved boundary is less successful.** If we use the approximation of (6.41) and expand in Taylor series we find the leading term of the truncation error is

$$\frac{p_1^2 - p_2^2}{1 + p_1 + p_2} u_{xy}. \tag{6.107}$$

We see at once that this is $O(1)$, and does not tend to zero as the mesh size goes to zero. It is now not difficult to use the maximum principle to obtain a bound on the error, and we find a bound which is $O(\Delta x)$. But our asymptotic expansion of the error is founded on the fact that the leading term in the truncation error is the product of a power of $\Delta x$ and a function of $x$ and $y$ only, independent of $\Delta x$. This is no longer true of the expression in (6.107), since $p_1$ and $p_2$ are functions of $x, y$ and $\Delta x$, and moreover are not smooth functions, but involve terms like the fractional part of $x/\Delta x$. This observation is illustrated in Fig. 6.4. This shows the results of a numerical solution of Poisson's equation on the annulus between circles of radii 1 and 0.3. In the first problem Dirichlet conditions are given on both boundaries, and these conditions and the function $f$ are so chosen that the solution is

$$u(x, y) = 1 + 3x^2 + 5y^2 + 7x^2y^2 + (x^2 + 2)^2. \tag{6.108}$$

The maximum error in this calculation is shown in the lower curve, which shows the error behaving like $(\Delta x)^2$. In the second problem the solution is the same, but a Neumann condition is given on the outer circular boundary. We notice that the error is now much larger, that there is a general trend of order $O(\Delta x)$, **but the detailed behaviour is very irregular.** To get a numerical solution of such a problem, with an error behaving smoothly like $O((\Delta x)^2)$, will require a more complicated approximation to the boundary condition.

**A major advantage of the finite element method, as we shall see in the next section, is that it deals with the application of Neumann boundary conditions in a simple and natural way, that leads to much better behaved errors.**

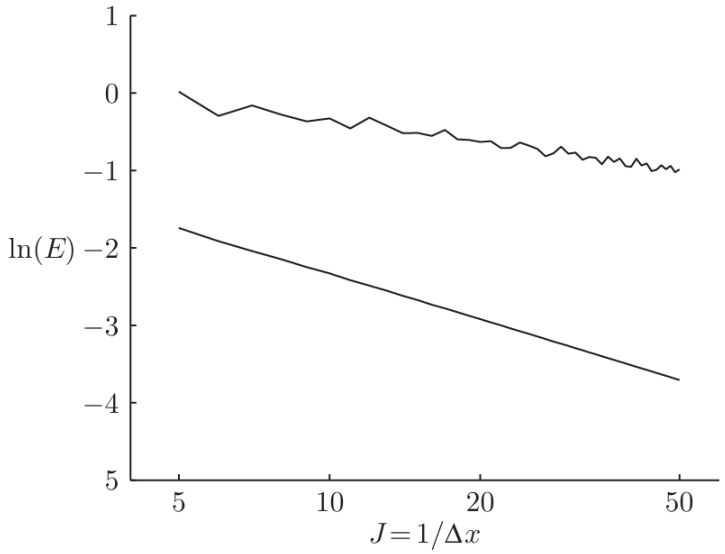


Fig. 6.4. Numerical solution of Poisson's equation in an annulus: lower curve, Dirichlet conditions; upper curve, Neumann condition on the outer boundary.

Figure 1.4: Numerical solution of Poisson' s equation in an annulus: lower curve, Dirichlet conditions; upper curve, Neumann condition on the outer boundary.

## 1.7 Variational formulation and the finite element method

Our general diffusion problem (6.20), (6.22) can be given a variational formulation. Consider first the case where Dirichlet boundary conditions are given at all points on the boundary. We define

$$I(v) := \int_\Omega \left[\frac{1}{2}a|\nabla v|^2 - fv\right] dx\, dy. \tag{6.108}$$

Then we assert that the solution of (6.20) satisfies a variational equation that we write as

$$\delta I(u) = 0. \tag{6.109}$$

This means that if $v = u + \delta u$ is any function for which $I(v)$ is defined, and which satisfies the boundary conditions, then

$$I(u + \delta u) - I(u) = o(\delta u). \tag{6.110}$$

We can show why this is so without attempting a completely rigorous proof. By expanding $I(u+\delta u)$ we find that

$$\begin{aligned} I(u + \delta u) - I(u) &= \int_\Omega \left[(a\nabla u)\cdot(\nabla \delta u) - f\delta u\right] dx\, dy \\ &\quad + \int_\Omega \frac{1}{2}a|\nabla \delta u|^2\, dx\, dy \\ &= \int_\Omega -[\nabla\cdot(a\nabla u) + f]\delta u\, dx\, dy + O\left((\delta u)^2\right), \end{aligned} \tag{6.111}$$

where we have used Gauss' theorem, and the fact that $\delta u = 0$ on the boundary of $\Omega$. The result follows.

