Acta Numerica (2005), pp. 445–508
DOI: 10.1017/S0962492904000261

Steady-state convection-diffusion problems

Martin Stynes
Department of Mathematics,
National University of Ireland,
Cork, Ireland
E-mail: `m.stynes@ucc.ie`

In convection-diffusion problems, transport processes dominate while diffusion effects are confined to a relatively small part of the domain. This state of affairs means that one cannot rely on the formal ellipticity of the differential operator to ensure the convergence of standard numerical algorithms. Thus new ideas and approaches are required.

The survey begins by examining the asymptotic nature of solutions to stationary convection-diffusion problems. This provides a suitable framework for the understanding of these solutions and the difficulties that numerical techniques will face. Various numerical methods expressly designed for convection-diffusion problems are then presented and extensively discussed. These include finite difference and finite element methods and the use of special meshes.

CONTENTS

1. Introduction

1.1. What are convection-diffusion problems?

Our interest is in elliptic operators whose second-order derivatives are multiplied by some parameter ε that is allowed to be close to zero. These derivatives model diffusion while first-order derivatives (which are assumed to be present) are associated with convective or transport processes. In classical problems where ε is not close to zero, diffusion is the dominant mechanism in the model and the first-order convective derivatives play a relatively minor rôle in the analysis. On the other hand, when ε is near zero and the elliptic differential operator has convective terms, it is called a convection-diffusion operator. Such operators, while still satisfying the definition of ellipticity, live dangerously by flirting with the non-elliptic world. Their convective terms have a significant influence on the theoretical and numerical solution of the problem and cannot be summarily dismissed as 'lower-order terms'.

We shall see that the solutions of convection-diffusion problems have a convective nature on most of the domain of the problem, and the diffusive part of the differential operator is influential only in certain narrow subdomains. In these subdomains the gradient of the solution is large: its magnitude is proportional to some negative power of the parameter ε. We describe such behaviour by saying that the solution has a *layer*.

The fact that the elliptic nature of the differential operator is disguised on most of the domain means that numerical methods designed for elliptic problems will not work satisfactorily. In practice they usually exhibit a certain degree of instability. The challenge then is to modify these methods into a stable form without compromising their accuracy.

A second-order differential operator in n variables whose highest-order derivatives are

$$-\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2(\cdot)}{\partial x_i \partial x_j},$$

where the a_{ij} are constants, is said to be elliptic if

$$\sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \geq \sigma \sum_{i=1}^{n} \xi_i^2 \quad \text{for all } \xi_i \text{ and } \xi_j, \tag{1.1}$$

where $\sigma > 0$ is called the ellipticity constant. The differential operators in convection-diffusion problems stretch this definition as far as they dare: their ellipticity constant is close to zero.

It is often assumed (certainly in introductory textbooks in both theoretical differential equations and numerical analysis) that σ is not close to zero; for example the Laplacian has $\sigma = 1$. This assumption avoids many difficulties. Consider, say, the proof of convergence of a finite difference method for the problem $-\sigma u''(x) + u'(x) = f(x)$ on $(0,1)$ with $u(0) = u(1) = 0$: if you allow the positive constant σ to take a value near zero, does the argument still work? In fact, on a more fundamental level, what happens to the solution u of this boundary value problem when σ becomes small? Taking into account this alteration in the behaviour of u, how can we modify the numerical method so that it remains stable and accurate? It is questions such as these that will preoccupy us for the duration of this survey.

Our task now is to make concrete these suspicions and assertions. We shall begin in Section 2 by recalling some ideas about maximum principles and asymptotic expansions. In Section 3 we use these tools to begin an examination of the asymptotic nature of solutions to convection-diffusion problems. Furthermore, to carry out any numerical analysis we need *a priori* to have some bounds on the derivatives of the solutions of these problems; such estimates, and useful decompositions of the solutions, are also given in this section. Finite difference methods and the accuracy of their solutions are examined in Section 4. This leads naturally to the question of constructing suitable meshes for convection-diffusion problems, and Section 5 is devoted to an epitome of this class: Shishkin meshes. We present in this section a full analysis of a finite difference method on a Shishkin mesh.

The discussion up to this point has dealt only with ordinary differential equations, where the theory is fairly complete. Now we move into deeper waters: in Section 6 we discuss the nature of solutions to convection-diffusion problems posed in two-dimensional domains. *A priori* estimates for such problems are presented in Section 7, then some preliminary comments on numerical methods are given in Section 8. Finite difference methods for such problems are considered in Section 9, but our main emphasis is on Section 10 which is devoted to finite element methods.

This survey cannot, for reasons of length, give a complete account of the many numerical methods used to solve steady-state convection-diffusion problems. Roos, Stynes and Tobiska (1996) give a comprehensive discussion of numerical methods in this area and a new edition of this book is at present in preparation.

1.2. A little motivation and history

Perhaps the most common source of convection-diffusion problems is as linearizations of Navier–Stokes equations with large Reynolds number. Morton (1996) points out that this is by no means the only place where they arise: in his opening chapter he lists ten examples involving convection-diffusion equations that include the drift-diffusion equations of semiconductor device modelling and the Black–Scholes equation from financial modelling. He also observes that 'Accurate modelling of the interaction between convective and diffusive processes is the most ubiquitous and challenging task in the numerical approximation of partial differential equations.'

The numerical solution of convection-diffusion problems goes back to the 1950s (Allen and Southwell 1955), but only in the 1970s did it acquire a research momentum that has continued to this day. A potted history of the development of numerical methods for convection-diffusion problems is presented in Stynes (2003). The field is still very active and, as we shall see in our later sections, much remains to be done.

1.3. Notation

Throughout this article, ε is a small positive parameter and C will denote a generic constant that is independent of ε and of any mesh used – it can take different values in different places (even sometimes in the same calculation). A subscripted C (*e.g.*, C_1) is also a constant that is independent of ε and of any mesh used, but takes one fixed value.

2. Analytical tools

Consider the second-order differential operator L in n variables defined on some bounded domain (open connected set) D by

$$Lu(x) = -\sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u(x)}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial u(x)}{\partial x_i} + h(x)u(x),$$

where the a_{ij} are constants. We assume that L is elliptic in the sense of (1.1). Denote the closure of D by $\bar{D}$ and its boundary by ∂D, and let $C^k(S)$ denote the space of functions that are defined on a set S and k-times differentiable on S.

Lemma 2.1. (maximum principle) Let $u \in C^0(\bar{D}) \cap C^2(D)$ satisfy the differential inequality $Lu \geq 0$ on D. Suppose that the functions b_i and h are bounded on D, and $h \geq 0$ on D. Suppose also that $u \geq 0$ on ∂D. Then $u \geq 0$ on $\bar{D}$.

This familiar result is proved in Protter and Weinberger (1984). It is the key to analysing the behaviour of solutions to convection-diffusion problems and proving the convergence to these solutions of the outputs of various numerical methods.

A maximum principle can be used to bound a function in absolute value.

Corollary 2.2. (barrier function) Suppose that the functions b_i and h are bounded on D, and $h(x) \geq 0$ on D. Let $u, v \in C^0(\bar{D}) \cap C^2(D)$. Suppose that $|Lu(x)| \leq Lv(x)$ for all $x \in D$ and $|u(x)| \leq v(x)$ for all $x \in \partial D$. Then $|u(x)| \leq v(x)$ for all $x \in \bar{D}$.

Proof. One cannot immediately apply Lemma 2.1 to the functions $|u|$ and v because $|u|$ may not be differentiable. Instead apply this lemma to the functions $u - v$ and $u + v$ and deduce the desired result. □

A function such as v in Corollary 2.2 is called a *barrier function* for u. This corollary is often applied to a function u that is a solution of a boundary value problem – so $u|_{\partial D}$ and Lu are known, but $u|_D$ is unknown. We then try to choose a suitable function v that satisfies the hypotheses of the corollary in order to deduce some worthwhile information about the behaviour of u inside D.

Putting barrier functions aside for the moment, we turn our attention to a useful descriptive tool: asymptotic expansions.

Let $\varepsilon > 0$ be a small parameter. If $f = f(x, \varepsilon)$ and $g = g(x, \varepsilon)$ with x lying in some domain D, we write $f(x, \varepsilon) = \mathcal{O}(g(x, \varepsilon))$ as $\varepsilon \to 0$ if there exist a positive number A that is independent of ε and an $\varepsilon_0 > 0$ such that $|f(x, \varepsilon)| \le A|g(x, \varepsilon)|$ for $0 < \varepsilon \le \varepsilon_0$. If in addition A and ε_0 are independent of x, we say that $f(x, \varepsilon) = \mathcal{O}(g(x, \varepsilon))$ as $\varepsilon \to 0$ uniformly for $x \in D$.

This notation is useful for comparing functions of similar size. For functions of greatly differing relative size, we use a 'small o' notation: we write $f(x, \varepsilon) = o(g(x, \varepsilon))$ as $\varepsilon \to 0$ if, given any $\delta > 0$, there exists an $\varepsilon_0 > 0$ such that $|f(x, \varepsilon)| \le \delta|g(x, \varepsilon)|$ for $0 < \varepsilon \le \varepsilon_0$. If in addition ε_0 is independent of x, we say that $f(x, \varepsilon) = o(g(x, \varepsilon))$ as $\varepsilon \to 0$ uniformly for $x \in D$.

An asymptotic sequence $\{\phi_n(\varepsilon)\}$, $n = 1, 2, \ldots$, is a sequence of functions of ε such that

$$\phi_{n+1}(\varepsilon) = o(\phi_n(\varepsilon)) \quad \text{as } \varepsilon \to 0 \quad \text{for each } n.$$

Asymptotic sequences are the building blocks from which one constructs asymptotic expansions.

Let $u(x, \varepsilon)$ be defined for all $x \in D$ and all sufficiently small ε. Let $\{\phi_n(\varepsilon)\}$ be an asymptotic sequence. The series $\sum_{n=1}^{N} u_n(x)\phi_n(\varepsilon)$, where N may be finite or infinite, is said to be the *asymptotic expansion* of u with respect to $\{\phi_n\}$ as $\varepsilon \to 0$, if for each $M \in \{1, \ldots, N\}$ we have

$$u(x, \varepsilon) - \sum_{n=1}^{M} u_n(x)\phi_n(\varepsilon) = o(\phi_M) \quad \text{as } \varepsilon \to 0. \tag{2.1}$$

In this case we write $u(x, \varepsilon) \sim \sum_{n=1}^{N} u_n(x)\phi_n(\varepsilon)$. This asymptotic expansion is *uniform in D* if (2.1) holds true uniformly for $x \in D$.

To introduce our final asymptotic concept, we take a simple example involving functions of ε that have no additional dependence on a variable x.

Example 2.3. One can easily show that one solution u_ε of the algebraic equation $u_\varepsilon^2 + \varepsilon u_\varepsilon - 1 = 0$, where ε is a small positive parameter, satisfies $u_\varepsilon = 1 + \mathcal{O}(\varepsilon)$. Thus, as $\varepsilon \to 0$, this solution approaches the solution $u_0^{(1)} = 1$ of the problem $u_0^2 - 1 = 0$. Similarly, the other solution of $u_\varepsilon^2 + \varepsilon u_\varepsilon - 1 = 0$ approaches the other solution $u_0^{(2)} = -1$ of $u_0^2 - 1 = 0$. Thus, as $\varepsilon \to 0$, the solutions of the original problem approach the solutions of the modified problem with ε set equal to zero.

The situation is different for the solutions $v_\varepsilon^{(1)}$ and $v_\varepsilon^{(2)}$ of the equation $\varepsilon v_\varepsilon^2 + v_\varepsilon - 1 = 0$. An application of the quadratic formula and binomial theorem shows that

$$v_\varepsilon^{(1)} = 1 - \varepsilon + 2\varepsilon^2 - 5\varepsilon^3 + \cdots, \qquad v_\varepsilon^{(2)} = -\varepsilon^{-1} - 1 + \varepsilon - 2\varepsilon^2 + \cdots.$$

Hence, as $\varepsilon \to 0$, one has $v_\varepsilon^{(1)} \to 1$ (the solution of the modified problem $v_0 - 1 = 0$ obtained by setting $\varepsilon = 0$) but $v_\varepsilon^{(2)} \to -\infty$.

The first part of this example is a *regular perturbation* problem: the behaviour of the solution when the perturbation parameter ε reaches its limit value of 0 is quite similar to the behaviour when ε is near but not equal to 0. The second part is a *singular perturbation* problem, where reaching the limit value of the parameter causes some significant change in the solution (here $v_\varepsilon^{(2)}$ is not close to $v_0 = 1$). As we shall see, convection-diffusion problems form a class of singular perturbation problems.

3. Convection-diffusion problems in one dimension

In this section we shall examine the asymptotic nature of solutions to convection-diffusion problems in one dimension, which will provide useful insights. The behaviour of the derivatives of these solutions, which is critical for the numerical analysis that follows later, is also discussed. Finally these two lines of attack are combined in the final subsection on Shishkin decompositions of solutions.

3.1. Asymptotic analysis

To avoid excessive detail, we do not begin with the most general situation but work instead with the two-point boundary-value problem

$$Lu(x) := -\varepsilon u''(x) + u'(x) = f(x) \quad \text{for } 0 < x < 1, \tag{3.1a}$$

$$u(0) = u(1) = 0, \tag{3.1b}$$

where we recall from Section 1.3 that ε is a small positive parameter. Assume that $f \in C^\infty[0,1]$. This is a convection-diffusion problem: the coefficient of the first-order derivative is much larger in magnitude than the coefficient of the second-order derivative.

It would be more precise to write $u(x, \varepsilon)$ for the solution of (3.1), but for convenience we use $u(x)$.

If we set $\varepsilon = 0$ then (3.1a) becomes a first-order differential equation – a significant change – so we expect that this problem is singularly perturbed. A more careful definition of singularly perturbed (with respect to

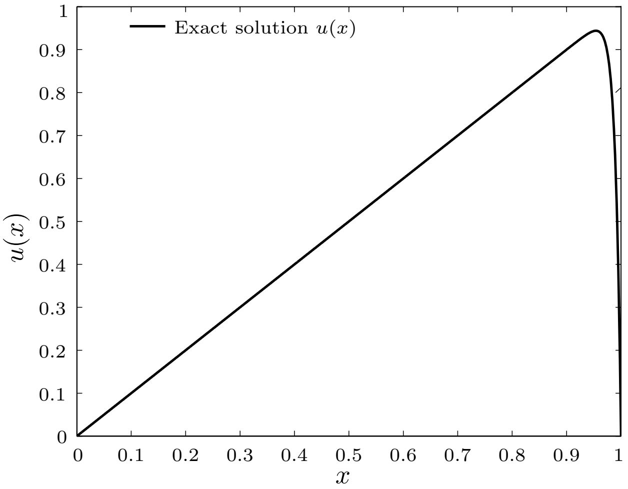

Figure 3.1. Graph of (3.3) with $\varepsilon = 0.01$.

the maximum norm) is that there exists $\hat{x} \in [0,1]$ (in fact $\hat{x} = 1$ for this problem) such that

$$\lim_{\varepsilon\to 0} \lim_{x\to\hat{x}} u(x) \neq \lim_{x\to\hat{x}} \lim_{\varepsilon\to 0} u(x). \tag{3.2}$$

Example 3.1. To get some immediate insight into (3.1), consider the simple case where $f(x) \equiv 1$. Then

$$u(x) = x - \frac{\mathrm{e}^{-(1-x)/\varepsilon} - \mathrm{e}^{-1/\varepsilon}}{1 - \mathrm{e}^{-1/\varepsilon}} \quad \text{for } 0 \le x \le 1. \tag{3.3}$$

See Figure 3.1.

One can check that (3.2) holds true with $\hat{x} = 1$. We say that $u(x)$ has a *boundary layer* at $x = 1$: this is a narrow region where u is bounded independently of ε but its derivatives blow up as $\varepsilon \to 0$ (differentiate (3.3) to see that $u'(1) \approx -1/\varepsilon,\ u''(1) \approx -1/\varepsilon^2$, *etc.*). All the important features of the general problem (3.1) are also present in (3.3).

For the differential operator L of (3.1), only with certain exceptional combinations of the boundary conditions and f does the problem fail to be singularly perturbed. For example, if $f(x) \equiv 1$ and the boundary conditions were changed to $u(0) = 0,\ u(1) = 1$, then the solution of (3.1) becomes the well-behaved function $u(x) = x$ and (3.2) is no longer satisfied for any $\hat{x} \in [0,1]$, *i.e.*, (3.1) is now a regular perturbation problem.

The standard way of generating an asymptotic expansion for the solution

$u(x)$ of a boundary-value problem such as (3.1) is to assume that

$$u(x)=\sum_{n=0}^{\infty}u_n(x)\varepsilon^n. \tag{3.4}$$

Substituting this into (3.1a) yields

$$-\varepsilon\sum_{n=0}^{\infty}u_n''(x)\varepsilon^n+\sum_{n=0}^{\infty}u_n'(x)\varepsilon^n=f(x).$$

Comparing coefficients of powers of ε, we get

$$u_0'(x)=f(x),\quad u_1'(x)=u_0''(x),\quad u_2'(x)=u_1''(x),\quad \textit{etc.}$$

Each of these is a first-order ordinary differential equation and should have associated with it a single boundary condition. But the boundary conditions (3.1b) seem to imply that $u_n(0)=u_n(1)=0$ for all n: twice as many conditions as we can handle! It turns out that if we require for all n that $u_n(0)=0$ and place no condition on $u_n(1)$, then we shall be able to build an asymptotic expansion – but no other way of using the boundary conditions (*e.g.*, $u_n(1)=0$ for all n) works. Recalling (3.3), which is qualitatively similar to the solution of (3.1), in forming the asymptotic expansion (3.4) *one must discard boundary conditions where a layer occurs.*

We can now solve for the $u_n(x)$:

$$u_0(x)=\int_0^x f(t)\,\mathrm{d}t,\quad u_1(x)=f(x)-f(0),\quad u_2(x)=f'(x)-f'(0),\quad \textit{etc.}$$

Thus (3.4) becomes

$$\sum_{n=0}^{\infty}\big(F^{(n)}(x)-F^{(n)}(0)\big)\varepsilon^n, \tag{3.5}$$

where $F(x):=\int_0^x f(t)\,\mathrm{d}t$. One can show that

$$u(x)=\sum_{n=0}^{M}\big(F^{(n)}(x)-F^{(n)}(0)\big)\varepsilon^n+o(\varepsilon^M)$$

for each $M\geq 0$, but this expansion is not uniform for $0\leq x\leq 1$; it is uniform only for $0\leq x\leq\delta$ where δ is any fixed constant in $(0,1)$. This situation is unsatisfactory since at $x=1$ we expect that $u(x)$ has a boundary layer, which is its most interesting feature. Of course the inadequacy of the expansion near $x=1$ is unsurprising because its construction has ignored the boundary condition $u(1)=0$ from (3.1b).

What can be done to improve the asymptotic expansion? Consider the special case $f(x)\equiv 1$. Then (3.5) collapses to the function x, but the exact solution is given by (3.3). In this formula the terms $\mathrm{e}^{-1/\varepsilon}$ are ‘exponentially small' (*i.e.*, negligible compared with any integer power of ε) and can safely be ignored. What is missing from (3.5) is some approximation of $\mathrm{e}^{-(1-x)/\varepsilon}$, that is, some function of the variable $(1-x)/\varepsilon$ must be added to (3.5).

A standard systematic way of introducing such a function is as follows: define the *stretched variable* $\rho := (1-x)/\varepsilon$ and rewrite the differential equation as a function of ρ instead of a function of x. (Note: in the formula for ρ, the number 1 appears as the location of the layer, but the division by ε is more subtle – the purpose of the change of variable is to achieve the same dependence on ε in all the relevant terms of the transformed differential operator, but the exact scaling to use in general singular perturbation problems is not always obvious.)

Thus set $\tilde{u}(\rho) \equiv u(x)$ for $0 < \rho < 1/\varepsilon$ (corresponding to $0 < x < 1$). In fact we work with $0 < \rho < \infty$ as it is slightly simpler. Now

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{\mathrm{d}\tilde{u}}{\mathrm{d}\rho}.\frac{\mathrm{d}\rho}{\mathrm{d}x} = -\frac{1}{\varepsilon}\tilde{u}_\rho \quad \text{and} \quad u''(x) = \frac{1}{\varepsilon^2}\tilde{u}_{\rho\rho},$$

so writing the differential operator in terms of ρ we get

$$-\varepsilon u'' + u' = -\frac{1}{\varepsilon}\big(\tilde{u}_{\rho\rho} + \tilde{u}_\rho\big) =: \tilde{L}u.$$

The original asymptotic expansion $\sum_{n=0}^{\infty} u_n(x)\varepsilon^n$ in (3.4) satisfied

$$L\left(\sum_{n=0}^{\infty} u_n(x)\varepsilon^n\right) = f,$$

so the correction $v(\rho)$ that is to be added to this expansion must satisfy $\tilde{L}v = 0$, *i.e.*, $v_{\rho\rho} + v_\rho = 0$. This second-order differential equation needs boundary conditions on $v(\rho)$ at both $\rho = 0$ (which corresponds to $x = 1$) and at $\rho = \infty$. We can now finally enforce the original boundary condition $u(1) = 0$ by requiring that our modified asymptotic expansion satisfies this condition, *i.e.*, that

$$\sum_{n=0}^{\infty} u_n(1)\varepsilon^n + v(0) = 0.$$

We want the function v to act like a boundary layer, which implies that it dies off rapidly as ρ becomes large. Thus it is natural to impose the boundary condition $v(\infty) = 0$.