In fact we have shown rather more, for since $a(x,y) > 0$ it follows from (6.111) that $I(u+\delta u) \geq I(u)$ for all functions $\delta u$ which vanish on the boundary. Hence the function $u$ gives the minimum of $I(v)$, taken over all functions $v$ which satisfy the boundary conditions.

Now suppose we take a finite expansion of the form

$$V(x,y) = \sum_{j=1}^{N} V_j \phi_j(x,y), \tag{6.112}$$

where the functions $\phi_j$ are given, and try to choose the coefficients $V_j$ so that this gives a good approximation to the solution $u$. We can expand $I(V)$ in the form

$$I(V) = \frac{1}{2} \sum_i \sum_j A_{ij} V_i V_j - \sum_i b_i V_i, \tag{6.113}$$

where

$$A_{ij} = \int_\Omega [a \nabla \phi_i \cdot \nabla \phi_j] dx\, dy \tag{6.114}$$

and

$$b_i = \int_\Omega f \phi_i dx\, dy. \tag{6.115}$$

Since the exact solution $u$ minimises $I(v)$, it is natural to define an approximation $U$ to the solution by choosing the coefficients $\{V_j\}$ to minimise $I(V)$. As (6.113) shows that $I(V)$ is a quadratic form in the coefficients, the determination of its minimum is a straightforward matter, applying the constraint that $V$ satisfies the boundary conditions.

**This approach was used by Rayleigh, Ritz and others in the nineteenth century with various choices of the functions $\phi_j(x,y)$. It is also the starting point for the finite element method which is now widely used to solve elliptic problems in preference to finite difference methods, especially by engineers.**

The particular feature of the finite element method lies in the choice of the $\phi_j(x,y)$, which are known as the **trial functions**, or **shape functions**. They are chosen so that each $\phi_j$ is non-zero only in a small part of the region $\Omega$. In the matrix $A$ most of the elements will then be zero, since $A_{ij}$ will only be non-zero if the shape functions $\phi_i$ and $\phi_j$ overlap. We will consider one simple case.

**Suppose the region $\Omega$ is a polygon and it is subdivided into triangular elements,** as in Fig. 6.5. The vertices $P$ of the triangles are known as **nodes**. We suppose that $V$ is piecewise linear, having a linear form in each triangle determined by its values at the vertices of the triangle. Then it is clearly continuous from one triangle to the next; and we suppose that the boundary conditions are similarly piecewise linear around the bounding polygon and $V$ takes on these values. The function $\phi_j(x,y)$ is defined to be the piecewise linear function which takes the value 1 at node $P_j$, and zero at all the other nodes. This defines the function uniquely, and it is clear that it is non-zero only in the triangles which have $P_j$ as a vertex. This function is for obvious reasons known as a **hat function**; drawn in three dimensions (Fig. 6.6) it has the form of a pyramid, with triangular faces.

This definition of the shape functions has the useful property that the coefficient $V_j$ gives the value of the function $V(x,y)$ at the node $P_j$, since all the other shape functions vanish at this node. Then to ensure that $V$ satisfies the boundary conditions we fix the coefficients $V_j$ corresponding to nodes on the boundary, and allow all the others to vary in the minimisation process.

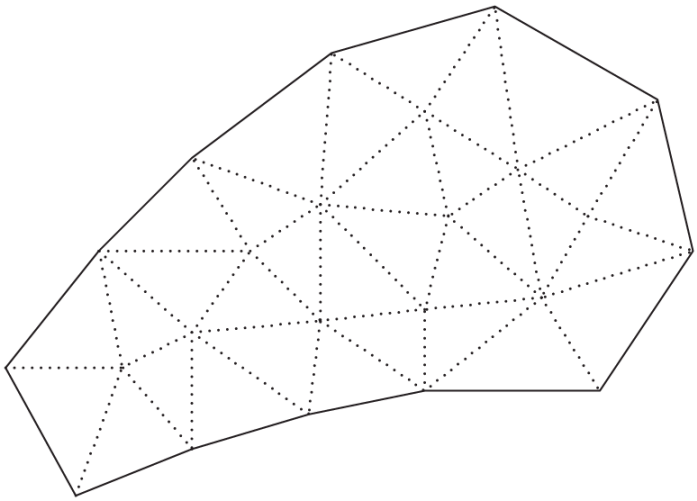

Fig. 6.5. Triangular elements in a polygon.