The two-point boundary value problem that defines v is now completely specified and can be solved explicitly:

$$\begin{aligned} v(\rho) = \mathrm{e}^{-\rho}v(0) &= -\mathrm{e}^{-(1-x)/\varepsilon}\sum_{n=0}^{\infty} u_n(1)\varepsilon^n \\ &= -\mathrm{e}^{-(1-x)/\varepsilon}\sum_{n=0}^{\infty}\big(F^{(n)}(1) - F^{(n)}(0)\big)\varepsilon^n. \end{aligned}$$

Adding this term to (3.5), the new proposed expansion is

$$u_{as}(x) := \sum_{n=0}^{\infty}\big(F^{(n)}(x) - F^{(n)}(0)\big)\varepsilon^n - \mathrm{e}^{-(1-x)/\varepsilon}\sum_{n=0}^{\infty}\big(F^{(n)}(1) - F^{(n)}(0)\big)\varepsilon^n. \tag{3.6}$$

To show that (3.6) is indeed a valid asymptotic expansion, *i.e.*, that $u(x) \sim u_{as}(x)$, set

$$\begin{aligned}\theta_M(x) = u(x) - \sum_{n=0}^{M}\big(F^{(n)}(x) - F^{(n)}(0)\big)\varepsilon^n \\ + \mathrm{e}^{-(1-x)/\varepsilon}\sum_{n=0}^{M}\big(F^{(n)}(1) - F^{(n)}(0)\big)\varepsilon^n\end{aligned}$$

for $M = 0, 1, 2, \ldots$. We shall bound θ_M by means of a suitably chosen barrier function. Now $\theta_M(1) = 0$ and $\theta_M(0) = \mathrm{e}^{-1/\varepsilon}\sum_{n=0}^{M}\big(F^{(n)}(1) - F^{(n)}(0)\big)\varepsilon^n = \mathcal{O}\big(\varepsilon^{M+1}\big)$. Also,

$$\begin{aligned}L\theta_M(x) &= f(x) - \sum_{n=0}^{M}\big[-\varepsilon F^{(n+2)}(x) + F^{(n+1)}(x)\big]\varepsilon^n \\ &= f(x) - F'(x) + \varepsilon^{M+1}F^{(M+2)}(x) \\ &= \varepsilon^{M+1}F^{(M+2)}(x),\end{aligned}$$

where the series telescoped. For each $w \in C[0,1]$, set

$$\|w\|_\infty = \max_{x\in[0,1]}|w(x)|.$$

Define the barrier function $b(x) = C\varepsilon^{M+1}(1+x)$, where the constant $C \geq \|F^{(M+2)}\|_\infty$ is chosen such that $b(0) = C\varepsilon^{M+1} \geq |\theta_M(0)|$. Then $Lb(x) = C\varepsilon^{M+1} \geq |L\theta_M(x)|$ for $0 < x < 1$. By Corollary 2.2, $|\theta_M(x)| \leq b(x) \leq 2C\varepsilon^{M+1}$ for $0 \leq x \leq 1$, and this is $o(\varepsilon^M)$ uniformly for $x \in [0,1]$.

Thus (3.6) is an asymptotic expansion of $u(x)$ that is valid uniformly for $0 \leq x \leq 1$.

Consider now the more general problem

$$\begin{gathered}-\varepsilon u''(x) + a(x)u'(x) + b(x)u(x) = f(x) \quad \text{for } 0 < x < 1, \\ u(0) = A, \quad u(1) = B,\end{gathered} \tag{3.7}$$

where $a(x) > \alpha > 0$ and $b(x) \geq 0$ on [0,1], and A, B are given constants.

Remark 3.2. In fact, given that $a(x) > 0$ on [0,1], one can assume without loss of generality that $b(x) \geq 0$ (so Corollary 2.2 can be invoked), provided that ε is sufficiently small. To see this, set $u(x) = v(x)\mathrm{e}^{kx}$ where the constant k is yet to be chosen; then $Lu = f$ is equivalent to

$$-\varepsilon v''(x) + [a(x) - 2\varepsilon k]v'(x) + [b(x) + ka(x) - \varepsilon k^2]v(x) = f(x)\mathrm{e}^{-kx},$$

and one can choose k such that the coefficients of v' and v are both positive, so v satisfies a differential equation of the desired type. Any numerical method for v will easily yield a numerical solution for u via the transformation $u(x) = v(x)\mathrm{e}^{kx}$.

Lemma 3.3. Let u be the classical solution of (3.7). There exists a constant C such that

$$\|u\|_\infty \le C \tag{3.8}$$

and

$$|u'(0)| \le C. \tag{3.9}$$

Proof. Set $z(x) = u(x) - A$ for $0 \le x \le 1$. Then $z(0) = 0$, $|z(1)| \le |B - A|$ and

$$|Lz(x)| = |f(x) - Ab(x)| \le \|f\|_\infty + |A|\,\|b\|_\infty.$$

Apply Corollary 2.2 to bound $|z(x)|$ by the barrier function

$$\theta(x) = \frac{x}{\alpha}\big(\alpha|B - A| + \|f\|_\infty + |A|\,\|b\|_\infty\big).$$

This immediately implies (3.8), and (3.9) follows from

$$|u'(0)| = \lim_{x\to 0^+} [|z(x)|/x] \le \lim_{x\to 0^+} [\theta(x)/x]. \qquad \square$$

Inequality (3.9) shows that the solution $u(x)$ of (3.7) has no boundary layer at $x = 0$. It will in general have a boundary layer at $x = 1$, like Example 3.1.

Away from $x = 1$, we have $u(x) \approx u_0(x)$, where $u_0(x)$ is the solution of the *reduced problem*

$$a(x)u_0'(x) + b(x)u_0(x) = f(x) \quad \text{for } 0 < x < 1, \qquad u(0) = A. \tag{3.10}$$

This is the same $u_0(x)$ as the first term in (3.4). An analysis similar to that for (3.1) will construct functions $u_n(x)$ and $v_n(x)$ such that, for $k = 0, 1, 2, \ldots,$

$$u(x) = \sum_{n=0}^{k} u_n(x)\varepsilon^n + \sum_{n=0}^{k} v_n(x)\varepsilon^n + \varepsilon^{k+1}R(x, \varepsilon, k), \tag{3.11}$$

where for all i and n we have

$$|u_n^{(i)}(x)| \le C = C(i, n), \qquad |v_n^{(i)}(x)| \le C\varepsilon^{-i}\mathrm{e}^{-\alpha(1-x)/\varepsilon}$$

with $C = C(i, n)$, and $|R(x, \varepsilon, k)| \le C = C(k)$ uniformly for $0 \le x \le 1$. Hence

$$\sum_{n=0}^{\infty} u_n(x)\varepsilon^n + \sum_{n=0}^{\infty} v_n(x)\varepsilon^n$$

is an asymptotic expansion of $u(x)$ that is valid uniformly for $0 \le x \le 1$.

Remark 3.4. If in (3.7) we have $a(x) < 0$ on [0,1], then the change of variable $x \mapsto 1 - x$ reduces the problem to the case $a(x) > 0$ already considered. Thus the essential nature of $u(x)$ remains unaltered except that the boundary layer is now at $x = 0$.

If $a(x)$ changes sign on [0,1] then the solution $u(x)$ may have interior layers and/or boundary layers; see Roos *et al.* (1996, §I.1.2).

Further examples of asymptotic expansions of solutions of singularly perturbed problems can be found in Kevorkian and Cole (1996). For a comprehensive discussion of the construction of asymptotic expansions for a large variety of convection-diffusion problems in n dimensions, see Il'in (1992).

3.2. Bounds on derivatives

Asymptotic expansions of the solution u of a convection-diffusion problem such as (3.1) give us a good idea of how u behaves. To analyse numerical methods, information on the derivatives of u is also needed, and this is now presented.

Consider the general convection-diffusion problem

$$\begin{gathered} Lu(x) := -\varepsilon u''(x) + a(x)u'(x) + b(x)u(x) = f(x) \quad \text{for } 0 < x < 1, \qquad (3.12) \\ u(0) = u(1) = 0, \end{gathered}$$

where $a(x) > \alpha > 0$ and $b(x) \geq 0$ on [0,1]. Assume that a and b lie in $C^\infty[0,1]$.

We already know from Lemma 3.3 that $|u'(0)| \leq C$; the next result, which is due to Kellogg and Tsan (1978), tells us what happens on all of [0,1].

Theorem 3.5. For $i = 0, 1, 2, \ldots$ and ε sufficiently small, there exists a constant $C = C(i)$ such that

$$|u^{(i)}(x)| \leq C\bigl(1 + \varepsilon^{-i}\mathrm{e}^{-\alpha(1-x)/\varepsilon}\bigr) \quad \text{for } 0 \leq x \leq 1. \qquad (3.13)$$

Proof. The case $i = 0$ is covered by Lemma 3.3. The case $i = 1$ is proved by a clever but elementary argument using integrating factors. Then the result can be deduced for $i = 2, 3, \ldots$ by an inductive argument. See Kellogg and Tsan (1978) or Roos *et al.* (1996, p. 9) for the details. □

Remark 3.6. If in (3.12) we replace the Dirichlet boundary condition $u(1) = 0$ at the layer by a *Neumann boundary condition* $u'(1) = k$ (for some constant k), then (3.13) becomes

$$|u^{(i)}(x)| \leq C\bigl(1 + \varepsilon^{1-i}\mathrm{e}^{-\alpha(1-x)/\varepsilon}\bigr) \quad \text{for } i = 0, 1, 2, \ldots \text{ and } 0 \leq x \leq 1.$$

That is, the first-order derivative of u is bounded at $x = 1$ as $\varepsilon \to 0$, but higher-order derivatives still blow up. On a plot of $u(x)$ there is no obvious layer at $x = 1$, but the function is nevertheless not entirely tame.

One might ask: Can we not obtain bounds on derivatives of u simply by differentiating uniform asymptotic expansions, such as (3.11)? This is tempting, but we have developed no theory that controls the difference between a derivative of u and the same derivative of its asymptotic expansion. In general the differentiation of asymptotic expansions of functions is not rigorously justified, but for solutions of elliptic differential equations a theory can be established. This approach is outlined in Theorem 3.7 below and leads not only to bounds on the derivatives of u but also to a convenient decomposition of u.

3.3. Decompositions of the solution

In Theorems 3.7 and 3.9 we show that $u(x)$ can be written as the sum of a well-behaved term and a layer term. Such decompositions of u aid our insight when constructing accurate numerical methods and are often needed in the rigorous analysis of such methods.

Theorem 3.7. (standard decomposition of $\boldsymbol{u}$) Let u be the solution of (3.12). Let q be a positive integer. Then there is a splitting $u = S + E$ such that, for $0 \le j \le q$, the inequalities

$$\|S^{(j)}\|_\infty \le C \quad \text{and} \quad |E^{(j)}(x)| \le C\varepsilon^{-j}\mathrm{e}^{-\alpha(1-x)/\varepsilon} \quad \text{for } 0 \le x \le 1$$

hold true for some constant $C = C(q)$.

Proof. Recall the standard asymptotic expansion of $u(x)$ given in (3.11), and for convenience write $R(x)$ for the remainder $R(x, \varepsilon, k)$. Observe that we have a bound only on $\|R\|_\infty$: no information is available on the derivatives of $R(x)$. As the u_n and v_n are computed explicitly and $Lu = f$, one can determine $LR(x)$ from (3.11). Now the deep *a priori* estimates of Schauder for elliptic differential equations (Ladyzhenskaya and Ural'tseva 1968, p. 110) will yield the bound $\|R^{(j)}\|_\infty \le C\varepsilon^{-j}$ for $0 \le j \le q$.

Choosing $k = q - 1$ in (3.11), set

$$S = \sum_{n=0}^{q-1} u_n(x)\varepsilon^n + \varepsilon^q R(x) \quad \text{and} \quad E(x) = \sum_{n=0}^{q-1} v_n(x)\varepsilon^n.$$

The result now follows immediately from what is known about the terms in S and E. □

In this theorem and other similar results, S is called the *smooth part* of u and E the *layer part*. In the literature dealing with singularly perturbed differential equations, 'smooth' is generally used in this non-standard way to mean that a function has certain low-order derivatives bounded independently of the perturbation parameter.

Theorem 3.5 is adequate when proving convergence of some numerical methods for (3.12), but for others it is convenient to invoke Theorem 3.7 in order to analyse separately the smooth and layer parts of u. At first sight Theorem 3.7 seems the stronger of the two results, but this is not the case, as Linß (2001) showed.

Theorem 3.8. Theorems 3.5 and 3.7 are equivalent.

Proof. Clearly Theorem 3.7 implies Theorem 3.5.

For the converse implication, assume that (3.13) holds true and let q be an arbitrary but fixed positive integer. Set $x^* = 1 - (q\varepsilon/\alpha)\ln(1/\varepsilon)$ and define $S(x) = u(x)$ for $0 \le x \le x^*$. Then (3.13) and the choice of x^* ensure that $|S^{(j)}(x)| \le C$ for $0 \le j \le q$ and $0 \le x \le x^*$. Consequently one can (using a Taylor expansion of $S(x)$ about $x = x^*$) extend S to $[0,1]$ with $|S^{(j)}(x)| \le C$ for $0 \le j \le q$ and $0 \le x \le 1$.

Now set $E = u - S$. Then $E(x) \equiv 0$ for $0 \le x \le x^*$, and for $x^* < x \le 1$ we have

$$|E^{(q)}(x)| \le |u^{(q)}(x)| + |S^{(q)}(x)| \le C\big(1 + \varepsilon^{-q}\mathrm{e}^{-\alpha(1-x)/\varepsilon}\big) \le C\varepsilon^{-q}\mathrm{e}^{-\alpha(1-x)/\varepsilon}$$

from the definition of x^*. Using induction, we integrate $E^{(k)}(x)$ for $k = q, q-1, \ldots, 1$ to get

$$\begin{aligned}|E^{(k-1)}(x)| &\le \left|\int_{x^*}^{x} E^{(k)}(s)\,\mathrm{d}s\right| \\ &\le C\int_{x^*}^{x} \varepsilon^{-k}\mathrm{e}^{-\alpha(1-s)/\varepsilon}\,\mathrm{d}s \\ &\le C\varepsilon^{-(k-1)}\mathrm{e}^{-\alpha(1-x)/\varepsilon}\end{aligned}$$

for $x^* < x \le 1$. □

For the analysis of certain finite difference methods on Shishkin meshes (which we shall meet in Section 6), we need a decomposition of u with a further property that is originally due to Shishkin: see the references in Farrell, Hegarty, Miller, O'Riordan and Shishkin (2000) and Miller, O'Riordan and Shishkin (1996). By a modification of the construction of the asymptotic expansion (3.11) as described in Dobrowolski and Roos (1997) and Miller *et al.* (1996), we can prove the following strengthening of Theorem 3.7.

Theorem 3.9. (Shishkin decomposition of u) Let u be the solution of (3.12). Let q be a nonnegative integer. Then there is a splitting $u = S + E$ such that, for $0 \le j \le q$, the inequalities

$$\|S^{(j)}\|_\infty \le C \quad \text{and} \quad |E^{(j)}(x)| \le C\varepsilon^{-j}\mathrm{e}^{-\alpha(1-x)/\varepsilon} \quad \text{for } 0 \le x \le 1 \tag{3.14}$$

hold true for some constant $C = C(q)$, and in addition

$$LS(x) = f(x) \quad \text{and} \quad LE(x) = 0 \quad \text{for } 0 \le x \le 1.$$

4. Finite difference methods in one dimension

Consider the convection-diffusion problem

$$Lu(x) := -\varepsilon u''(x) + a(x)u'(x) + b(x)u(x) = f(x) \quad \text{for } 0 < x < 1, \tag{4.1}$$
$$u(0) = u(1) = 0,$$

where $0 < \varepsilon \ll 1$, $a(x) > \alpha > 0$ and $b(x) \geq 0$ on [0,1]. Assume that a and b lie in $C^\infty[0,1]$.

Let N be a positive integer. Partition [0,1] by the equidistant mesh $x_i = ih$ for $i = 0, \ldots, N$, where $h := 1/N$. We aim to compute an approximation $\{u_i^N\}_{i=0}^N$ of $\{u_i\}$; here and subsequently we write u_i for $u(x_i)$, a_i for $a(x_i)$, *etc.*

Standard discretizations of differential equations use a *central difference approximation* of the convective term. That is, one approximates $u'(x_i)$ by $(u_{i+1}^N - u_{i-1}^N)/(2h)$. Using this discretization and the standard approximation $(u_{i-1}^N - 2u_i^N + u_{i+1}^N)/h^2$ of $u''(x_i)$ produces a difference scheme whose matrix B is tridiagonal with ith row

$$\left(0 \ldots 0 \quad -\frac{\varepsilon}{h^2} - \frac{a_i}{2h} \quad \frac{2\varepsilon}{h^2} + b_i \quad -\frac{\varepsilon}{h^2} + \frac{a_i}{2h} \quad 0 \ldots 0\right) \tag{4.2}$$

for $i = 1, \ldots, N-1$. The 0th and Nth rows of B, which incorporate the boundary conditions, are $(1\ 0 \ldots 0)$ and $(0 \ldots 0\ 1)$. The right-hand side of the scheme is $(0\ f_1\ f_2 \ldots f_{N-1}\ 0)^T$.

In the particular case where $a(x) \equiv f(x) \equiv 1$ and $b(x) \equiv 0$, the solution of this difference scheme is

$$u_i^N = x_i - \frac{r^{N-i} - r^N}{1 - r^N}, \quad \text{where } r = \frac{2\varepsilon - h}{2\varepsilon + h}.$$

In practice one usually has $N \ll 1/\varepsilon$, so $\varepsilon \ll h$ and $r \approx -1$. Consequently the computed solution will oscillate as i varies, quite unlike the true solution (3.3). See Figure 4.1.

Remark 4.1. To see that the computed solution is inaccurate near $x = 1$ in the general case (4.1), consider (4.2) with $j = N - 1$. Taking $\varepsilon \ll h^2$, this equation is essentially

$$f_{N-1} = \frac{a_{N-1}(u_N^N - u_{N-2}^N)}{2h} + b_{N-1}u_{N-1}^N = -\frac{a_{N-1}u_{N-2}^N}{2h} + b_{N-1}u_{N-1}^N,$$

on applying the boundary condition. That is, $u_{N-2}^N = \mathcal{O}(h)$; but because of the boundary layer in $u(x)$ at $x = 1$ we expect that u_{N-2} is not close to zero. Thus u_{N-2}^N is far from u_{N-2}, and this is due to oscillations in the computed solution.

What has gone wrong?

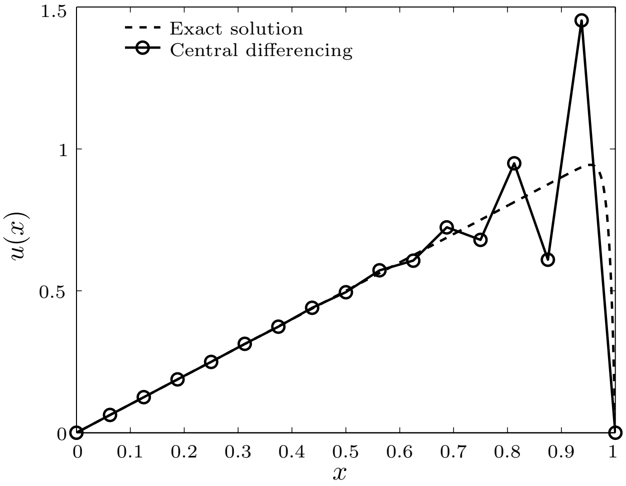

Figure 4.1. Solution to (4.1) with $\varepsilon = 0.01$, $a \equiv 1$, $b \equiv 0$, $f \equiv 1$ computed by central differencing with $N = 16$.

A square matrix $A = (A_{ij})$ is said to be an *M-matrix* if $A_{ij} \leq 0$ for all $i \neq j$ and A^{-1} exists with $(A^{-1})_{ij} \geq 0$ for all i, j. Difference schemes that employ M-matrices are common because they are desirable: they are generally stable, and more amenable to analysis. Our central difference scheme above fails to satisfy the M-matrix sign condition on the off-diagonal entries since $B_{i,i+1} > 0$ when ε is small relative to h. If $h\|a\|_\infty \leq 2\varepsilon$, then the sign condition is satisfied and it turns out that the difference method gives an acceptable computed solution, but to enforce this inequality when ε is small is impractical in many problems (especially in partial differential equations) since it can lead to an intolerably large number of mesh points.

The second M-matrix requirement – that A^{-1} exists with $(A^{-1})_{ij} \geq 0$ for all i and j – does not seem easy to verify in practice. Fortunately there are more tractable alternatives, as stated in the next result.

Lemma 4.2. *Suppose that the $n \times n$ matrix $A = (A_{ij})$ satisfies $A_{ij} \leq 0$ for all $i \neq j$. Then A^{-1} exists and $(A^{-1})_{ij} \geq 0$ for all i, j if either of the following two conditions is satisfied:*

(i) *there exists a vector $\mathbf{v} > 0$ such that $A\mathbf{v} > 0$ (here and below, inequalities like this are understood to hold true component-wise)*

(ii) *A is strictly diagonally dominant with $a_{ii} > 0$ for all i.*

Proof. For a proof that the first condition is sufficient, see Bohl (1981); for the second, see, *e.g.*, Quarteroni and Valli (1994, Lemma 2.1.1). □

One can often construct a vector that satisfies condition (i) of Lemma 4.2 by finding a function $w(x)$ such that $w > 0$ and $Lw > 0$, then forming $\mathbf{v}$ by restricting w to the mesh.

For M-matrices we have discrete analogues of Lemma 2.1 and Corollary 2.2.

Lemma 4.3. (discrete maximum principle) Let A be an M-matrix. If $\mathbf{v}$ is a vector with $A\mathbf{v} \geq 0$, then $\mathbf{v} \geq 0$.

Proof. $\mathbf{v} = (A^{-1})(A\mathbf{v}) \geq 0$, because $A^{-1} \geq 0$ and $A\mathbf{v} \geq 0$. □

Lemma 4.4. (discrete barrier function) Let A be an M-matrix. If $\mathbf{v}_1, \mathbf{v}_2$ are vectors such that $|A\mathbf{v}_1| \leq A\mathbf{v}_2$, then $|\mathbf{v}_1| \leq \mathbf{v}_2$.

Proof. Now $A(\mathbf{v}_2 - \mathbf{v}_1) \geq 0$, so $\mathbf{v}_2 - \mathbf{v}_1 \geq 0$ by Lemma 4.3. Similarly $\mathbf{v}_2 + \mathbf{v}_1 \geq 0$, and the result follows. □

The boundary data requirement of Corollary 2.2 seems to be absent from Lemma 4.4, but this is deceptive: the first and last rows of A include this information (see the construction of the matrix B above).

Returning to our difference scheme, we see that the 'incorrect' sign of $B_{i,i+1}$ comes from the central difference approximation of $u'(x_i)$. This approximation is generally recommended in basic courses in numerical methods because it gives $O(h^2)$ consistency error when $\varepsilon = 1$, but this is irrelevant when the method is (as we found here) unstable. To cure the instability, for convection-diffusion problems one can approximate $u'(x_i)$ by the *simple upwinding* formula $(u_i^N - u_{i-1}^N)/h$. Although the consistency error is now only $O(h)$ when $\varepsilon = 1$, the ith row of the scheme is

$$\left(0 \ldots 0 \quad -\frac{\varepsilon}{h^2} - \frac{a_i}{h} \quad \frac{2\varepsilon}{h^2} + \frac{a_i}{h} + b_i \quad -\frac{\varepsilon}{h^2} \quad 0 \ldots 0\right).$$

Hence, writing A for the associated $(N+1)\times(N+1)$ matrix that incorporates the boundary conditions, we have $A_{ij} \leq 0$ for $i \neq j$, as desired.

Lemma 4.5. The matrix A associated with the simple upwind scheme is an M-matrix.

Proof. Clearly $A_{ij} \leq 0$ for $i \neq j$. Define the vector $\mathbf{v}$ by $v_i = 1 + x_i$ for $i = 0, \ldots, N$. Then a simple calculation shows that $(Av)_i \geq \min\{1, \alpha\} > 0$ for all i. The result now follows from Lemma 4.2. □

Note that upwinding for (3.12) uses the one-sided difference $(u_i^N - u_{i-1}^N)/h$ to approximate $u'(x_i)$, but the alternative choice of $(u_{i+1}^N - u_i^N)/h$ would not give the correct sign pattern in the matrix. Upwinding means taking a one-sided difference *on the side away from the layer*, so for ε small relative to h^2 the scheme essentially decouples the boundary condition at $x = 1$

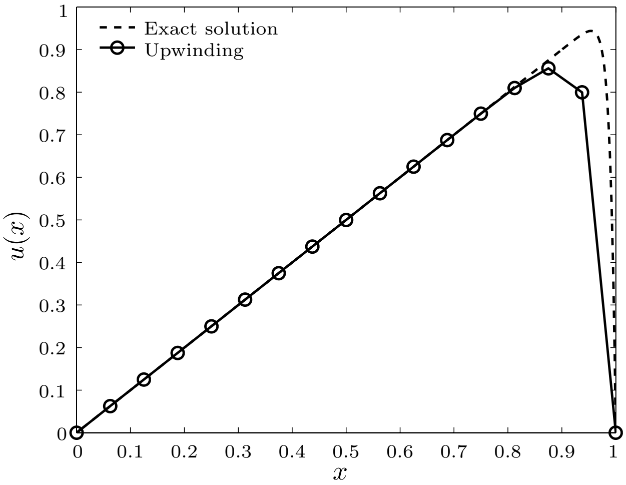

Figure 4.2. Solution to (4.1) with $\varepsilon = 0.01$, $a \equiv 1$, $b \equiv 0$, $f \equiv 1$ computed by simple upwinding with $N = 16$.

from the values at the interior nodes; this is exactly what we need to avoid computational infelicities like that of Remark 4.1.

The first satisfactory investigation into the accuracy of simple upwinding is due to Kellogg and Tsan (1978). Their delicate analysis derived a tight bound on the consistency error of the method, then converted this to the following convergence result by means of discrete barrier functions.

Theorem 4.6. (error bound for simple upwinding on an equidistant mesh) Let $\{u_i^N\}_{i=0}^N$ be the solution to (4.1) computed using simple upwinding on an equidistant mesh with N subintervals. Suppose that $h \geq \varepsilon$. Then there exists a constant C such that

$$|u_i - u_i^N| \leq C\left[h + \exp\left(\frac{-\alpha(1-x_i)}{\alpha h + 2\varepsilon}\right)\right] \quad \text{for } i = 0, \ldots, N.$$

No proof of this result is given here since it can be found in Kellogg and Tsan (1978) or Roos *et al.* (1996, §I.2.1.2), and in any case we shall present a related analysis in Section 5.

If x_i is bounded away from 1, then Theorem 4.6 implies that

$$|u_i - u_i^N| \leq C\left[h + \exp\left(\frac{-\alpha(1-x_i)}{\alpha h + 2h}\right)\right] \leq Ch. \tag{4.3}$$

That is, the upwind scheme yields an $\mathcal{O}(h)$-accurate solution away from $x = 1$. But at interior mesh points that lie close to or inside the layer, the scheme is only $\mathcal{O}(1)$-accurate. See Figure 4.2.

Remark 4.7. The error bound (4.3) is sharp and can lead to disconcerting and puzzling results in numerical experiments. Suppose that for a given convection-diffusion problem, initially we have an equidistant mesh with $h \gg \varepsilon$, so all mesh points in (0,1) lie well outside the layer. Now consider what happens if we repeatedly bisect each interval and compute a fresh solution. At first the interior mesh points remain outside the layer, so by (4.3) the numerical results show first-order convergence of the maximum nodal error. But as we continue to bisect the mesh, eventually mesh points begin to move into the layer – where the accuracy of the computed solution is only $O(1)$ – so at this stage mesh bisection causes the maximum nodal error to *increase*.

While upwinding does remove unnatural oscillations from the computed solution, we pay a price for this: the layers in the computed solution are excessively smeared, *i.e.*, are not as steep as they should be. See Figure 4.2. To put this another way, upwinding seems to produce an accurate solution for a different problem where the diffusion coefficient is much greater than ε. We now make this visual observation more precise.

The simple upwinding discretization is

$$\begin{aligned}
&(-\varepsilon u'' + au' + bu)(x_i)\\
&\quad\mapsto \frac{-\varepsilon}{h^2}\big(u_{i+1}^N - 2u_i^N + u_{i-1}^N\big) + \frac{a_i}{h}\big(u_i^N - u_{i-1}^N\big) + b_i u_i^N\\
&\quad= -\left(\varepsilon + \frac{ha_i}{2}\right)\frac{1}{h^2}\big(u_{i+1}^N - 2u_i^N + u_{i-1}^N\big) + \frac{a_i}{2h}\big(u_{i+1}^N - u_{i-1}^N\big) + b_i u_i^N.
\end{aligned}$$

That is, upwinding applied to $Lu = f$ is the same method as standard central differencing applied to the modified differential equation $\tilde{L}u := -(\varepsilon + ha/2)u'' + au' + bu = f$. The diffusion coefficient in this modified differential equation is so large (relative to ε) that central differencing produces an M-matrix and yields an accurate approximation of the true solution of $\tilde{L}u = f$, but of course near $x = 1$ this solution is not close to the solution of $Lu = f$.

The amount $ha(x)/2$ by which the diffusion coefficient was apparently increased is called the *artificial diffusion* introduced by upwinding.

This relationship between simple upwinding, $Lu = f$ and $\tilde{L}u = f$ opens the door to a flood of possibilities: one can choose a certain amount of artificial diffusion to add to the problem $Lu = f$, then apply a standard (non-upwinded) numerical method, with the aim of retaining stability (*i.e.*, excluding oscillations) while minimizing the smearing of layers in the computed solution. Pursuing this approach turns out to be quite fruitful; in fact, stable numerical methods on uniform meshes for convection-diffusion ODEs are usually equivalent to modifying the diffusion in the original differential equation, then applying a standard method such as central differencing – but for PDEs, the connection may be less straightforward.

To summarize: when a standard numerical method is applied to a convection-diffusion problem, if there is too little diffusion then the computed solution is often oscillatory, while if there is too much diffusion, then the computed layers are smeared.

We now consider difference schemes that are accurate both outside and inside the boundary layer. A difference scheme on a family of meshes is said to be *robust* or *uniformly convergent* (*with respect to* ε) *of order* $\beta > 0$ *in the discrete* L^∞ *norm* if its solution $\{u_i^N\}$ satisfies $|u_i - u_i^N| \le CN^{-\beta}$ for $i = 0, \ldots, N$ and all sufficiently small H, independently of ε. Here N is the number of mesh intervals, H is the mesh diameter and β is some positive constant that is independent of the mesh and of ε.

A uniformly convergent scheme must address explicitly the exponential nature of the layer part of the solution u, as the next result shows.

Theorem 4.8. (necessary conditions for uniform convergence on an equidistant mesh) Assume that we have an equidistant mesh of width h. Suppose that a difference scheme for the problem $-\varepsilon u'' + au' = f$, $u(0) = u(1) = 0$, with a and f positive constants, can be written in the form

$$\theta_- u_{i-1}^N + \theta_0 u_i^N + \theta_+ u_{i+1}^N = hf_i \quad \text{for } i = 1, \ldots, N-1, \qquad u_0^N = u_N^N = 0, \tag{4.4}$$

where each $\theta = \theta(h, \varepsilon)$ depends only on the ratio h/ε. If the scheme is uniformly convergent for some $\beta > 0$, then

$$\theta_- + \theta_0 + \theta_+ = 0 \quad \text{and} \quad \mathrm{e}^{-ah/\varepsilon}\theta_- + \theta_0 + \mathrm{e}^{ah/\varepsilon}\theta_+ = 0. \tag{4.5}$$

Proof. The idea is to use uniform convergence to replace the u_j^N in (4.4) by u_j, then investigate what happens as $h \to 0$ in the special case where h/ε is held constant, so each θ remains constant. See Roos *et al.* (1996, p. 40) for the details. □

The hypothesis of Theorem 4.8 that each θ depend only on the ratio h/ε is not restrictive. The first condition in (4.5) is satisfied by all plausible difference schemes; it is the second condition that distinguishes uniformly convergent schemes. Simple upwinding fails to satisfy that second condition.

Example 4.9. On equidistant meshes, the best-known uniformly convergent scheme for (4.1) is the *Il'in–Allen–Southwell difference scheme*. Allen and Southwell (1955) proposed it without any analysis of its behaviour, then it was independently rediscovered by Il'in (1969), who gave a complicated analysis of its convergence. The scheme is

$$-\frac{a_i \mathrm{e}^{\rho_i}}{h(\mathrm{e}^{\rho_i} - 1)} u_{i-1}^N + \left[\frac{a_i(\mathrm{e}^{\rho_i} + 1)}{h(\mathrm{e}^{\rho_i} - 1)} + b_i\right] u_i^N - \frac{a_i}{h(\mathrm{e}^{\rho_i} - 1)} u_{i+1}^N = f_i$$
$$\text{for } i = 1, \ldots, N-1,$$

where $\rho_i = ha_i/\varepsilon$, with $u_0^N = u_N^N = 0$. It computes $\{u_i\}$ exactly in the special case where a, b and f are constants. Recalling our discussion above of adding artificial diffusion, this scheme is obtained if central differencing is applied to the modified differential equation

$$-\varepsilon\left(\frac{ha(x)}{2\varepsilon}\coth\frac{ha(x)}{2\varepsilon}\right)u''(x)+a(x)u'(x)+b(x)u(x)=f(x).$$

Il'in's scheme can be generated in a wide variety of ways (Roos 1994). In Kellogg and Tsan (1978) discrete barrier functions were used for the first time in the convection-diffusion literature to show that the solution $\{u_i^N\}$ computed by this scheme is uniformly convergent: $|u_i - u_i^N| \le CN^{-1}$ for all i.

The more complicated El Mistikawy–Werle 3-point scheme has the form

$$r_i^- u_{i-1}^N + r_i^0 u_i^N + r_i^+ u_{i+1}^N = q_{i-1}f_{i-1} + q_i^0 f_i + q_{i+1}^+ f_{i+1} \quad \text{for } i=1,\dots,N-1.$$

It achieves second-order uniform convergence on equidistant meshes, *i.e.*, $\max_i |u_i - u_i^N| \le CN^{-2}$.

See Roos *et al.* (1996, §I.2.1.3) for more information on both of these schemes.

Numerical methods like these, whose coefficients involve exponential functions of h/ε, are known collectively as *exponentially fitted* schemes. While they have become less popular in recent years, nevertheless exponential fitting is the mainstay of the FEM package PLTMG and is still widely used in semiconductor device modelling (where the Il'in scheme is known as the Scharfetter–Gummel scheme).

Remark 4.10. In the case of a Neumann boundary condition the layer is weaker (Remark 3.6). Simple upwinding on an equidistant mesh then yields (Linß 2005)

$$|u_i - u_i^N| \le Ch \quad \text{for } i=0,\dots,N.$$

5. Shishkin meshes

When numerically solving a convection-diffusion problem, it seems reasonable to cluster mesh points in the layer – where the solution $u(x)$ is most troublesome – instead of spreading them equidistantly over [0,1]. Graded meshes, where the mesh width gets finer and finer as one moves closer and closer to $x = 1$, have been advocated by several authors; see Roos *et al.* (1996, §I.2.4.2) for references. Since the early 1990s a simpler piecewise-equidistant mesh has been enthusiastically propagated by Shishkin and other authors (Farrell *et al.* 2000, Miller *et al.* 1996).

Consider the convection-diffusion problem (4.1). Set

$$\sigma = \min\{1/2,\ (2/\alpha)\varepsilon\ln N\}.$$

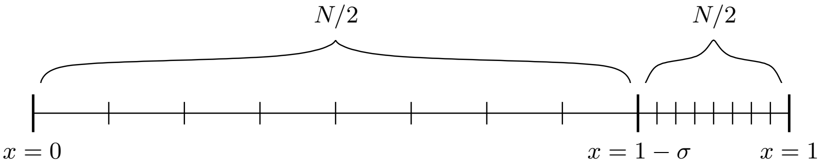

Figure 5.1. Shishkin mesh with $N = 16$.

We shall assume that $\sigma = (2/\alpha)\varepsilon \ln N$, as the other value of σ occurs only when N is exponentially large relative to ε, which is rare in practice. Then the *mesh transition point* is defined to be $1 - \sigma$. Let N be an even integer. Divide each of $[0, 1 - \sigma]$ and $[1 - \sigma, 1]$ by an equidistant mesh with $N/2$ subintervals; see Figure 5.1.

The coarse part of this Shishkin mesh has spacing $H = 2(1 - \sigma)/N$, so $N^{-1} \leq H \leq 2N^{-1}$. The fine part has spacing $h = 2\sigma/N = (4/\alpha)\varepsilon N^{-1} \ln N$, so $h \ll \varepsilon$. On the mesh, $x_i = iH$ for $i = 0, \ldots, N/2$ and $x_i = 1 - (N - i)h$ for $i = N/2 + 1, \ldots, N$. Set $h_i = x_i - x_{i-1}$ for each i. Note that the mesh width h_i changes abruptly at $i = N/2$, and $H/h = \alpha(1 - \sigma)/(2\varepsilon \ln N)$ can be very large.

Remark 5.1. Nonequidistant meshes for convection-diffusion problems are sometimes described as 'layer-resolving' meshes. One might presume that this terminology means that wherever the derivatives of $u(x)$ are large, the mesh is sufficiently fine to control the truncation error of the difference scheme. But the Shishkin mesh does not fully resolve the layer: $|u'(x)| \approx C\varepsilon^{-1} \exp(-\alpha(1 - x)/\varepsilon)$, so $|u'(1 - \sigma)| \approx C\varepsilon^{-1} \exp(-2 \ln N) = C\varepsilon^{-1} N^{-2}$, which in general is large since typically $\varepsilon \ll N^{-1}$. Thus $|u'(x)|$ is still large on part of the first coarse-mesh interval $[x_{N/2-1}, x_{N/2}]$.

This is not a drawback: it is in fact the genius of the Shishkin mesh. For if one set out to construct a two-stage piecewise-equidistant mesh as we have done, but with the additional requirement that the mesh be fine enough to control the local truncation error wherever $|u'(x)|$ is very large, then the number of mesh points required would have to grow like $\ln(1/\varepsilon)$ as ε got smaller. Shishkin's insight was that one could achieve satisfactory theoretical and numerical results without resolving all of the layer. His construction enables us to work with a fixed number $(N + 1)$ of mesh points that is independent of the value of ε.

We apply simple upwinding. For each mesh function $\{v_i\}_{i=0}^N$, set $D_- v_i = (v_i - v_{i-1})/h_i$ and

$$\delta^2 v_i = \frac{2}{h_i + h_{i+1}} \left(\frac{v_{i+1} - v_i}{h_{i+1}} - \frac{v_i - v_{i-1}}{h_i} \right).$$

Our difference scheme is

$$-\varepsilon\delta^2 u_i^N + a_i D_- u_i^N + b_i u_i^N = f_i \quad \text{for } i = 1, \dots, N-1, \qquad u_0^N = u_N^N = 0. \tag{5.1}$$

It is straightforward to check (*cf.* Lemma 4.5) that the matrix L^N associated with (5.1) is an M-matrix. To analyse the convergence of the method, recall the Shishkin decomposition $u = S + E$ of Theorem 3.9 and split the discrete solution $\{u_i^N\}$ in an analogous manner: define $\{S_i^N\}$ and $\{E_i^N\}$ by

$$\begin{aligned} L^N S_i^N &= (LS)_i \quad \text{for } i = 1, \dots, N-1, & S_0^N &= S(0), \quad S_N^N = S(1), \\ L^N E_i^N &= (LE)_i = 0 \quad \text{for } i = 1, \dots, N-1, & E_0^N &= E(0), \quad E_N^N = E(1). \end{aligned}$$

Then $u_i^N = S_i^N + E_i^N$ for all i, and

$$|u_i - u_i^N| = |(S+E)_i - \big(S_i^N + E_i^N\big)| \le |S_i - S_i^N| + |E_i - E_i^N|. \tag{5.2}$$

We shall bound each difference separately.

Lemma 5.2. There exists a constant C_0 such that

$$|S_i - S_i^N| \le C_0 N^{-1} \quad \text{for } i = 0, \dots, N.$$

Proof. As the derivatives of S are bounded, a standard consistency error analysis shows that

$$\begin{aligned} |L^N(S_i - S_i^N)| &= |L^N S_i - (LS)_i| \\ &\le 2\varepsilon \int_{x_{i-1}}^{x_{i+1}} |S'''(x)|\,\mathrm{d}x + a_i \int_{x_{i-1}}^{x_i} |S''(x)|\,\mathrm{d}x \\ &\le C(x_{i+1} - x_{i-1}) \\ &\le CN^{-1} \end{aligned} \tag{5.3}$$

for $i = 1, \dots, N-1$. Set $w_i = C_0 N^{-1} x_i$ for all i, where the positive constant C_0 will be chosen so that $\{w_i^N\}$ is a discrete barrier function for $\{S_i - S_i^N\}$. Now

$$L^N w_i = a_i C_0 N^{-1} + b_i w_i > \alpha C_0 N^{-1} \ge |L^N S_i - (LS)_i|$$

by (5.3), provided that C_0 is a sufficiently large constant. Clearly $w_0 = 0 = |S_0 - S_0^0|$ and $w_N = C_0 N^{-1} \ge 0 = |S_N - S_N^N|$. Thus Lemma 4.4 can be applied and we get $|S_i - S_i^N| \le w_i \le C_0 N^{-1}$ for all i, as desired. □

To bound $|E_i - E_i^N|$ one again invokes Lemma 4.4, but the approach is less direct because $E(x)$ has large derivatives on part of the coarse mesh (see Remark 5.1). We show first that $|E_i|$ and $|E_i^N|$ are small on $[0, 1-\sigma]$ because they decay rapidly away from $x = 1$, then on $[1-\sigma, 1]$ the mesh is so fine that $|E_i - E_i^N|$ can be bounded by a consistency error analysis like that of Lemma 5.2.