Figure 1.5: Triangular elements in a polygon.

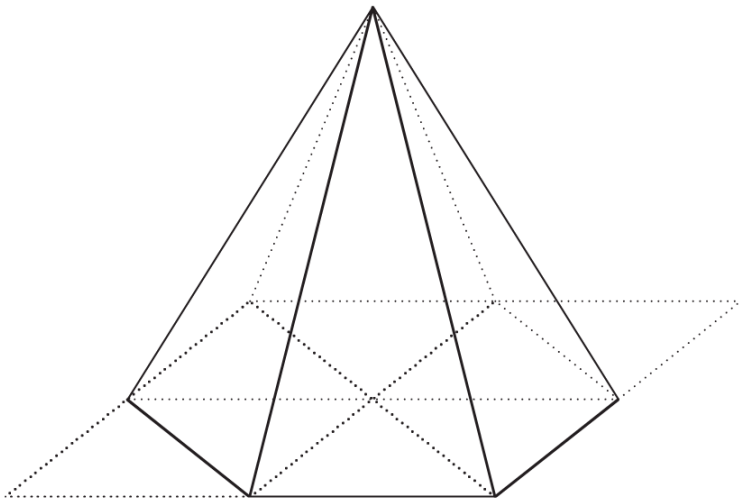

Fig. 6.6. A hat basis function.

Figure 1.6: A hat basis function.

One of the main advantages of the finite element method is that it adapts quite easily to difficult geometrical shapes, as a triangulation can easily follow a boundary of almost any shape. **Once a set of triangular elements has been constructed, the elements of the matrix $A$ can be evaluated in a routine way, taking no particular account of the shape of the triangles.**

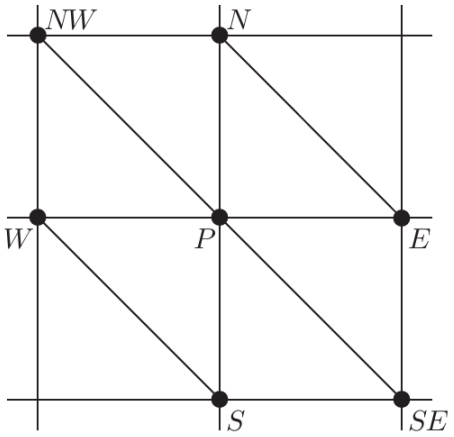


Fig. 6.7. Triangular elements on a square mesh.

Figure 1.7: Triangular elements on a square mesh.

To illustrate the procedure we shall consider the simplest model problem, the solution of Poisson's equation in a square with $u = 0$ on the boundary, as in (6.1). We cover the square with a uniform square mesh as before, and then divide each small square into two triangles by drawing a diagonal, as in Fig. 6.7. A general row of the matrix $A$ will then contain seven non-zero elements, as the hat function centred at the node $P$ will overlap with the hat functions centred at $P$ and its six neighbours, labelled $N, S, E, W$, $NW$ and $SE$ in the figure. Since each $\phi$ is a linear function of $x$ and $y$ in each triangle, its gradient $\nabla\phi$ is constant in each triangle. With $a(x, y) = 1$ in this example, the evaluation of

$$A_{ij} = \int_\Omega [\nabla\phi_i \cdot \nabla\phi_j] dx\, dy \tag{6.116}$$

is a simple matter. The partial derivatives of $\phi$ in each triangle each take one of the values 0, $1/\Delta x$ or $-1/\Delta x$, and we find that

$$A_{PP} = 4, \tag{6.117a}$$

$$A_{PN} = A_{PS} = A_{PE} = A_{PW} = -1 \tag{6.117b}$$

and

$$A_{P,NW} = A_{P,SE} = 0. \tag{6.117c}$$

We also require the value of $b_i$ from (6.115). If we approximate this by replacing the function $f(x, y)$ by a constant, the value of $f$ at the point $P$, we obtain

$$b_P = (\Delta x)^2 f_P. \tag{6.118}$$

However, we should point out here that in a finite element program the integrations of (6.114) and (6.115) will be carried out over each element before being assembled into the global equation of the form (6.119)