From (3.14),

$$|E_i| \le C\mathrm{e}^{-\alpha(1-(1-\sigma))/\varepsilon} = CN^{-2} \le CN^{-1} \quad \text{for } i = 0, \dots, N/2. \tag{5.4}$$

In the next lemma a discrete barrier function is used to show that $|E_i^N|$ also is small when $i \le N/2$. Set

$$Z_i = \prod_{j=1}^{i} \left(1 + \frac{\alpha h_j}{2\varepsilon}\right) \quad \text{for } i = 0, \dots, N.$$

Lemma 5.3. There exists a constant C such that

$$|E_i^N| \le CN^{-1} \quad \text{for } i = 0, \dots, N/2.$$

Proof. For $i = 1, \dots, N$, a calculation shows that there exists a constant $C_1 > 0$ such that

$$L^N Z_i \ge \frac{C_1}{\max\{\varepsilon, h_i\}} Z_i. \tag{5.5}$$

Now $\mathrm{e}^t \ge 1 + t$ for all $t \ge 0$, so

$$\frac{Z_i}{Z_N} = \prod_{j=i+1}^{N} \left(1 + \frac{\alpha h_j}{2\varepsilon}\right)^{-1} \ge \prod_{j=i+1}^{N} \mathrm{e}^{-\alpha h_j/(2\varepsilon)} = \mathrm{e}^{-\alpha(1-x_i)/(2\varepsilon)}. \tag{5.6}$$

Set $Y_i = C_2 Z_i / Z_N$ for $i = 0, \dots, N$. Then $L^N Y_i = (C_2/Z_N) L^N Z_i \ge 0 = |L^N E_i^N|$ for $i = 1, \dots, N-1$, by (5.5) and the definition of $\{E_i^N\}$. Also $Y_N = C_2 \ge |E(1)| = |E_N^N|$ if the constant C_2 is chosen sufficiently large, by the bound on $|E(x)|$ given by inequality (3.14). Finally, (5.6) implies that

$$Y_0 = \frac{C_2 Z_0}{Z_N} \ge C_2 \mathrm{e}^{-\alpha/(2\varepsilon)} \ge C_2 \mathrm{e}^{-\alpha/\varepsilon} \ge |E(0)| = |E_0^N|$$

provided that the constant C_2 is chosen sufficiently large, where we appealed again to (3.14). Thus we can choose C_2 so that the conditions of Lemma 4.4 are satisfied, *i.e.*, $\{Y_i\}$ is a discrete barrier function for $\{E_i^N\}$, and it follows that

$$|E_i^N| \le Y_i = \frac{C_2 Z_i}{Z_N} \quad \text{for all } i. \tag{5.7}$$

But for $i = 0, \dots, N/2$,

$$\begin{aligned} \frac{Z_i}{Z_N} \le \frac{Z_{N/2}}{Z_N} &= \prod_{j=1+N/2}^{N} \left(1 + \frac{\alpha h}{2\varepsilon}\right)^{-1} \\ &= \left(1 + 2N^{-1} \ln N\right)^{-N/2} \\ &\le N^{-1} \mathrm{e}^{(\ln^2 N)/N} \le CN^{-1} \end{aligned}$$

for some constant C (to prove the penultimate inequality, take a logarithm of the left-hand side and notice that $\ln(1+t) \geq t-t^2/2$ for $t \geq 0$). Combining this inequality with (5.7), the proof is complete. □

Corollary 5.4. There exists a constant C such that

$$|E_i - E_i^N| \leq CN^{-1} \quad \text{for } i = 0, \dots, N/2.$$

Proof. This is immediate from (5.4) and Lemma 5.3. □

It remains only to bound $|E_i - E_i^N|$ for $i > N/2$.

Lemma 5.5. There exists a constant C such that

$$|E_i - E_i^N| \leq CN^{-1}\ln N \quad \text{for } i = N/2+1, \dots, N.$$

Proof. We shall apply a discrete barrier function argument at the nodes $\{x_i\}_{i=N/2}^N$ by considering the discretization of a two-point boundary value problem on the interval $[1-\sigma, 1]$. Observe that when L^N is restricted to the interior nodes of this interval it still yields an M-matrix.

Recalling the bounds on $|E^{(j)}(x)|$ in (3.14), a standard consistency error analysis shows that for $i = N/2+1, \dots, N-1$,

$$\begin{aligned} |L^N(E_i - E_i^N)| &= |L^N E_i - (LE)_i| \\ &\leq 2\varepsilon \int_{x_{i-1}}^{x_{i+1}} |E'''(x)| \,\mathrm{d}x + a_i \int_{x_{i-1}}^{x_i} |E''(x)| \,\mathrm{d}x \\ &\leq C \int_{x_{i-1}}^{x_{i+1}} \varepsilon^{-2} \mathrm{e}^{-\alpha(1-x)/\varepsilon} \,\mathrm{d}x \\ &= C\varepsilon^{-1}\mathrm{e}^{-\alpha(1-x_i)/\varepsilon} \sinh(\alpha h/\varepsilon) \\ &\leq C\varepsilon^{-1}N^{-1}(\ln N)\mathrm{e}^{-\alpha(1-x_i)/\varepsilon}, \end{aligned}$$

since $\sinh(\alpha h/\varepsilon) = \sinh(4N^{-1}\ln N) \leq CN^{-1}\ln N$ for all $N \geq 2$.

Set $\phi_i = C_3N^{-1}(\ln N)(1+Z_i/Z_N)$ for $i = N/2, \dots, N$, where the constant C_3 will be chosen later. By (5.5) and (5.6),

$$\begin{aligned} L^N\phi_i &\geq C_3N^{-1}(\ln N)(L^N Z_i)/Z_N \\ &\geq C_3C_1\varepsilon^{-1}N^{-1}(\ln N)Z_i/Z_N \\ &\geq C_3C_1\varepsilon^{-1}N^{-1}(\ln N)\mathrm{e}^{-\alpha(1-x_i)/(2\varepsilon)} \end{aligned}$$

for $i = N/2+1, \dots, N$. Consequently $L^N\phi_i \geq |L^N(E_i - E_i^N)|$ if the constant C_3 is sufficiently large. Furthermore, we can choose C_3 such that

$$\phi_{N/2} = C_3N^{-1}(\ln N)(1 + Z_{N/2}/Z_N) \geq C_3N^{-1}(\ln N) \geq |E_{N/2} - E_{N/2}^N|$$

by Corollary 5.4, and $\phi_N = 2C_3N^{-1}(\ln N) > 0 = |E_N - E_N^N|$.

Thus $\{\phi_i\}$ is a discrete barrier function for $\{E_i - E_i^N\}$, and Lemma 4.4 now implies that for $i = N/2, \dots, N$ we have $|E_i - E_i^N| \leq \phi_i \leq 2C_3N^{-1}\ln N$. □

The final convergence result can now be stated.

Theorem 5.6. (uniform convergence of simple upwinding on a Shishkin mesh) There exists a constant C such that the solution $\{u_i^N\}$ of (5.1) satisfies

$$|u_i - u_i^N| \leq CN^{-1}\ln N \quad \text{for } i = 0, \ldots, N.$$

Proof. Combine (5.2), Lemma 5.2, Corollary 5.4 and Lemma 5.5. □

Observe that uniform convergence is attained even though the consistency error in the maximum norm is not bounded uniformly in ε.

Roos (1996) shows that the condition number of the discrete linear system associated with (5.1) is $O(\varepsilon^{-2}N^2\ln^{-2}N)$, which is uncomfortably large when ε is small, but that an easy preconditioning by diagonal scaling (approximate equilibration) reduces this condition number to $O(N^2\ln^{-1}N)$.

Remark 5.7. The precise choice of mesh transition point $1-\sigma$ in the Shishkin mesh is of both theoretical and computational interest. A careful examination of the proof of Theorem 5.6 reveals that σ should have the form $(k/\alpha)\varepsilon\phi(N)$, where $\phi(N) \to \infty$ but $N^{-1}\phi(N) \to 0$ as $N \to \infty$, and k is some constant. The simplest choice for $\phi(N)$ is $\ln N$. The choice $k = 2$ used in our definition of σ subtly enters the proof of Lemma 5.3 during the final chain of inequalities that bound Z_i/Z_N. How to choose k in an optimal way is discussed in Stynes and Tobiska (1998). It is shown there, using an argument close to our proof of Theorem 5.6, that for a variant of simple upwinding one has

$$|u_i - u_i^N| \leq C\max\{N^{-k}, kN^{-1}\ln N\} \quad \text{for } i = 0, \ldots, N.$$

The sharpness of this bound is confirmed by numerical experiments. Consequently choosing k larger than 1 only slightly diminishes the numerical accuracy of the method, but choosing k smaller than 1 causes a noticeable deterioration in the numerical rate of convergence.

Andreev and Kopteva (1996) show that for central differencing on a Shishkin mesh, the computed solution $\{u_i^N\}$ satisfies $|u_i - u_i^N| \leq CN^{-2}\ln^2 N$ for all i. The proof is difficult as the scheme does not satisfy a discrete maximum principle. Numerical experience (Linß and Stynes 2001*b*) with analogues of this approach for two-dimensional problems reveals that it is quite expensive to solve the discrete linear system efficiently, so we shall not pursue it further.

Remark 5.8. Error estimates in various norms for numerical methods on Shishkin meshes usually include a multiplicative factor $\ln^\beta N$ for some $\beta > 0$. This factor is asymptotically unimportant relative to the main convergence factor N^{-k}, where $k > 0$, but its effect is evident in numerical experiments.

If we work with certain graded meshes (*e.g.*, Bakhvalov meshes) then the $\ln N$ factor disappears so these meshes yield a higher rate of convergence but they are more complicated to construct.

The result of Theorem 5.6 can be extended to more general forms of upwinding and to other non-equidistant layer-adapted meshes that are designed for convection-diffusion problems. For an excellent survey of such generalizations for problems in one and two dimensions, see Linß (2003).

6. Convection-diffusion problems in two dimensions

In two dimensions, the convection-diffusion equation takes the form

$$Lu(x,y) := -\varepsilon\Delta u(x,y) + \mathbf{a}(x,y).\nabla u(x,y) + b(x,y)u(x,y) = f(x,y) \quad (6.1a)$$

on $\Omega \subset \mathbb{R}^2$, with

$$u(x,y) = g(x,y) \quad \text{on } \partial\Omega, \quad (6.1b)$$

where $0 < \varepsilon \ll 1$, and the functions $\mathbf{a}, b$ and f are assumed to be Hölder continuous on $\bar{\Omega}$, the closure of Ω. We also assume that $b \geq 0$ on $\bar{\Omega}$. Here Ω is any bounded domain in $\mathbb{R}^2$ with a piecewise Lipschitz-continuous boundary $\partial\Omega$ (*e.g.*, a rectangle or a domain with differentiable boundary). Assume that g is continuous except perhaps for a jump discontinuity at a single point. Il'in (1992) gives asymptotic expansions of the solutions to several specific cases of (6.1).

The differential operator L is elliptic, so (6.1) has a solution in $C^2(\Omega)$; see for example Gilbarg and Trudinger (2001). Recall that L satisfies the maximum principle of Lemma 2.1.

Assume that $|\mathbf{a}| \approx 1$, so that convection dominates diffusion. In the problems that we consider, the solution $u(x,y)$ of (6.1) has an asymptotic structure similar to that for one-dimensional problems. That is, analogously to the case $k = 0$ in (3.11), one can write u as the sum of the solution to a first-order PDE, plus layer(s), plus an $\mathcal{O}(\varepsilon)$ term.

To make this more precise, divide the boundary $\partial\Omega$ into 3 parts:

$$\text{inflow boundary } \partial^-\Omega = \{x \in \partial\Omega \colon \mathbf{a}.\mathbf{n} < 0\}, \quad (6.2a)$$

$$\text{outflow boundary } \partial^+\Omega = \{x \in \partial\Omega : \mathbf{a}.\mathbf{n} > 0\}, \quad (6.2b)$$

$$\text{characteristic (tangential) flow boundary } \partial^0\Omega = \{x \in \partial\Omega : \mathbf{a}.\mathbf{n} = 0\}, \quad (6.2c)$$

where $\mathbf{n}$ is the outward-pointing unit normal to $\partial\Omega$. See Figure 6.1.

A typical solution u will have *boundary layers* – narrow regions close to $\partial\Omega$ where $|\nabla u|$ is large – along $\partial^+\Omega$ and $\partial^0\Omega$. As in one-dimensional problems, exceptional Dirichlet boundary conditions g can eliminate these layers; recall the comments following Example 3.1. Also, Neumann boundary conditions on some or all of $\partial^+\Omega$ and $\partial^0\Omega$ mean that layers are no longer

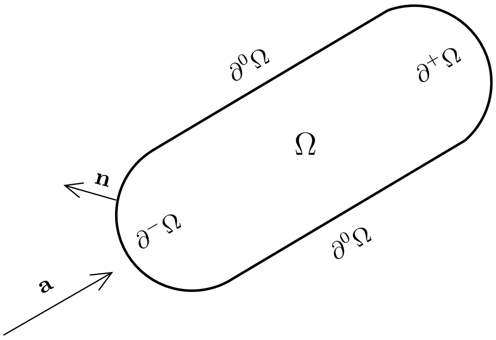

Figure 6.1. Partition of $\partial\Omega$.

visible there (*cf.* Remark 3.6). We shall exploit this property in some numerical examples where Neumann boundary conditions are introduced so that boundary layers will not distract the reader from other visual phenomena.

On most of Ω, u is approximately equal to $u_0(x, y)$, the solution of the *reduced problem*

$$\mathbf{a}(x,y).\nabla u_0(x,y) + b(x,y)u_0(x,y) = f(x,y) \quad \text{on } \Omega, \qquad u_0 = g \quad \text{on } \partial^-\Omega. \tag{6.3}$$

This first-order problem is the two-dimensional analogue of (3.10). Following the standard theory of such PDEs, the *characteristic traces* or *characteristic curves* or *characteristics* of (6.3) are the parametrized curves $(x(t), y(t))$ in Ω defined by

$$x'(t) = a_1(x,y), \quad y'(t) = a_2(x,y), \tag{6.4}$$

with initial data $(x(0), y(0)) = (\hat{x}, \hat{y})$, where $(\hat{x}, \hat{y})$ is any point in $\partial^-\Omega$. Thus one such curve emanates into Ω from each point in $\partial^-\Omega$. The function $u_0(x, y)$ propagates itself along these curves: on each characteristic, (6.3) simplifies to the ordinary differential equation

$$\frac{\mathrm{d}u_0(t)}{\mathrm{d}t} + bu_0 = f \tag{6.5}$$

with initial data $u_0(0) = g(\hat{x}, \hat{y})$, where we have abused the notation by writing u_0 as a function of t along each characteristic. As in fluid dynamics, the direction of propagation $\mathbf{a}$ is often called the *flow*; this explains the terminology of (6.2).

We shall refer to the characteristics of (6.3) as the *subcharacteristics* of (6.1).

Just like in one dimension, boundary layers occur where there is a mismatch between the reduced solution u_0 and the boundary data. This can happen only along $\partial^+\Omega$ and $\partial^0\Omega$. While all layers look much the same when plotted, nevertheless there can be significant analytical differences between them.

Layers along $\partial^+\Omega$ are called *regular* or *exponential boundary layers*. Writing $\vec{n} = (n_1, n_2)$ for the unit outward-pointing normal to $\partial\Omega$, then near $\partial^+\Omega$, exponential layers are essentially multiples of the function

$$\exp[-(\mathbf{a}.\mathbf{n})\, d((x,y), \partial^+\Omega)/\varepsilon],$$

where $d((x,y), \partial^+\Omega)$ denotes the distance from the point (x,y) to the outflow boundary. Thus in cross-section perpendicular to $\partial^+\Omega$ these layers are very similar to the boundary layers that we met in one dimension. Their first-order derivatives in the direction perpendicular to the boundary have magnitude $\mathcal{O}(1/\varepsilon)$, and the width of the layer (*i.e.*, the distance one must travel from the boundary before all first-order derivatives are bounded by some constant C) is $\mathcal{O}(\varepsilon \ln(1/\varepsilon))$.

Layers along $\partial^0\Omega$ are called *parabolic* or *characteristic boundary layers*. In asymptotic expansions of u, these layers can be written as the solution of a parabolic PDE but not as the solution to an ODE; they have a much more complicated structure than exponential boundary layers. Their first-order derivatives in the direction perpendicular to the boundary are $\mathcal{O}(1/\sqrt{\varepsilon})$ – not as large as for exponential layers – but the width of the layer is $\mathcal{O}(\sqrt{\varepsilon}\ln(1/\varepsilon))$, so they are wider than exponential layers.

Example 6.1. In Figure 6.2 we plot the solution $u(x,y)$ to the problem

$$-\varepsilon\Delta u(x,y) + u_x(x,y) = 1 \quad \text{on } \Omega := (0,1)\times(0,1), \qquad u(x,y) \equiv 0 \quad \text{on } \partial\Omega,$$

where $\varepsilon = 0.01$.

The inflow boundary $\partial^-\Omega$ is the side $x = 0$ of $\bar{\Omega}$; the tangential flow boundary comprises the sides $y = 0$ and $y = 1$; the outflow boundary is the remaining side $x = 1$.

From (6.4) each subcharacteristic is parametrized by $x'(t) = 1,\ y'(t) = 0$, so we can take $x = t$ and the subcharacteristics are the lines $y = k$ for arbitrary constant k. Then by (6.5) the reduced problem u_0, written as a function of the parameter t, satisfies $u_0'(t) = 1$, with initial data $u_0(0) = 0$. Hence $u_0(t) = t$, *i.e.*, $u_0(x,y) = x$ for all $(x,y) \in \Omega$.

On most of Ω one therefore has $u(x,y) \approx x$. The side $x = 1$ of $\bar{\Omega}$ is the outflow boundary $\partial^+\Omega$ and an exponential layer appears there. The tangential flow boundaries $y = 0$ and $y = 1$ have characteristic boundary layers that grow in strength as x moves from 0 to 1 because of the increasing discrepancy between u_0 and the boundary condition.

In an asymptotic expansion of u, the leading term describing the layer along $y = 0$ (the layer along $y = 1$ is of course analogous) is

$$v_0\biggl(x, \frac{y}{\sqrt{\varepsilon}}\biggr) = -\sqrt{\frac{2}{\pi}} \int_{s=y/\sqrt{2\varepsilon x}}^{\infty} \mathrm{e}^{-s^2/2}\, u_0\biggl(x - \frac{y^2}{2\varepsilon s^2}, 0\biggr)\, \mathrm{d}s.$$

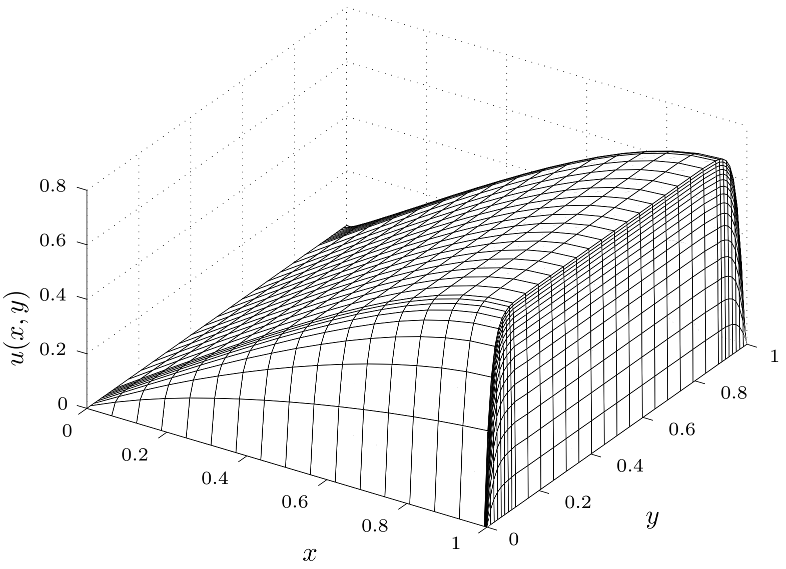

Figure 6.2. Exponential boundary layer with two characteristic boundary layers.

This is much more complicated than for an exponential layer, but at least we can see that when deriving this term the correct choice for the local stretched variable is $(x,\, y/\sqrt{\varepsilon})$.

As well as boundary layers, solutions of convection-diffusion problems in two-dimensional domains can have *interior layers* if there is a discontinuity in the boundary data on $\partial^-\Omega$. This phenomenon has no analogue in one-dimensional problems. From the theory of first-order PDEs, if g has a jump discontinuity at a point $(\hat{x}, \hat{y}) \in \partial^-\Omega$, then u_0 will be discontinuous across the subcharacteristic $\Gamma(\hat{x}, \hat{y})$ that passes through $(\hat{x}, \hat{y})$. Now first-order PDEs preserve Dirichlet boundary data discontinuities but second-order elliptic PDEs smooth out such discontinuities, so the solution $u(x, y)$ of (6.1) will be continuous across $\Gamma(\hat{x}, \hat{y})$. At the same time, u must be close to u_0 once we are a small distance away from $\Gamma(\hat{x}, \hat{y})$. Combining these facts, we deduce that u has an interior layer along the subcharacteristic $\Gamma(\hat{x}, \hat{y})$. Such layers have an asymptotic structure similar to characteristic boundary layers; they are often referred to as *parabolic* or *characteristic interior layers*.

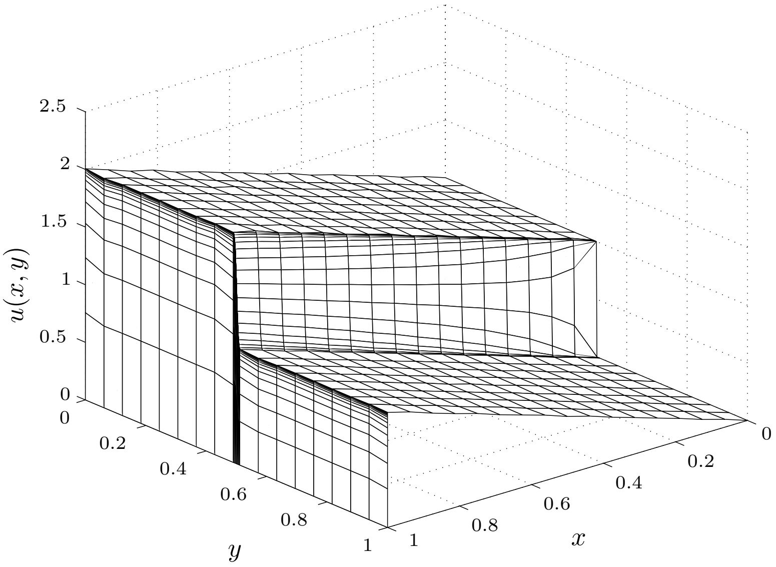

Figure 6.3. Straight interior layer.

Example 6.2. In Figure 6.3 we use the same differential operator as in Example 6.1, with $\varepsilon = 10^{-6}$. A jump discontinuity has been introduced in the inflow boundary data:

$$g(0, y) = \begin{cases} 1 & \text{for } 0 \leq y < 0.5, \\ 0 & \text{for } 0.5 < y \leq 1. \end{cases}$$

Consequently the reduced solution is

$$u_0(x, y) = \begin{cases} 1 + x & \text{for } 0 \leq y < 0.5, \\ x & \text{for } 0.5 < y \leq 1. \end{cases}$$

This yields an interior layer along the subcharacteristic passing through the discontinuity at $(0.5, 0)$, that is, along the line $y = 0.5$. Neumann boundary conditions have been applied on the sides $y = 0$ and $y = 1$ so no layers are visible there, unlike Figure 6.2. A homogeneous Dirichlet boundary condition is still assumed at $x = 1$, and again produces an exponential outflow layer, but this layer is sharper than in Figure 6.2 because ε is much smaller in the present example.

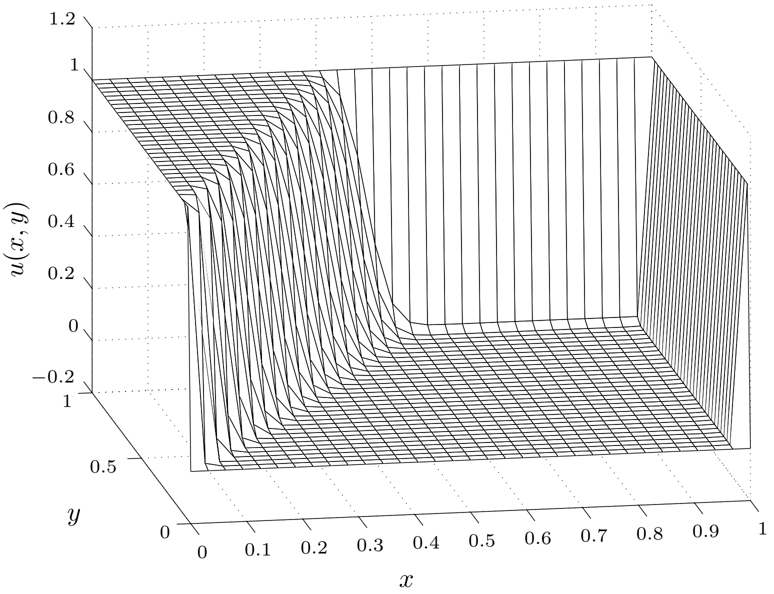

Figure 6.4. Solution of Example 6.3.

Example 6.3. Consider the problem

$$Lu(x,y) := -\varepsilon \Delta u(x,y) + u_x(x,y) + 2u_y(x,y) = 0 \quad \text{on } \Omega := (0,1)\times(0,1),$$

where the boundary condition is $u(x,y) = g(x,y)$ with

$$g(x,y) = \begin{cases} 0 & \text{when } y = 0, \\ 1 & \text{otherwise.} \end{cases}$$

There is no tangential flow boundary. The inflow boundary $\partial^-\Omega$ comprises the sides $x = 0$ and $y = 0$ of $\bar\Omega$. In (6.5) the functions b and f are both zero, so the reduced solution $u_0(x,y)$ is just the initial data on $\partial^-\Omega$ propagated along the subcharacteristics of L without change. These subcharacteristics are the lines $y = 2x + k$ for arbitrary constant k.

The solution $u(x,y)$ is as usual very close to u_0 away from layers. The outflow boundary $\partial^+\Omega$ comprises the sides $x = 1$ and $y = 1$ of $\bar\Omega$. Along the portion $0 \le x \le 1/2$ of the side $y = 1$ there is no layer because $u_0 = g$ there. There are exponential boundary layers along the rest of $\partial^+\Omega$. An interior layer emanates across Ω from the discontinuity in g at the point $(0,0)$, *i.e.*, along the line $y = 2x$. See Figure 6.4, where $\varepsilon = 0.001$. The slightly diffuse nature of the interior layer in this figure is an artifact of the method used to compute u; see Remark 10.8.

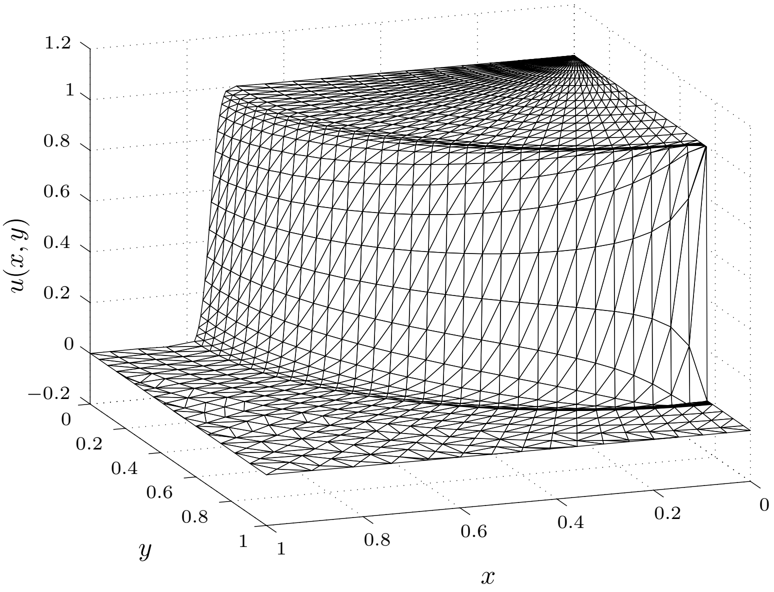

Figure 6.5. Curved interior layer.

Example 6.4. Finally we consider a problem with a curved interior layer:

$$\begin{aligned} -\varepsilon\Delta u + \mathbf{a}.\nabla u &= 0 \quad \text{on } \Omega := (0,1)\times(0,1),\\ \nabla u.\mathbf{n} &= 0 \quad \text{on } \{(x,0) : 0 \le x \le 1\} =: \partial\Omega_1,\\ u &= g \quad \text{on } \partial\Omega \setminus \partial\Omega_1, \end{aligned}$$

where $\mathbf{n}$ is the outward-pointing unit normal to $\partial\Omega$ and

$$\mathbf{a}(x,y) = (\sin\theta, -\cos\theta)$$

with θ the argument of the point (x,y) in polar coordinates. The function g is defined by

$$g = \begin{cases} 1 & \text{for } x = 0,\ 0 \le y \le 0.75,\\ 0 & \text{for } x = 0,\ 0.75 < y \le 1,\\ 0 & \text{for } 0 \le x \le 1,\ y = 1,\\ 0 & \text{for } x = 1,\ 0 \le y \le 1. \end{cases}$$

The Neumann condition on $\partial\Omega_1$ ensures that no outflow boundary layer appears there.

The subcharacteristics are quarter-circles centred at the origin. Since $b = f = 0$, the reduced solution merely propagates the inflow boundary values along these quarter-circles without changing their values. A computed solution to this problem is shown in Figure 6.5 with $\varepsilon = 0.0001$.

7. *A priori* estimates

In this section various *a priori* results for the solution of (6.1) are presented.

Many *a priori* analyses in the literature assume the condition

$$\mathbf{a}(x,y) = \big(a_1(x,y), a_2(x,y)\big) > (\alpha_1, \alpha_2) > (0,0) \quad \text{on } \Omega, \tag{7.1}$$

which in the case where Ω is the unit square ensures that no characteristic boundary layers are present.

Lemma 7.1. Assume that (7.1) holds true. Then the following results hold.

(i) There exists a constant C, which depends on the domain Ω, such that

$$\|u\|_{L^\infty(\Omega)} \le \|g\|_{L^\infty(\partial\Omega)} + \frac{C\|f\|_{L^\infty(\Omega)}}{\max\{\alpha_1, \alpha_2\}}\,. \tag{7.2}$$

If Ω is the unit square, then $C = 1$.

(ii) For each $\delta > 0$, define $\Omega_\delta = \{x \in \Omega : \ \operatorname{dist}(x, \partial^+\Omega \cup \partial^0\Omega) > \delta\}$. Let $g \in C(\partial\Omega)$. Then there exists a constant $C = C(\delta)$ such that $|u(x,y) - u_0(x,y)| \le C\varepsilon$ for all $(x,y) \in \Omega_\delta$.

Proof. The proof of (i) is similar to the proof of Lemma 3.3.

The hypothesis of (ii) ensures that there are no interior layers. The proof can be found in Goering, Felgenhauer, Lube, Roos and Tobiska (1983). □

Let $\|\cdot\|_k$ and $|\cdot|_k$ denote the usual norm and seminorm on the Sobolev space $H^k(\Omega)$ for all nonnegative integers k. In particular $\|\cdot\|_0 = \|\cdot\|_{L^2(\Omega)}$.

The presence of layers in u means that one does not have $\|u\|_k \le C$ for any $k \ge 1$. Even in one dimension, the H^k norm of the function $\mathrm{e}^{-(1-x)/\varepsilon}$ is easily seen to be $\mathcal{O}\big(\varepsilon^{-k+1/2}\big)$, and exponential layers in two-dimensional problems have a similar magnitude. This observation motivates the following definition of a weighted energy norm that is commonly used in finite element analyses of convection-diffusion problems: for all $w \in H^1(\Omega)$, set

$$\|w\|_{1,\varepsilon} = \sqrt{\varepsilon|w|_1^2 + \|w\|_0^2}.$$

Then typically $\|u\|_{1,\varepsilon} \le C$, uniformly in ε.

Lemma 7.2. Let u be the solution of (6.1). Assume that $b - (\operatorname{div}\mathbf{a})/2 \ge C_5 > 0$ on $\bar\Omega$ for some constant C_5. Assume also that Ω is convex or has smooth boundary. Then there exists a constant C such that

$$\varepsilon^{3/2}|u|_2 + \varepsilon^{1/2}|u|_1 + \|u\|_0 \le \varepsilon^{3/2}|u|_2 + \sqrt{2}\,\|u\|_{1,\varepsilon} \le C.$$

Proof. Let G be the solution of the problem $\Delta G = 0$ on Ω, $G = g$ on $\partial\Omega$. Then the hypotheses on the domain Ω ensure that $\|G\|_2 \le C$ by a classical inequality (see, *e.g.*, Gilbarg and Trudinger (2001)). Subtract G

from u to reduce the problem to the case of homogeneous Dirichlet boundary conditions. Now use a standard energy norm argument: multiply $Lu = f$ by u then integrate by parts, obtaining

$$\varepsilon|u|_1^2 + \int_\Omega \big(b - \tfrac{1}{2}\operatorname{div}\mathbf{a}\big)u^2 = \int_\Omega fu \leq \|f\|_0\|u\|_0 \leq \frac{1}{2C_5}\|f\|_0^2 + \frac{C_5}{2}\|u\|_0^2$$

and $\|u\|_{1,\varepsilon} \leq C$ follows.

The PDE (6.1) and this inequality now yield

$$\varepsilon\|\Delta u\|_0 \leq C(|u|_1 + \|u\|_0 + \|f\|_0) \leq C(\varepsilon^{-1/2} + 1) \leq C\varepsilon^{-1/2},$$

so $\varepsilon^{3/2}\|\Delta u\|_0 \leq C$. But the classical inequality $|u|_2 \leq C(\|\Delta u\|_0 + \|u\|_0)$ holds true (Gilbarg and Trudinger 2001), and we get $\varepsilon^{3/2}|u|_2 \leq C$. □

Remark 7.3. Analogously to Remark 3.2, if (7.1) holds true then one can assume without loss of generality that $b - (\operatorname{div}\mathbf{a})/2 \geq C_5 > 0$ on $\bar{\Omega}$ also holds true.

We now give some idea of the behaviour of derivatives of the solution u of (6.1) near exponential boundary layers and corners. Suppose that Ω is the unit square and the differential operator is as in Example 6.3, so that (7.1) holds true. Then the sides $x = 1$ and $y = 1$ form the outflow boundary $\partial^+\Omega$. Assuming that no extra complications such as interior layers are present, near $x = 1$ one expects the solution u to satisfy the bound

$$\left|\frac{\partial^{i+j}u(x,y)}{\partial^i x\partial^j y}\right| \leq C\big(1 + \varepsilon^{-i}\mathrm{e}^{-(1-x)/\varepsilon}\big), \tag{7.3}$$

while near $y = 1$ one expects

$$\left|\frac{\partial^{i+j}u(x,y)}{\partial^i x\partial^j y}\right| \leq C\big(1 + \varepsilon^{-j}\mathrm{e}^{-2(1-y)/\varepsilon}\big). \tag{7.4}$$

Close to the corner (1,1) there will be an *outflow corner layer*, which is like a product of exponential boundary layers, and satisfies the bound

$$\left|\frac{\partial^{i+j}u(x,y)}{\partial^i x\partial^j y}\right| \leq C\big(1 + \varepsilon^{-(i+j)}\mathrm{e}^{-(1-x)/\varepsilon}\mathrm{e}^{-2(1-y)/\varepsilon}\big). \tag{7.5}$$

Despite the extra negative powers of ε in (7.5), corner layers of this type rarely cause difficulty for numerical methods because they decay so rapidly as one moves away from the corner.

A rigorous proof of bounds such as (7.3)–(7.5) is a delicate and lengthy matter. Such a proof is given by Linß and Stynes (2001*a*) for problems like the one under discussion, but with the extra assumptions that the Dirichlet boundary condition $g(x, y)$ is a continuous function and that a sufficient number of *compatibility conditions* hold true at the corners of $\bar{\Omega}$.

Compatibility conditions are relationships between the data of the problem and the differential operator that ensure that derivatives of u up to a desired order are continuous on $\bar{\Omega}$. They arise only at corners and are not caused by the singularly perturbed nature of the problem. Grisvard (1985) provides a general exposition of compatibility conditions for elliptic operators on polygonal domains and Han and Kellogg (1990) write down the precise form that they take when applied to convection-diffusion problems posed on the unit square.

If compatibility conditions beyond a certain order are not satisfied at a corner of a domain, then certain derivatives of that order and higher orders must blow up as one approaches this corner. Kellogg and Stynes (2005) derive bounds on the derivatives of the solution of a generalization of Example 6.1 in terms of the number of compatibility conditions that are satisfied at each corner. Near $x = 1$, but away from corners, we have (7.3). Near the characteristic boundary $y = 1$, we find that

$$\left| \frac{\partial^{i+j} u(x,y)}{\partial^i x \partial^j y} \right| \leq C\big[1 + (\sqrt{\varepsilon}\,)^{-j} \mathrm{e}^{-2(1-y)/\sqrt{\varepsilon}}\,\big]$$

provided we stay away from corners. Near the corners, singularities in the derivatives begin to appear; we do not give the details here.

The data of Example 6.3 are not fully compatible at the corner (1,1) with the differential operator L. This incompatibility will cause singularities in the derivatives of u at (1,1). The interaction between these singularities and the exponential and corner layers is not yet fully understood. That is, we are currently unable to write down reliable sharp pointwise bounds on the derivatives of u near the point (1,1), but one expects that sharp bounds are at least as bad as (7.5) and will blow up as (x, y) approaches $(1, 1)$.

It is in general difficult to derive bounds on derivatives of solutions of convection-diffusion problems inside characteristic boundary and interior layers. Although such bounds are of great interest to numerical analysts, few rigorous results appear in the literature. Kellogg and Stynes (2005) provide pointwise bounds for characteristic boundary layers. In a subsequent paper (Kellogg and Stynes 2004) they consider a convection-diffusion problem in a half-plane with a discontinuity in an arbitrary specified derivative of the boundary data and derive pointwise bounds on derivatives of the solution, including the behaviour along the interior layer emanating from the point of discontinuity.

Dörfler (1999) gives bounds on u and its derivatives in various norms (both isotropic and anisotropic) and for a variety of convection-diffusion problems on bounded domains. Shishkin (1990) contains pointwise bounds on derivatives of u for many variants of (6.1) but the arguments are presented in a very concise style and it is difficult to ascertain the precise assumptions made.

8. General comments on numerical methods

Numerical methods (such as central differencing on equidistant meshes) that contain no mechanism for stabilizing solutions in exponential layers will usually have wild oscillations in their computed solutions on much of Ω, like in Section 4. As we shall see, this problem can be handled by modifying the approximation of the convective terms (*e.g.*, using some form of finite difference upwinding or special choices of finite element trial and test spaces) or by modifying the mesh (*e.g.*, a two-dimensional Shishkin mesh). When this is done correctly, one can compute accurate solutions inside these layers.

Characteristic layers, on the other hand, differ in both respects:

- if the method has no stabilizing mechanism specifically designed to address characteristic layers, then the layer will induce small oscillations in the computed solution, but these oscillations usually appear only inside and near the characteristic layer, so the solution can still be computed accurately on the rest of Ω;
- it is often difficult – at least in the case of interior layers – to compute accurate solutions inside characteristic layers.

Thus one could use some form of upwinding (*i.e.*, some discrete approximation of $\mathbf{a}.\nabla u$ that is skewed away from the outflow boundary) to stabilize the method for exponential layers, combined with some heuristic mesh refinement near characteristic layers. Whether or not the mesh refinement yields an accurate solution inside the characteristic layers, nevertheless the solution elsewhere will be accurate.

The following pair of examples are related to our observation that one can to a certain extent neglect characteristic layers but not exponential layers.

Consider again Example 6.3 but with $g(x, y) \equiv 1$. Then the solution $u(x, y)$ has exponential boundary layers along $x = 1$ and $y = 1$. The reduced solution $u_0(x, y)$ will of course ignore these layers, and we find that $\|u - u_0\|_{1,\varepsilon} = \mathcal{O}(1)$.

On the other hand the solution u of Example 6.1 has two characteristic layers and one exponential layer. Schieweck (1986) proves that if one sets $v(x, y) = u_0(x, y) - u_0(1, y)\mathrm{e}^{-(1-x)/\varepsilon}$ (this is the reduced solution plus an appropriate exponential layer term, so it ignores only the parabolic layers), then $\|u - v\|_{1,\varepsilon} \le C\varepsilon^{1/4}$.

Nevertheless, in some applications characteristic layers cannot be neglected.

9. Finite difference methods in two dimensions

Assume that Ω is the unit square and the mesh $\{(x_i, y_j)\}$ is rectangular and equidistant in each coordinate direction: $x_i = ih$ and $y_j = jk$ for

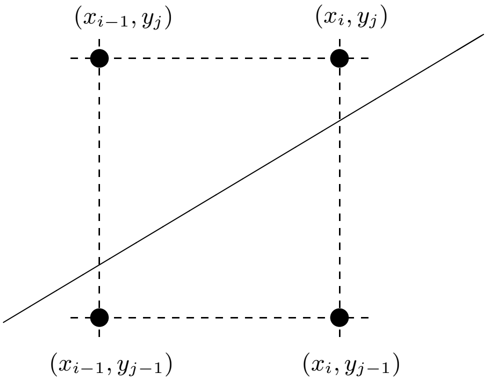

Figure 9.1. Mesh points and line indicating nearby interior layer.

$i = 0, \dots, N$ and $j = 0, \dots, M$ with $h := 1/N$ and $k := 1/M$. We use a standard approximation of the second-order derivatives:

$$\begin{aligned} u_{xx}(x_i, y_j) &\approx \frac{u^N_{i+1,j} - 2u^N_{ij} + u^N_{i-1,j}}{h^2}, \\ u_{yy}(x_i, y_j) &\approx \frac{u^N_{i,j+1} - 2u^N_{ij} + u^N_{i,j-1}}{k^2}, \end{aligned} \tag{9.1}$$

where u^N_{ij} is the computed solution at each mesh point (x_i, y_j).