below: in particular, this will mean that (6.115) will commonly be approximated **by centroid quadrature** in each triangle, so that in (6.118) $f_P$ will be replaced by the mean of the six values at the centroids of the triangles sharing the vertex $P$. Now the minimum of the quadratic form (6.113) is given by the vector $U$, with values that satisfy the system of linear equations $AU = b$. The number of interior nodes corresponding to the $V_j$ which are allowed to vary, determines the dimension of $U$, $b$ and hence of $A$; there are no other contributions to the right-hand side of the equations because the boundary data are zero. A general equation of this system is

$$4U_P - U_N - U_S - U_E - U_W - (\Delta x)^2 f_P = 0, \tag{6.119}$$

which is **the same as the familiar finite difference scheme introduced in Section 6.1 – apart from the change of sign throughout.**

The error analysis of the finite element method has a different character from that of the finite difference schemes. We have seen that $u$ satisfies

$$\int_\Omega [a\nabla u \cdot \nabla w - fw]\, dx\, dy = 0 \tag{6.120}$$

for any function $w$ which vanishes on the boundary. Also the function $U$ would satisfy

$$\int_\Omega [a\nabla U \cdot \nabla W - fW]\, dx\, dy = 0 \tag{6.121}$$

if the integrals were carried out exactly (or $a$ and $f$ were constant), for any function $W$ which can be expressed as a finite sum of the form (6.112) and vanishes on the boundary. Now we can choose any $V$ of the form (6.112) which satisfies the boundary conditions and take both $w$ and $W$ to be $V - U$ because this difference vanishes on the boundary; then by subtraction we obtain

$$\int_\Omega [a\nabla(U - u) \cdot \nabla(V - U)] dx\, dy = 0. \tag{6.122}$$

Thus

$$\begin{aligned}\int_\Omega a|\nabla(V-u)|^2 dx\, dy &= \int_\Omega a|\nabla[(V-U)+(U-u)]|^2 dx\, dy \\ &= \int_\Omega a|\nabla(V-U)|^2 dx\, dy + \int_\Omega a|\nabla(U-u)|^2 dx\, dy,\end{aligned}$$

because the cross-product terms drop out by (6.122). This means that

$$\int_\Omega a|\nabla(U-u)|^2 dx\, dy \le \int_\Omega a|\nabla(V-u)|^2 dx\, dy \quad \forall V. \tag{6.123}$$

We can define a special norm, so-called **energy-norm**, for a function $w(x, y)$ which vanishes on the boundary of $\Omega$,

$$||w||_E^2 = \int_\Omega a|\nabla w|^2 dx\, dy. \tag{6.124}$$

We have thus shown that

$$||U - u||_E \le ||V - u||_E \quad \forall V. \tag{6.125}$$

This key result means that $U$ is the best possible approximation to $u$ of the form (6.112) in this norm. Its error can then be estimated very sharply by application of approximation theory. To do so here in any detail

would take us beyond the scope of this book, as would consideration of the effects of the quadrature needed for variable $a$ and $f$; we merely quote the main results as

$$||U - u||_E \leq C_1 h |u|_*, \tag{6.126a}$$

$$||U - u||_{L_2(\Omega)} \leq C_2 h^2 |u|_*, \tag{6.126b}$$

where $h$ is the maximum diameter of all the triangles in the triangulation,

$$|u|_* = ||u_{xx}|| + ||u_{xy}|| + ||u_{yy}||$$

and $||\cdot||$ denotes the $||\cdot||_{L_2(\Omega)}$ norm.

Finally, suppose that homogeneous Neumann boundary conditions are imposed on part of the boundary; these are dealt with as a natural part of the variational process and in a correspondingly automatic way by the finite element method. In the derivation of (6.111) a boundary integral

$$\int_{\partial\Omega} a \frac{\partial u}{\partial n} \delta u \, dl \tag{6.127}$$

is obtained from the application of Gauss' theorem. This is zero if at all boundary points either Dirichlet boundary conditions are imposed, giving $\delta u = 0$, or a homogeneous Neumann boundary condition is required, in which case $\partial u/\partial n = 0$ for the true solution. Thus when we minimise $I(v)$ given by (6.108), at the minimum not only is (6.20) satisfied at all interior points, but also $\partial u/\partial n = 0$ at all boundary points which are not constrained by a Dirichlet boundary condition. This is called a **natural** treatment of the Neumann boundary condition.

# Bibliography