As for one-dimensional problems, approximating the first-order derivatives in (6.1) by central differences

$$u_x(x_i, y_j) \approx \frac{u^N_{i+1,j} - u^N_{i-1,j}}{2h} \quad \text{and} \quad u_y(x_i, y_j) \approx \frac{u^N_{i,j+1} - u^N_{i,j-1}}{2k}$$

leads to an unstable method. Instead one can use simple upwinding,

$$u_x(x_i, y_j) \approx \frac{u^N_{i,j} - u^N_{i-1,j}}{h} \quad \text{and} \quad u_y(x_i, y_j) \approx \frac{u^N_{i,j} - u^N_{i,j-1}}{k},$$

and this yields an M-matrix. Combining this with (9.1) and the approximation $u(x_i, y_j) \approx u^N_{ij}$ for the zero-order term in (6.1), the resulting method is stable but we expect from our experience with ODEs that it will smear exponential boundary layers.

In fact, one can foresee heuristically that this method will also smear interior layers. In Figure 9.1, the value of $u(x_i, y_j)$ depends strongly on the u values along the upstream portion of the subcharacteristic that passes through (x_i, y_j) – this is a line through (x_i, y_j) parallel to the line drawn – but simple upwinding makes $u(x_i, y_j)$ depend on $u(x_i, y_{j-1})$, which introduces inaccuracies because the value of $u(x_i, y_j)$ has little to do with the values of u on the other side of the interior layer indicated by the line in Figure 9.1.

A difference scheme on a family of arbitrary rectangular meshes of $(N+1)^2$ points (we take the same number of mesh points in each coordinate direction for simplicity) is said to be *robust* or *uniformly convergent* (*with respect to* ε) *of order* $\beta > 0$ *in the discrete* L^∞ *norm* if its solution $\{u_{ij}^N\}$ satisfies

$$|u_{ij} - u_{ij}^N| \le CN^{-\beta} \quad \text{for } i, j = 0, \dots, N$$

and all sufficiently small H, independently of ε. Here we take $N+1$ mesh points in each coordinate direction for simplicity, H is the mesh diameter, β is some positive constant that is independent of the mesh and of ε, and we write u_{ij} instead of $u(x_i, y_j)$ (we shall do likewise for all other functions in $C(\bar{\Omega})$).

For uniform convergence on an equidistant mesh, an analogue of Theorem 4.8 shows that once again the coefficients in the scheme must have a certain exponential character (Roos *et al.* 1996, p. 194). One can define a five-point scheme that is a two-dimensional analogue of the Il'in scheme of Example 4.9. When the data of (6.1) are smooth and some compatibility conditions are satisfied at the corners of Ω, this scheme can be proved to achieve uniform convergence of order β, where β is almost 1/2, in the discrete L^∞ norm (Roos *et al.* 1996, p. 195). Nevertheless this scheme, which is a form of upwinding, smears interior layers quite badly and is rarely used.

Continuing in the footsteps of our earlier sections, we now consider a two-dimensional Shishkin mesh for a problem on the unit square that satisfies (7.1) and consequently has exponential boundary layers along $x = 1$ and $y = 1$. Let N, an even integer, be the number of mesh intervals in each coordinate direction. Define the transition points on the x- and y-axes to be $1 - \lambda_x$ and $1 - \lambda_y$ respectively, where $\lambda_x = (2\varepsilon/\alpha_1)\ln N$ and $\lambda_y = (2\varepsilon/\alpha_2)\ln N$. The fine and coarse mesh regions on the coordinate axes each contain $N/2$ mesh intervals. See Figure 9.2 for the mesh with $N = 8$.

One can define simple upwinding on non-equidistant meshes similarly to the formulas of Section 5. Writing u_{ij}^N for the solution computed using this method on the Shishkin mesh, then under compatibility assumptions from Linß and Stynes (2001*a*) guaranteeing that the solution u can be decomposed as a sum of the reduced solution, an exponential layer at $x = 1$, an exponential layer at $y = 1$ and a corner layer at (1,1), where these layers satisfy bounds similar to (7.3)–(7.5), an analysis similar to that of Section 5 shows that

$$|u_{ij} - u_{ij}^N| \le CN^{-1}\ln N \quad \text{for all } i, j.$$

That is, we get almost first-order uniform convergence in the discrete L^∞ norm.

If we modify this scheme by using central differencing instead of upwinding wherever the Shishkin mesh is fine in the relevant coordinate direction, then the M-matrix property is retained and a variant of the upwind analysis

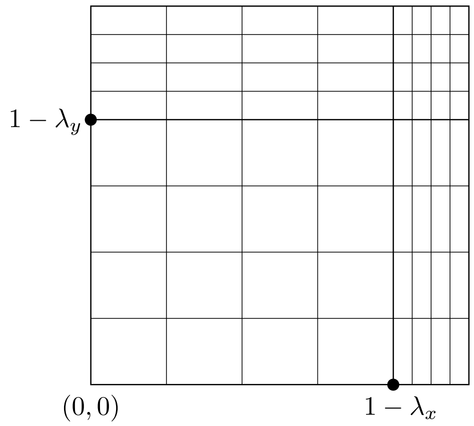

Figure 9.2. Shishkin mesh with $N = 8$.

yields (Linß and Stynes 1999) the improved bound

$$|u_{ij} - u_{ij}^{N,\text{hybrid}}| \leq CN^{-1} \quad \text{for all } i, j,$$

where $u_{ij}^{N,\text{hybrid}}$ is the solution computed by this hybrid scheme.

Kopteva (2003) shows, under some extra compatibility assumptions at the corners, that one iteration of Richardson extrapolation applied to the simple upwind solution u_{ij}^N on the Shishkin mesh yields a solution v_{ij}^N for which

$$|u_{ij} - v_{ij}^N| \leq CN^{-2} \ln^2 N \quad \text{for all } i, j.$$

Approximation of the first-order derivatives of u is also discussed in this paper.

Remark 9.1. (Shishkin's obstacle theorem) The above convergence results are all proved under hypotheses that exclude characteristic layers. The difficulty of accurately approximating characteristic boundary layers is underlined by a remarkable result of Shishkin (1989): suppose that one has a problem whose solution has a characteristic boundary layer. Suppose also that one applies any difference scheme on an equidistant mesh whose coefficients are drawn from a fixed class of functions (*e.g.*, the Il'in scheme, whose coefficients are all exponentials and polynomials; the point is that one is forbidden to vary the difference scheme by choosing the type of coefficients to correspond exactly to the precise nature of each new set of boundary data). Then *this scheme cannot yield uniform convergence of any positive order in the discrete L^∞ norm inside the characteristic boundary layer for all smooth and compatible boundary data g.* The essential reason for this negative result is that at each point (x, y) near $\partial^0\Omega$ a characteristic boundary layer depends on all the data along that connected component of $\partial^0\Omega$; this is

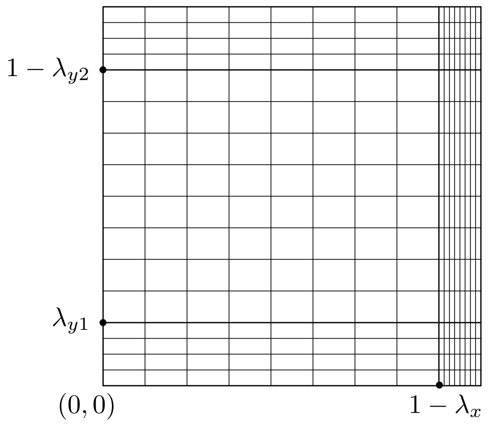

Figure 9.3. Shishkin mesh for Example 6.1 with $N = 16$.

quite unlike an exponential boundary layer, whose behaviour at (x, y) near $\partial^+\Omega$ depends only on the difference between the reduced solution u_0 and the boundary data at the nearest boundary point – a much simpler situation. A consequence of Shishkin's result is that special schemes on equidistant meshes are unsatisfactory inside characteristic boundary layers; instead, we must use meshes that are adapted either *a priori* or *a posteriori*.

For a problem on the unit square (such as Example 6.1) that has an exponential boundary layer at $x = 1$ and characteristic boundary layers at $y = 0$ and $y = 1$, a suitable Shishkin mesh is constructed as follows: use an x-axis transition point exactly as in Figure 9.2. Place y-axis transition points at λ_{y1} and $1-\lambda_{y2}$ where each λ_{yk} is $\mathcal{O}\big(\varepsilon^{1/2}\ln N\big)$, then use $N/4$ equidistant mesh intervals in each of $[0, \lambda_{y1}]$ and $[1-\lambda_{y2}, 1]$ and $N/2$ equidistant mesh intervals in $[\lambda_{y1}, 1-\lambda_{y2}]$. See Figure 9.3. Then for simple upwinding, Shishkin (1990) shows that under certain fairly strong hypotheses on the smoothness and compatibility of the data of the problem, simple upwinding yields

$$|u_{ij} - u_{ij}^N| \le CN^{-1}\ln N \quad \text{for all } i, j,$$

where u_{ij}^N is the computed solution.

A large collection of numerical computations on Shishkin meshes for various problems can be found in Farrell *et al.* (2000). While the assembly of Shishkin meshes for boundary layers along straight portions of $\partial\Omega$ is straightforward once the asymptotic nature of the layer has been ascertained, for general (curved) interior layers there are practical difficulties in the construction of these meshes and the only examples in the literature are for problems like Example 6.4, where the layer has a certain symmetry. Nevertheless one can achieve satisfactory visual results with heuristic approximations of Shishkin meshes in such situations; see Madden and Stynes (1997). Shishkin's second doctoral thesis (Shishkin 1990) contains a wealth of theoretical results for finite differences applied on piecewise uniform meshes for many convection-diffusion problems. It is at present being translated into English, but is written in an extremely condensed style.

Remark 9.2. (Defect correction method) This technique seeks to generate a useful higher-order scheme by combining a stable low-order scheme with a higher-order but unstable scheme.

Consider an arbitrary rectangular mesh. Compute an initial approximation $\hat{u}^N$ using simple upwinding: $L^N_{up}\hat{u}^N = f^N$. Obtain the 'defect' σ^N by means of the formally higher-order central difference scheme L^N_c: set $\sigma^N = f^N - L^N_c\hat{u}^N$. Compute the defect correction δ^N by solving $L^N_{up}\delta^N = \sigma^N$. Form the final solution $u^N := \hat{u}^N + \delta^N$.

This method avoids instability by solving only discrete systems that involve the upwind operator L^N_{up}, yet aims to attain the higher-order convergence associated with the operator L^N_c. The idea can be placed in a more general setting and has been applied to many problems unrelated to convection-diffusion (Bohmer and Rannacher 1984). For convection-diffusion the only satisfactory analysis of the method, which shows that it does indeed achieve second-order convergence on a Shishkin mesh, is in Fröhner, Linß and Roos (2001) where a one-dimensional problem is treated. Defect correction is related to Richardson extrapolation, and to obtain a rigorous proof of its validity in two dimensions on a Shishkin mesh like that of Figure 9.2 would require, *e.g.*, some extension of the delicate analysis in Kopteva (2003). Nevertheless numerical results for the method are encouraging (see Remark 10.2).

Finally, we point out that when one no longer assumes hypotheses such as $\mathbf{a}(\cdot,\cdot) > (0,0)$, then although simple upwinding remains stable (*i.e.*, the computed solution is free of non-physical oscillations), it can give dangerously misleading results. Brandt and Yavneh (1991) give an example of linearized recirculating flow in an annulus where the subcharacteristics are circles and, except near the boundary of the domain, the solution computed by a version of simple upwinding is $\mathcal{O}(1)$ distant from the true solution!

10. Finite element methods

If one attempts to solve a convection-diffusion problem by means of a standard Galerkin finite element method with linear or bilinear elements on an equidistant mesh, then a typical computed solution will display large oscillations. This is analogous to our experience in Section 9 with central differencing. Thus some mechanism is needed to stabilize a FEM: a special choice of trial or test functions, or a special mesh, or a modification of the standard bilinear form, or a combination of these devices. In the subsections that follow we discuss each in turn.

Throughout this section we shall assume (*cf.* Remark 7.3) that

$$b(x,y) - \frac{\operatorname{div} \mathbf{a}(x,y)}{2} \geq C_5 > 0 \quad \text{on } \bar{\Omega} \text{ for some constant } C_5. \tag{10.1}$$

For convenience also assume that $u \equiv 0$ on $\partial\Omega$.

10.1. L^*-splines

We did not discuss finite element methods for one-dimensional convection-diffusion problems such as (3.12) since often they are merely an alternative way of generating finite difference schemes.

For example, one can generate the Il'in scheme of Example 4.9 by a finite element method on the same equidistant mesh. It is a Petrov–Galerkin FEM, that is, the trial space S^N and test space T^N are not identical, unlike standard (Bubnov–)Galerkin methods. One takes S^N to be the standard space of piecewise linear functions on the mesh $x_i = i/N$, for $i = 0, 1, \ldots, N$, that vanish at $x = 0, 1$ to satisfy the boundary conditions in (3.12). Recall that the differential equation in (3.12) is $Lu(x) := -\varepsilon u''(x) + a(x)u'(x) + b(x)u(x) = f(x)$, with $a(\cdot) > 0$. Define the test space T^N to be the space of approximate *L^*-splines* spanned by $\{\psi_i\}_{i=1}^{N-1}$, where

$$\bar{L}^*(\psi_i)(x) := -\varepsilon\psi_i''(x) - \bar{a}(x)\psi_i'(x) + \bar{b}(x)\psi_i(x) = 0 \tag{10.2}$$
$$\text{on each subinterval } (x_{j-1}, x_j)$$

and $\psi_i(x_j) = \delta_{ij}$, the discrete Kronecker delta. Here $\bar{a}$ is some approximation of $a(x)$ that is constant on each mesh subinterval, and b and f are approximated by $\bar{b}$ and $\bar{f}$ in a similar way. As usual in FEMs, the computed solution $u^N(x) \in S^N$ is generated by a weak form of the differential equation:

$$\int_0^1 \left[\varepsilon(u^N)'(x)\psi_i'(x) + \bar{a}(x)(u^N)'(x)\psi_i(x) + \bar{b}u^N(x)\psi_i(x)\right] \mathrm{d}x$$
$$= \int_0^1 \bar{f}(x)\psi_i(x)\,\mathrm{d}x \quad \text{for } i = 1, \ldots, N-1.$$

If one defines $\bar{a}$ by the quadrature rule

$$\int_0^1 \bar{a}(x)(u^N)'(x)\psi_i(x)\,\mathrm{d}x = a_i \int_0^1 (u^N)'(x)\psi_i(x)\,\mathrm{d}x,$$

with similar definitions for $\bar{b}$ and $\bar{f}$, then one obtains the Il'in scheme. The alternative choice

$$\bar{a}\Big|_{(x_{j-i}, x_j)} = \frac{a_{j-1} + a_j}{2} \quad \text{for each } j$$

(with similar definitions for $\bar{b}$ and $\bar{f}$) yields the El Mistikawy–Werle scheme of Section 4.

Both of these are successful schemes, and the only special construction we made when generating them in a FEM context was to use L^*-splines. Why do L^*-splines make such good test functions?

The explanation is to be found by considering Green's functions for the differential operator L. For each mesh point $x_i \in (0,1)$ let $G(\cdot, x_i)$ denote the Green's function associated with that point, that is,

$$L^*G(\xi, x_i) = \delta(\xi - x_i) \quad \text{for } 0 < \xi < 1, \qquad G(0, x_i) = G(1, x_i) = 0,$$

where we define

$$L^*G(\xi, x_i) := -\varepsilon G_{\xi\xi}(\xi, x_i) - \big(a(\xi)G(\xi, x_i)\big)_\xi + b(\xi)G(\xi, x_i).$$

Then

$$\begin{aligned} u_i &= \int_0^1 f(\xi)G(\xi, x_i)\,\mathrm{d}\xi \\ &= \int_0^1 (Lu)(\xi)G(\xi, x_i)\,\mathrm{d}\xi \\ &= \int_0^1 \big[\varepsilon u'(\xi))G_\xi(\xi, x_i) + a(x)u'(\xi)G(\xi, x_i) + bu(\xi)G(\xi, x_i)\big]\,\mathrm{d}\xi. \end{aligned}$$

Note both the resemblance between this identity and the weak form of the differential equation that was used above to generate the FEM and the similarity between the definitions of G and ψ_i. The key idea of this FEM was to choose the ψ_i in such a way that the test space T^N was capable of producing a decent approximation of the Green's function, and this property can be exploited in the analysis of the method.

The Green's function exhibits layers at $\xi = 0$ and at $\xi = x_i$; on each subinterval $[0, x_i]$ and $[x_i, 1]$ these layers occur at the left-hand end, unlike the layer in $u(x)$ at $x = 1$, because of the negative coefficient $-a(\xi)$ in the convective term appearing in the definition of L^*. See Figure 10.1.

Remark 10.1. When piecewise linears or bilinears are used as the trial space for convection-diffusion problems in one or two dimensions, useful numerical methods on general meshes are based on some test space that is constructed to approximate the Green's function of the continuous operator. This Green's function is skewed away from the outflow boundary; see Morton (1996) for a discussion of its properties in two dimensions.

Alternatively, one can shift the work from the test space to the trial space by using trial functions ϕ that are approximate L-splines (*i.e.*, satisfy some approximate version of $L\phi = 0$), together with some standard space of test functions such as piecewise linears. The relationship between this dual

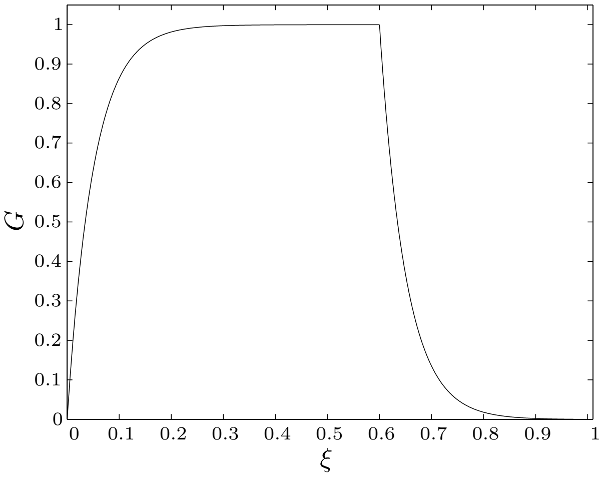

Figure 10.1. Green's function $G(\xi, x_i)$ with $a \equiv 1$, $b \equiv 0$, $x_i = 0.6$ and $\varepsilon = 0.05$.

approach and the use of L^*-spline test functions is discussed at length in Roos *et al.* (1996, §I.2.2.3).

Some authors have generalized the L^*-splines of (10.2) to two dimensions by taking their tensor product on rectangular grids, but this method is applicable only on domains whose boundary comprises straight-line segments each of which is parallel to one of the coordinate axes, and so negates one of the main advantages of finite element methods over finite differences. Consequently we do not discuss this approach here but refer the reader to Roos *et al.* (1996, §II.3.4).

A more useful generalization that is genuinely two-dimensional is found in Sacco, Gatti and Gotusso (1999): to solve (6.1) on an arbitrary triangular mesh one uses a trial space with local basis

$$1, \quad \mathrm{e}^{(\bar{a}_1 x + \bar{a}_2 y)/\varepsilon}, \quad \bar{a}_1 y - \bar{a}_2 x,$$

where $(\bar{a}_1, \bar{a}_2)$ is a piecewise-constant approximation of $\mathbf{a} = (a_1, a_2)$. Here the functions 1 and $\mathrm{e}^{\bar{a}_1 x + \bar{a}_2 y}$ come from the functions that appear in approximate L-splines for the corresponding one-dimensional problem (3.12), but the third function $\bar{a}_1 y - \bar{a}_2 x$ is new. Observe that all three functions lie in the null space of the operator $-\varepsilon\Delta(\cdot) + \bar{a}_1(\cdot)_x + \bar{a}_2(\cdot)_y$, *i.e.*, they are approximate L-splines. Piecewise linears are used in the test space. It is shown in Sacco and Stynes (1998) that this method is essentially equivalent to the unusual exponentially upwinded scheme used in PLTMG.

10.2. Shishkin meshes

FEMs can of course be implemented on Shishkin meshes like those of Figures 9.2 and 9.3 (the mesh rectangles can be bisected into triangles to permit the use of, *e.g.*, a piecewise linear FEM). Note that some mesh rectangles have a high aspect ratio, *i.e.*, their length greatly exceeds their width. To analyse such methods, the highly anisotropic nature of the mesh necessitates the use of sharp anisotropic interpolation estimates like those of Apel and Dobrowolski (1992) and Apel (1999), which we now describe.

Suppose that each element τ (triangle or rectangle) of the mesh is contained in a rectangle with side lengths (h_x, h_y) and contains a rectangle with side lengths (Ch_x, Ch_y) for some fixed constant $C > 0$. In the case of triangles, assume also a maximum angle condition: the interior angles are bounded away from π. (A triangular Shishkin mesh satisfies this maximum angle condition.)

Let $v \in H^2(\tau)$. Let v^I denote the nodal interpolant (linear or bilinear) of v. Write $\|\cdot\|_{0,\tau}$ for the norm in $L^2(\tau)$. Then

$$\begin{aligned}\|v - v^I\|_{0,\tau}^2 &\leq C \sum_{|\alpha|=2} h^{2\alpha} \|D^\alpha v\|_{0,\tau}^2,\\ \|\partial_x(v - v^I)\|_{0,\tau}^2 &\leq C \sum_{|\alpha|=1} h^{2\alpha} \|D^\alpha \partial_x v\|_{0,\tau}^2,\\ \|\partial_y(v - v^I)\|_{0,\tau}^2 &\leq C \sum_{|\alpha|=1} h^{2\alpha} \|D^\alpha \partial_y v\|_{0,\tau}^2.\end{aligned}$$

Here α is the multi-index (α_1, α_2), $|\alpha| = \alpha_1 + \alpha_2$, $h^\alpha = h_x^{\alpha_1} h_y^{\alpha_2}$, and

$$D^\alpha = \frac{\partial^{\alpha_1}}{\partial x^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial y^{\alpha_2}}.$$

Bounds of this type are useful on Shishkin meshes because of the very small mesh width in precisely the coordinate direction whose derivative is large, and because no term v_{yy} appears in the bound on $\|\partial_x(v - v^I)\|_{0,\tau}$. Standard isotropic interpolation error estimates use only the diameter of the element and thereby lose the benefit of the Shishkin mesh in the analysis of interpolation error.

With these estimates, Dobrowolski and Roos (1997) show that if Ω is the unit square and the solution u of (6.1) can be written as the sum of a reduced solution and exponential boundary and corner layers, then for piecewise linear or bilinear interpolation on a Shishkin mesh,

$$\|u - u^I\|_{L^\infty(\Omega)} \leq CN^{-2} \ln^2 N \quad \text{and} \quad \|u - u^I\|_0 \leq CN^{-2} + C\sqrt{\varepsilon} N^{-2} \ln^2 N$$

so

$$\|u - u^I\|_0 \leq CN^{-2} \text{ when } \sqrt{\varepsilon} \leq C \ln^{-2} N, \quad \text{and} \quad \|u - u^I\|_{1,\varepsilon} \leq CN^{-1} \ln N.$$

These bounds give us some idea of what convergence rates one can hope for when devising FEMs for convection-diffusion problems on Shishkin meshes.

Define the bilinear form

$$B(v,w) = (\varepsilon\nabla v, \nabla w) + (\mathbf{a}.\nabla v, w) + (bv, w) \quad \text{for all } v, w \in H^1(\Omega), \tag{10.3}$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product. Then (10.1) implies that

$$B(v,v) \geq \min\{1, C_5\}\|v\|_{1,\varepsilon}^2 \quad \text{for all } v \in H_0^1(\Omega). \tag{10.4}$$

Remark 10.2. Linß and Stynes (2001*b*) perform numerical experiments that compare several methods on the same Shishkin mesh for a test problem on the unit square that has exponential outflow layers along $x = 1$ and $y = 1$. The methods considered are central differencing, simple upwinding, the hybrid difference scheme of Linß and Stynes (1999), defect correction (see Remark 9.2), linear and bilinear Galerkin FEMs, and linear and bilinear streamline-diffusion FEMs (which we will discuss in Section 10.3). Graphs of the computed solutions, errors and convergence rates in the discrete $L^\infty(\Omega)$ norm are given, and known theoretical convergence results for the various methods are listed. It is concluded that, taking into account any difficulties that arise in solving the discrete linear systems, the methods that performed best for this problem are the defect correction method and the two streamline-diffusion FEMs, and that inside the layers bilinears yield more accuracy than linears.

10.3. The streamline-diffusion FEM

With linear or bilinear Galerkin methods, one has coercivity only with respect to the norm $\|\cdot\|_{1,\varepsilon}$ as shown in (10.4). This alone is insufficient to guarantee the stability of the method: numerical experiments on equidistant meshes will produce large oscillations like those seen for central differencing. Thus several finite element practitioners have devised FEMs that are coercive with respect to a stronger norm. Of these, the most commonly used is the *streamline-diffusion FEM* (SDFEM), which dates from 1979 (Hughes and Brooks 1979); it is also called the *streamline upwind Petrov–Galerkin* (SUPG) method.

Given a partition Ω^N of Ω, let S^N be a conforming space of piecewise polynomials of degree $k \geq 1$ defined on Ω^N. Define the SDFEM solution $u_{SD} \in S^N$ by

$$\begin{aligned} &B_{SD}(u_{SD}, w^N) \\ &\qquad := B(u_{SD}, w^N) + \sum_{\tau\in\Omega^N} \delta_\tau(-\varepsilon\Delta u_{SD} + \mathbf{a}.\nabla u_{SD} + bu_{SD},\, \mathbf{a}.\nabla w^N)_\tau \\ &\qquad = (f, w^N) + \sum_{\tau\in\Omega^N} \delta_\tau(f,\, \mathbf{a}.\nabla w^N)_\tau \quad \text{for all } w^N \in S^n. \end{aligned} \tag{10.5}$$

Here $B(\cdot,\cdot)$ is the standard bilinear form defined in (10.3), $(\cdot,\cdot)_\tau$ is the $L^2(\tau)$ inner product, and δ_τ is a nonnegative user-chosen piecewise constant that will be used to stabilize the method (if $\delta_\tau = 0$ for all $\tau \in \Omega^N$ then we return to the standard Galerkin method). The term $\sum_{\tau\in\Omega^N} \delta_\tau(f,\, \mathbf{a}.\nabla w^N)$ is included in the right-hand side of (10.5) to give the standard FEM property of *Galerkin orthogonality*:

$$B_{SD}(u - u_{SD}, w^N) = 0 \quad \text{for all } w^N \in S^N. \tag{10.6}$$

In the particular case when S^N comprises piecewise linears and $b \equiv 0$, the bilinear form simplifies to

$$\begin{aligned} B_{SD}(u_{SD}, w^N) = (\varepsilon\nabla u_{SD}, \nabla w^N) + (\mathbf{a}.\nabla u_{SD}, w^N) \\ + \sum_{\tau\in\Omega^N} \delta_\tau(\mathbf{a}.\nabla u_{SD},\, \mathbf{a}.\nabla w^N)_\tau, \end{aligned}$$

which is the same as the standard Galerkin bilinear form $B(\cdot,\cdot)$ associated with the differential operator $-\varepsilon\Delta u - \delta|\mathbf{a}|^2 u_{\mathbf{aa}} + \mathbf{a}.\nabla u$, where δ is a piecewise-constant function and $u_{\mathbf{a}}$ denotes the directional derivative in the subcharacteristic direction. That is, we have added artificial diffusion to the PDE, but only in the direction of the subcharacteristics, which for stationary problems are the same as the so-called *streamlines* of the differential operator. This is the explanation of the name SDFEM.

The SDFEM can be regarded as a Petrov–Galerkin method with trial space S^N and test space $\{w^N + \sum_{\tau\in\Omega^N} \delta_\tau \mathbf{a}.\nabla w^N : w^N \in S^N\}$, *i.e.*, the test functions are obtained by 'upwinding' the trial functions along the subcharacteristics. For this reason it is also known as the SUPG method.

Assume that the mesh is quasi-uniform, so that (Brenner and Scott 2002, §4.4) on each element $\tau \in \Omega^N$ one has the standard interpolation property

$$|u - u^I|_{m,\tau} \le C h_\tau^{k+1-m} |u|_{k+1,\tau} \quad \text{for } m = 0, 1, 2 \tag{10.7}$$

and the local inverse inequality

$$\|\Delta w^N\|_{0,\tau} \le C_{\text{inv}} h_\tau^{-1} |w^N|_{1,\tau} \quad \text{for all } w^N \in S^N, \tag{10.8}$$

where the $|\cdot|_{\ell,\tau}$ are local Sobolev seminorms on the element τ, the norm on $L^2(\tau)$ is $\|\cdot\|_{0,\tau}$, and h_τ denotes the diameter of τ.

Define a norm that is stronger than $\|\cdot\|_{1,\varepsilon}$ and natural for the analysis of the SDFEM: for each $v \in H^1(\Omega)$, set

$$\|v\|_{SD} = \left(\varepsilon|v|_1^2 + \sum_{\tau\in\Omega^N} \delta_\tau \|\mathbf{a}.\nabla v\|_{0,\tau}^2 + C_5\|v\|_0^2\right)^{1/2}.$$

Lemma 10.3. Suppose that the SDFEM parameter δ_τ satisfies

$$0 \le \delta_\tau \le \frac{1}{2} \min\left\{ \frac{C_5}{\|b\|_{L^\infty(\tau)}^2}, \frac{h_\tau^2}{\varepsilon C_{\text{inv}}^2} \right\} \quad \text{for each } \tau \in \Omega^N. \tag{10.9}$$

Then the bilinear form $B_{SD}(\cdot,\cdot)$ is coercive with respect to $\|\cdot\|_{SD}$ over

$S^N \times S^N$, that is,

$$B_{SD}(w^N, w^N) \geq \frac{1}{2}\|w^N\|_{SD}^2 \quad \text{for all } w^N \in \Omega^N.$$

Proof. For each $w^N \in \Omega^N$, we get easily

$$\begin{aligned} B_{SD}(w^N, w^N) \geq \varepsilon|w^N|_1^2 + C_5\|w^N\|_0^2 &+ \sum_{\tau\in\Omega^N} \delta_\tau\|\mathbf{a}.\nabla w^N\|_{0,\tau}^2 \\ &+ \sum_{\tau\in\Omega^N} \delta_\tau(-\varepsilon\Delta w^N + bw^N,\, \mathbf{a}.\nabla w^N)_\tau. \end{aligned} \tag{10.10}$$

Now the inequality $st \leq s^2 + t^2/4$ for s and $t \geq 0$, inequality (10.8) and the hypothesis on δ_τ yield

$$\begin{aligned} &\left|\sum_{\tau\in\Omega^N} \delta_\tau(-\varepsilon\Delta w^N + bw^N,\, \mathbf{a}.\nabla w^N)_\tau\right| \\ &\quad\leq \sum_{\tau\in\Omega^N}\left[\varepsilon^2\delta_\tau\|\Delta w^N\|_{0,\tau}^2 + \delta_\tau\|b\|_{L^\infty(\tau)}^2\|w^N\|_{0,\tau}^2 + \frac{1}{2}\delta_\tau\|\mathbf{a}.\nabla w^N\|_{0,\tau}^2\right] \\ &\quad\leq \frac{1}{2}\left[\varepsilon|w^N|_1^2 + C_5\|w^N\|_0^2 + \sum_{\tau\in\Omega^N}\delta_\tau\|\mathbf{a}.\nabla w^N\|_{0,\tau}^2\right]. \end{aligned}$$

Applying this bound in (10.10), the lemma is proved. □

One can exploit this result to derive an error estimate in a fairly standard way. Let $u^I \in S^N$ denote the nodal interpolant of u. Then, under the hypothesis of Lemma 10.3,

$$\|u^I - u_{SD}\|_{SD}^2 \leq 2\,B_{SD}(u^I - u_{SD}, u^I - u_{SD}) = 2\,B_{SD}(u^I - u, u^I - u_{SD}),$$

by the Galerkin orthogonality property (10.6). Applying Cauchy–Schwarz-type inequalities to the right-hand side here and invoking (10.7) and $\varepsilon\delta_\tau \leq Ch_\tau^2$ from (10.9), we arrive at (Roos *et al.* 1996, p. 232)

$$\|u^I - u_{SD}\|_{SD} \leq Ch^k\left[\sum_\tau(\varepsilon + \delta_\tau + \delta_\tau^{-1}h_\tau^2 + h_\tau^2)|u|_{k+1,\tau}^2\right]^{1/2},$$

where $h := \max_\tau h_\tau$ is the mesh diameter. In order to extract the best possible rate of convergence from this inequality while honouring the constraint on δ_τ in (10.9), set

$$\delta_\tau = \begin{cases} \delta_0 h_\tau & \text{for } \mathrm{Pe}_\tau > 1, \\ \delta_1 h_\tau^2/\varepsilon & \text{for } \mathrm{Pe}_\tau \leq 1, \end{cases} \tag{10.11}$$

where we define the *mesh Péclet number* $\mathrm{Pe}_\tau := \|\mathbf{a}\|_{L^\infty(\tau)}h_\tau/\varepsilon$. Here δ_0 and

δ_1 are user-chosen positive constants. The more important case $\mathrm{Pe}_\tau > 1$ is usually referred to as the convection-dominated case.

Remark 10.4. No precise general formula for an 'optimal' (in some sense) value of the SDFEM parameter δ_τ is known; the choice (10.11) seems to be the best statement that one can make. There has been much research into this question. For discussions of how to choose δ_τ see, *e.g.*, Akin and Tezduyar (2004), Brezzi and Russo (1994), Fischer, Ramage, Silvester and Wathen (1999), Houston and Süli (2001), Madden and Stynes (1996) and Roos *et al.* (1996).

The above analysis leads to the following bound (Roos *et al.* 1996, p. 233).

Theorem 10.5. Let each δ_τ be chosen according to (10.11) while satisfying the hypotheses of Lemma 10.3. Then there exists a constant C such that

$$\|u - u_{SD}\|_{SD} \le \|u - u^I\|_{SD} + \|u^I - u_{SD}\|_{SD} \le C(\varepsilon^{1/2} + h^{1/2})h^k|u|_{k+1}, \tag{10.12}$$

where u^N is the solution of the SDFEM method (10.5).

In a very technical paper Sangalli (2003) shows that in the one-dimensional case (3.12), on an equidistant grid the SDFEM yields a solution that is quasi-optimal with respect to a certain interpolated norm that is roughly similar to our norm $\|\cdot\|_{SD}$.

When the mesh is coarse everywhere, so we are in the convection-dominated case on all elements, then $\varepsilon \le Ch_\tau$ for all $\tau \in \Omega^N$ and the bound (10.12) becomes

$$\|u - u_{SD}\|_{SD} \le Ch^{k+1/2}|u|_{k+1}.$$

This implies that

$$\|u - u_{SD}\|_0 + \left(\sum_{\tau\in\Omega^N} \delta_\tau \|\mathbf{a}.\nabla(u - u_{SD})\|_{0,\tau}^2\right)^{1/2} \le Ch^{k+1/2}|u|_{k+1}, \tag{10.13}$$

Here the term $|u|_{k+1}$ is typically $\mathcal{O}\big(\epsilon^{-k-1/2}\big)$. In general this will dominate the $h^{k+1/2}$ term and consequently (10.13) does not imply that the error $u - u_{SD}$ is small in some norm. Thus this estimate is of limited value. Nevertheless one can choose some maximal subset $\hat\Omega$ of Ω that excludes all layers, restrict the norms in (10.13) to $\hat\Omega$, then prove essentially the same bounds again (in terms of the new norms) by means of cut-off functions (Roos *et al.* 1996, §II.3.2.1).

Recalling that $\delta_\tau = \mathcal{O}(h_\tau)$ in the convection-dominated case, we see that in (10.13) the error bound for the *streamline derivative* $\mathbf{a}.\nabla u$ is of optimal order, but the estimate of $\|u - u^N\|_0$ is order 1/2 less than optimal. This apparent loss of accuracy in the L^2 norm has attracted much attention.

Whether or not the bound on $\|u-u^N\|_0$ was sharp remained unresolved for many years until Zhou (1997) constructed a simple example for piecewise linears on a special mesh where the SDFEM converged with order only 1.5. For bilinears the situation is different. (Recall the comments on the numerical results for linears versus bilinears in Linß and Stynes (2001*b*).) For the SDFEM on the unit square Ω, under the usual hypotheses that u has only exponential boundary and corner layers, Stynes and Tobiska (2003) prove convergence results on a Shishkin mesh that imply, *inter alia*, $\|u-u^N\|_0 \leq CN^{-2}\ln^2 N$. The fundamental difference between bilinears and linears in the analysis is that for bilinears one has sharp interpolation error identities (Lin 1991) that enable the analysis to be carried out separately on each rectangle, while for triangles the corresponding identities require one to combine neighbouring elements to obtain an optimal error bound and this is not feasible on, *e.g.*, a Shishkin mesh.

Remark 10.6. Lemma 10.3 implies an *a priori* estimate for the SDFEM solution u_{SD}:

$$\|u_{SD}\|_{SD} \leq C\biggl(\|f\|_0^2 + \sum_\tau \delta_\tau \|f\|_{0,T}^2\biggr)^{1/2}. \tag{10.14}$$

Thus the method retains some control over the streamline derivative $\mathbf{a}.\nabla u_{SD}$ of the computed solution. In the more interesting convection-dominated case, with $\delta_\tau = \delta_0 h_\tau$, inequality (10.14) says essentially that, on a quasi-uniform mesh, $\|\mathbf{a}.\nabla u_{SD}\|_{0,\tau}$ can be at most $\mathcal{O}(h_\tau^{1/2})$. It is this property that distinguishes the SDFEM from a standard Galerkin method, for whose oscillatory solution u^N one can have $\|\mathbf{a}.\nabla u^N\|_{0,\tau} = \mathcal{O}(1)$, since the slope of u^N is locally $\mathcal{O}(h_\tau^{-1})$.

This enhanced stability in the subcharacteristic direction means that the SDFEM can compute fairly satisfactory exponential layers in solutions of convection-diffusion problems, provided that δ_τ is chosen carefully. Note however that the method contains no mechanism for stabilization perpendicular to the subcharacteristics, so along characteristic layers the computed solution typically displays oscillations; as usual with such layers, these oscillations are confined to a fairly small neighbourhood of the layer.

Kopteva (2004) gives a detailed analysis of the accuracy of the SDFEM inside characteristic layers.

Figure 10.2 shows a solution computed by the SDFEM for a problem with an interior layer and two outflow exponential layers. The computed solution has oscillations along the interior layer and at one of the outflow boundary layers. In this example δ_τ is for simplicity set equal to the same value on all triangles of the equidistant mesh, but this is not in general the best approach. The same problem is solved again in Figure 10.3 but the common value of δ_τ has been increased judiciously to the value recommended

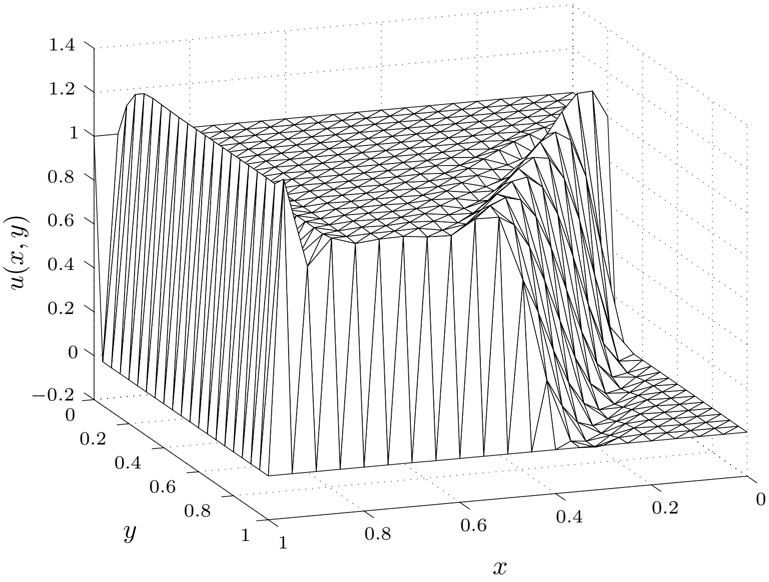

Figure 10.2. SDFEM I: δ_τ is the same for all τ.

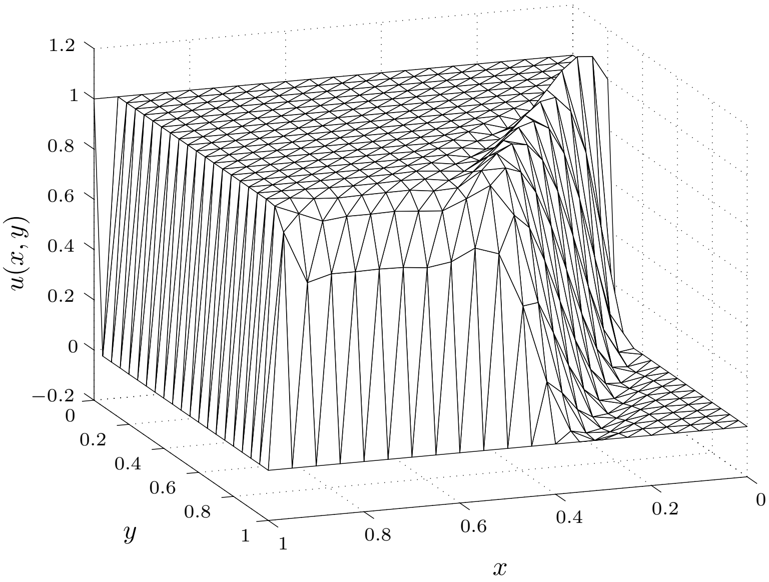

Figure 10.3. SDFEM II: increased value for δ_τ.

in Madden and Stynes (1996) to compute a sharp layer without oscillations along $x = 1$; unfortunately the outflow layer at $y = 1$ is smeared since the value of δ_T is now larger than optimal for that layer. In line with the discussion in Remark 10.6, the increase in δ_T has had little effect on the computed interior layer compared with the dramatic changes in the computed outflow layers.

Remark 10.7. In order to reduce or remove any oscillations that appear along characteristic layers, several authors have modified the SDFEM by adding artificial crosswind diffusion to the PDE or even by introducing nonlinear 'shock-capturing' terms into the SDFEM formulation. Expert opinion is divided on the value of this approach; for contrasting views see Shih and Elman (2000) and Knopp, Lube and Rapin (2002).

Remark 10.8. The SDFEM can of course be combined with a Shishkin mesh, and this technique was used to compute $u(x, y)$ in Figures 6.2, 6.3 and 6.5. In Figure 6.4 the SDFEM was also used but on an equidistant mesh; one can see that the interior layer in this figure is less sharp.

10.4. Discontinuous Galerkin finite element method

Recently, the *discontinuous Galerkin FEM* (DGFEM) has attracted a great deal of attention from many distinguished researchers. Like the SDFEM it achieves stability by a judicious choice of bilinear form, but the details of the construction are very different from Section 10.3.

Its name comes from its use of a standard piecewise polynomial trial space that is not required to be continuous across element boundaries. This local nature means the method is more readily parallelizable than (say) the SDFEM, and clearly permits the use of polynomials of different degrees on different elements, which can be exploited to gain increased accuracy when the problem is quite smooth on only part of the domain – as is usually the case with convection-diffusion problems. A drawback is the much larger number of degrees of freedom compared with finite element spaces that lie in $C(\Omega)$.

Methods of this type were first introduced in the 1970s and today there are several prominent variants. Arnold, Brezzi, Cockburn and Marini (2001/02) consider the problem $-\Delta u = f$ on Ω with $u = 0$ on $\partial\Omega$ and show that nine distinct versions of the DGFEM can be placed in the framework of a mixed-method weak formulation. They go on to analyse the stability of these methods, but this is of limited value in the context of convection-diffusion problems where the Laplacian is multiplied by a small parameter. This paper also gives an account of the historical development of DGFEMs that includes methods specifically designed for convection-diffusion problems.

Given the diversity of methods described as DGFEMs, we shall not attempt to give a thorough survey of this area. Instead we concentrate on one variant and the references appearing in this subsection will assist the reader who wishes to broaden his or her knowledge of the DGFEM.

Consider the nonsymmetric interior penalty DGFEM (NIPD) from Houston, Schwab and Süli (2002); related methods appear in, *e.g.*, Oden, Babuška and Baumann (1998) and Rivière, Wheeler and Girault (2001).

Assume that Ω is polygonal. Let $\mathcal{T}$ be a partition of Ω into elements κ (*e.g.*, triangles or rectangles). Houston *et al.* (2002) permit up to one hanging node for each κ, but for simplicity we shall assume that our partition has no hanging nodes. Assume also that each $\kappa \in \mathcal{T}$ is an affine image of a fixed master element $\hat{\kappa}$, *i.e.*, that $\kappa = F_\kappa(\hat{\kappa})$ where $\hat{\kappa}$ is either the open unit simplex or the open unit square in $\mathbb{R}^2$. For each nonnegative integer k, let $\mathcal{P}_k(\hat{\kappa})$ denote the set of polynomials of total degree k on $\hat{\kappa}$. (If $\hat{\kappa}$ is the unit square, one can also consider $\mathcal{Q}_k(\hat{\kappa})$, the set of all tensor-product polynomials on $\hat{\kappa}$ of degree k in each coordinate direction.) For each $\kappa \in \mathcal{T}$ write p_κ for the local polynomial degree. Set $\mathbf{p} = \{p_\kappa : \kappa \in \mathcal{T}\}$ and $\mathbf{F} = \{F_\kappa : \kappa \in \mathcal{T}\}$ and define the finite element space

$$S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F}) = \{v \in L^2(\Omega) : v|_\kappa \circ F_\kappa \in \mathcal{R}_{p_\kappa}(\hat{\kappa})\},$$

where $\mathcal{R}$ is either $\mathcal{P}$ or $\mathcal{Q}$.

For $s = 0, 1$ define the broken Sobolev spaces

$$H^s(\Omega, \mathcal{T}) = \{v \in L^2(\Omega) : v|_\kappa \in H^s(\kappa) \text{ for all } \kappa \in \mathcal{T}\}.$$

Let $\partial\kappa$ denote the boundary of κ for each $\kappa \in \mathcal{T}$. Define the inflow and outflow parts of $\partial\kappa$ by

$$\begin{aligned} \partial^-\kappa &= \{(x,y) \in \partial\kappa : \mathbf{a}(x,y).\mu_\kappa(x,y) < 0\}, \\ \partial^+\kappa &= \{(x,y) \in \partial\kappa : \mathbf{a}(x,y).\mu_\kappa(x,y) \geq 0\} \end{aligned}$$

respectively, where $\mu_\kappa(x,y)$ denotes the outward-pointing unit normal to $\partial\kappa$ at $(x,y) \in \partial\kappa$.

Let $v \in H^1(\Omega, \mathcal{T})$. For each $\kappa \in \mathcal{T}$, denote by v_κ^+ the inner trace of $v|_\kappa$ on $\partial\kappa$. If $\partial^-\kappa \setminus \partial\Omega$ is nonempty, then for almost every point $(x,y) \in \partial^-\kappa \setminus \partial\Omega$ there exists a unique $\kappa' \in \mathcal{T}$ (which depends on (x,y)) such that $x \in \partial^+\kappa'$ and $\kappa' \cap (\partial^-\kappa \setminus \partial\Omega)$ has nonzero one-dimensional measure, and we define the outer trace v_κ^- of v on $\partial^-\kappa \setminus \partial\Omega$ relative to κ to be the inner trace $v_{\kappa'}^+$ relative to κ'. Then define the jump of v across $\partial^-\kappa \setminus \partial\Omega$ by $\lfloor v \rfloor_\kappa = v_\kappa^+ - v_\kappa^-$.

We shall drop the subscript κ from the above notation when it is clear from the context what is intended.

Let $\mathcal{E}_{\text{int}}$ be the set of all open one-dimensional edges of the partition $\mathcal{T}$ that lie in Ω. Set $\Gamma_{\text{int}} = \{x \in \Omega : x \in e \text{ for some } e \in \mathcal{E}_{\text{int}}\}$. Numbering the elements κ consecutively, for each $e \in \mathcal{E}_{\text{int}}$ there exist indices i and j such that $i > j$ and the elements κ_i and κ_j share the interface e. Define the (element-numbering-dependent) jump of $v \in H^1(\Omega, \mathcal{T})$ across e and the mean value of v on e by

$$[v]_e = v|_{\partial\kappa_i\cap e} - v|_{\partial\kappa_j\cap e} \quad \text{and} \quad \langle v\rangle_e = \tfrac{1}{2}\big(v|_{\partial\kappa_i\cap e} + v|_{\partial\kappa_j\cap e}\big)$$

respectively. Furthermore, for each $e \in \mathcal{E}_{\text{int}}$ let ν denote the unit normal vector pointing from κ_i to κ_j ; if $e \subset \partial\Omega$, take $\nu = \mu$.

The bilinear form associated with the NIPD for (6.1a) with $u \equiv 0$ on $\partial\Omega$ is

$$\begin{aligned} B_{DG}(v,w) = \sum_{\kappa\in\mathcal{T}} \bigg(&\varepsilon \int_\kappa \nabla v.\nabla w \,\mathrm{d}x + \int_\kappa (\mathbf{a}.\nabla v + bv)w \,\mathrm{d}x \\ &- \int_{\partial^-\kappa\cap\partial^-\Omega} (\mathbf{a}.\mu)v^+w^+ \,\mathrm{d}s - \int_{\partial^-\kappa\backslash\partial\Omega} (\mathbf{a}.\mu_\kappa)\lfloor v\rfloor w^+ \,\mathrm{d}s \bigg) \\ &+ \varepsilon \int_{\partial\Omega} \big(v(\nabla w.\mu) - (\nabla v.\mu)w\big) \,\mathrm{d}s + \int_{\partial\Omega} \sigma vw \,\mathrm{d}s \\ &+ \varepsilon \int_{\Gamma_{\text{int}}} ([v]\langle\nabla w.\nu\rangle - \langle\nabla v.\nu\rangle[w]) \,\mathrm{d}s + \int_{\Gamma_{\text{int}}} \sigma[v][w] \,\mathrm{d}s, \end{aligned}$$

for all $v, w \in H^1(\Omega, \mathcal{T})$. Here σ, the user-chosen nonnegative *discontinuity-penalization parameter*, is defined by

$$\sigma|_e = \sigma_e \quad \text{for each } e \in \mathcal{E}_{\text{int}} \cup \partial\Omega.$$

Houston *et al.* (2002) choose $\sigma_e = \mathcal{O}(\varepsilon/h_e)$ where h_e is the length of edge e.

The NIPD method is then: find $u_{DG} \in S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$ such that

$$B(u_{DG}, w^N) = \sum_{\kappa\in\mathcal{T}} \int_\kappa f w^N \,\mathrm{d}x\,\mathrm{d}y \quad \text{for all } w^N \in S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F}). \qquad (10.15)$$

Existence and uniqueness of a solution to (10.15) are shown in Houston *et al.* (2002) by combining results from earlier papers of these authors.

Assuming that $u \in H^2(\Omega, \mathcal{T})$ and ∇u is continuous across each edge $e \in \mathcal{E}_{\text{int}}$, one can deduce the Galerkin orthogonality property

$$B_{DG}(u - u_{DG}, w^N) = 0 \quad \text{for all } w^N \in S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F}).$$

For all $v \in H^2(\Omega, \mathcal{T})$ define the norm $\|\cdot\|_{DG}$ by $\|v\|_{DG}^2 = B_{DG}(v, v)$. Setting $(v, w)_e = \int_e |\mathbf{a}.\mu_\kappa| vw \,\mathrm{d}s$ for each $e \subset \partial\kappa$ and $\|v\|_e^2 = (v, v)_e$, after some manipulation we get

$$\begin{aligned} \|v\|_{DG}^2 = &\sum_{\kappa\in\mathcal{T}} \big(\varepsilon\|\nabla v\|_{0,\kappa}^2 + \|c_0 v\|_{0,\kappa}^2\big) + \int_{\partial\Omega} \sigma v^2 \,\mathrm{d}s + \int_{\Gamma_{\text{int}}} \sigma[v]^2 \,\mathrm{d}s \\ &+ \tfrac{1}{2} \sum_{\kappa\in\mathcal{T}} \big(\|v^+\|_{\partial^-\kappa\cap\partial\Omega}^2 + \|v^+ - v^-\|_{\partial^-\kappa\backslash\partial\Omega}^2 + \|v^+\|_{\partial^+\kappa\cap\partial\Omega}^2\big), \end{aligned}$$

where $\|\cdot\|_{0,\kappa}$ is the $L^2(\kappa)$ norm and we set

$$c_0(x, y) = \sqrt{b(x, y) - \operatorname{div} \mathbf{a}(x, y)/2};$$

by (10.1) the function c_0 is well defined. Clearly $\|\cdot\|_{DG}$ is stronger than $\|\cdot\|_{1,\varepsilon}$.

Now Houston *et al.* (2002) write $u - u_{DG} = (u - \Pi u) + (\Pi u - u_{DG})$ where Π is the orthogonal projector in L^2 into $S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$. From Galerkin orthogonality we have

$$\|\Pi u - u_{DG}\|_{DG}^2 = B_{DG}(\Pi u - u_{DG}, \Pi u - u_{DG}) = B_{DG}(\Pi u - u, \Pi u - u_{DG}), \tag{10.16}$$

and, under the assumption that $\mathbf{a}.\nabla w^N|_\kappa$ lies in $S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$ for all $w^N \in S^{\mathbf{p}}(\Omega, \mathcal{T}, \mathbf{F})$, some analysis of the right-hand side of (10.16) enables $\|\Pi u - u_{DG}\|_{DG}$ to be estimated in terms of various norms of $u - \Pi u$. Invoking the triangle inequality $\|u - u_{DG}\|_{DG} \le \|u - \Pi u\|_{DG} + \|\Pi u - u_{DG}\|_{DG}$ then leads to a bound on $\|u - u_{DG}\|_{DG}$. In the particular case where the mesh elements are rectangles, piecewise polynomials of degree p are used, h is the mesh diameter and the solution u lies in $H^{p+1}(\Omega)$, the bound becomes

$$\|u - u_{DG}\|_{DG} \le C(\varepsilon^{1/2} h^p + h^{p+1/2})\|u\|_{H^{p+1}(\Omega)}, \quad \text{where } C = C(p).$$

Note that the right-hand side here depends on a Sobolev norm of u that is typically $\mathcal{O}(\varepsilon^{-p-1/2})$. It may be possible to use cut-off functions to localize this result away from layers, removing this undesirable feature.

The above analysis from Houston *et al.* (2002) assumes that the mesh is nondegenerate (Brenner and Scott 2002, §4.4), which excludes the long thin elements one expects in any mesh that is specifically designed to improve the behaviour of the method inside layers. Roos and Zarin (2003) apply this DGFEM to a problem on the unit square that has exponential layers along $x = 1$ and $y = 1$ and no other layers. Working with piecewise bilinears on a rectangular Shishkin mesh like that of Figure 9.2 with N mesh intervals in each coordinate direction, they adapt the analysis of Houston *et al.* (2002) to this situation (which entails a different choice for σ_e on part of the mesh) and prove that

$$\|u - u_{DG}\|_{DG} \le CN^{-1} \ln^{3/2} N. \tag{10.17}$$

A related paper (Zarin and Roos 2005) considers a problem similar to Example 6.1 and, using a Shishkin mesh similar to the one in Figure 9.3 with N mesh intervals in each coordinate direction, again obtains the bound (10.17).

We remind the reader that there is no universal agreement on a 'best' form of the DGFEM. For example, Gopalakrishnan and Kanschat (2003) consider a symmetric version of our bilinear form $B_{DG}(v, w)$ that is obtained by changing the signs of the terms $\varepsilon \int_{\partial\Omega} v(\nabla w.\mu)\,\mathrm{d}s$ and $\varepsilon \int_{\Gamma_{\text{int}}} [v]\langle\nabla w.\nu\rangle\,\mathrm{d}s$. A good sense of the breadth of interest in the DGFEM and the variety of its manifestations can be inferred from the collection of papers in Cockburn, Karniadakis and Shu (2000).

10.5. Adaptive methods

Adaptive FEMs compute a solution to a boundary-value problem on some conventional (*e.g.*, equidistant) mesh using some stable method such as SDFEM, then use this solution to compute *a posteriori* some local error estimator that gives guidance on where one should refine or coarsen the mesh to obtain a mesh better suited to the boundary-value problem. On this new mesh one then computes a fresh solution to the problem, then the mesh is again modified based on the local error estimator. The process is continued iteratively until some stopping criterion is reached. See Ainsworth and Oden (2000) or Brenner and Scott (2002, Chapter 9) for a more precise description.

There is perhaps a general consensus that in the long run adaptive methods will provide the most satisfactory approach to solving convection-diffusion problems, but today their behaviour when applied to such problems is still poorly understood, despite many published numerical experiments. John (2000) gives numerical examples of how apparently reasonable error estimators can yield inaccurate solutions to convection-diffusion problems.

A difficulty with the theory of *a posteriori* error estimators for convection-diffusion problems is that published inequalities relating the estimator to the true error frequently contain multiplicative factors that depend badly on the small diffusion parameter ε. This seriously undermines the validity of the estimator. Below we shall confine our discussion to a few ε-independent results that have been obtained.

For the one-dimensional problem (3.12), an adaptive-mesh algorithm that is based on arc-length equidistribution (where mesh points are moved but no points are created or deleted) is analysed by Kopteva and Stynes (2001), using earlier *a posteriori* bounds from Kopteva (2001). It is shown that, starting from an equidistant mesh with N subintervals, after $\mathcal{O}(\ln(1/\varepsilon)/(\ln N))$ iterations one obtains a computed solution u^N that resolves the layer with moreover $|u(x_i) - u_i^N| \leq CN^{-1}$ for all i. The underlying numerical method is simple upwinding so this is a finite difference approach, but we include it here since it is a clear convergence result for an adaptive method and few such results exist for convection-diffusion problems. It seems difficult to extend this type of result to two-dimensional problems.

In Sangalli (2001) the *residual-free bubble* FEM is considered; this method is related to the SDFEM (Brezzi, Marini and Süli 2000). An error estimator based on element residuals and jumps in the normal derivative of the solution across edges is shown to be robust for (6.1), *i.e.*, the global value of the estimator is equivalent to the true error up to a constant factor that is independent of ε, but the norm in which the true error is measured is

$$w \mapsto \varepsilon |w|_{H^1(\Omega)} + \|\mathbf{a}.\nabla w\|_{H^{-1}(\Omega)},$$

which is weak: the factor multiplying $|\cdot|_{H^1(\Omega)}$ is ε, not the more natural $\varepsilon^{1/2}$ that appears in the weighted energy norm $\|\cdot\|_{1,\varepsilon}$ of Section 7.

The *dual-weighted-residual* method for goal-oriented error estimation has been successfully applied to convection-diffusion problems by various authors; see Eriksson, Estep, Hansbo and Johnson (1996) and Bangerth and Rannacher (2003). Here the aim is to adapt the mesh in order to compute accurately some functional of the solution but not the solution itself. The theoretical basis for this method has recently been surveyed in *Acta Numerica* (Giles and Süli 2002) so we shall not discuss it further here.

Finally, Verfürth (2004) shows that for the SDFEM the error in the computed solution is equivalent (up to a constant factor that is independent of ε) to the global value of each of three different estimators (one based on element and edge residuals; one based on the solution of local Dirichlet problems; one based on the solution of local Neumann problems). The true error is measured in a norm

$$w \mapsto \|w\|_{1,\varepsilon} + \|w\|_*,$$

where $\|\cdot\|_*$ is the dual norm on $H^{-1}(\Omega)$ defined by

$$\|w\|_* = \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{(w,v)}{\|v\|_{1,\varepsilon}},$$

with $(\cdot,\cdot)$ the corresponding duality pairing. (This special norm is used to bound the convective term.) But the paper assumes that the mesh is quasi-uniform, which excludes the long thin elements that one expects an adaptive code to construct when solving a convection-diffusion problem.

In summary, we do not have today a satisfactory adaptive method for two-dimensional convection-diffusion problems that, starting from an ordinary coarse mesh, is guaranteed to produce a layer-adapted mesh with a bound on the error in the computed solution in some reasonably strong norm.

11. Concluding remarks

Our survey has not been exhaustive. For example, the *hp* finite element method appeared only in an incidental way in the title of Houston *et al.* (2002) in Section 10.4. For general surveys of methods for convection-diffusion problems see Morton (1996) and Roos, Stynes and Tobiska (2005). (For the *hp* finite element method see Schwab (1998), and also Melenk (2002), where singularly perturbed linear reaction-diffusion problems are examined in great detail.)

Time-dependent convection-diffusion problems are of great practical importance but space constraints did not allow their discussion here. As well as the general references cited above, see Ewing and Wang (2001) and Hundsdorfer and Verwer (2003).

The numerical analysis and solution of convection-diffusion problems on polygonal regions, where the solution is assumed to exhibit boundary but not interior layers and one has sufficient compatibility of the data at the corners of the domain, is by now fairly well understood in the framework of Shishkin meshes combined with finite difference or finite element methods. When we consider interior layers (and the effects of data incompatibilities at corners) our grasp is much less sure and there are several competing methods. In the long run the view of this author is that adaptive methods will triumph over all types of convection-diffusion problem, but much work remains to be done.

Acknowledgements

A first version of this survey was written while the author was on a sabbatical visit at the Department of Mathematics and the Industrial Mathematics Institute at the University of South Carolina, USA, for whose support I am deeply grateful. I am greatly indebted to Niall Madden, who cheerfully and efficiently provided all the figures. Hans-Görg Roos and Lutz Tobiska kindly read the entire article and offered helpful criticisms, for which I wish to express my gratitude.

REFERENCES

M. Ainsworth and J. T. Oden (2000), *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience, New York.

J. E. Akin and T. E. Tezduyar (2004), 'Calculation of the advective limit of the SUPG stabilization parameter for linear and higher-order elements', *Comput. Methods Appl. Mech. Engrg.* **193**, 1909–1922.

D. N. d. G. Allen and R. V. Southwell (1955), 'Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder', *Quart. J. Mech. Appl. Math.* **8**, 129–145.

V. B. Andreev and N. V. Kopteva (1996), 'Investigation of difference schemes with an approximation of the first derivative by a central difference relation', *Zh. Vychisl. Mat. i Mat. Fiz.* **36**(8), 101–117.

T. Apel (1999), *Anisotropic Finite Elements: Local Estimates and Applications*, Advances in Numerical Mathematics, B. G. Teubner, Stuttgart.

T. Apel and M. Dobrowolski (1992), 'Anisotropic interpolation with applications to the finite element method', *Computing* **47**, 277–293.

D. N. Arnold, F. Brezzi, B. Cockburn and L. D. Marini (2001/02), 'Unified analysis of discontinuous Galerkin methods for elliptic problems', *SIAM J. Numer. Anal.* **39**, 1749–1779 (electronic).

W. Bangerth and R. Rannacher (2003), *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser, Basel.

E. Bohl (1981), *Finite Modelle gewöhnlicher Randwertaufgaben*, Teubner, Stuttgart.

K. Bohmer and R. Rannacher (1984), *Defect Correction Methods: Theory and Applications*, Springer, Berlin.

A. Brandt and I. Yavneh (1991), 'Inadequacy of first-order upwind difference schemes for some recirculating flows', *J. Comput. Phys.* **93**, 128–143.

S. C. Brenner and L. R. Scott (2002), *The Mathematical Theory of Finite Element Methods*, Vol. 15 of *Texts in Applied Mathematics*, 2nd edn, Springer, New York.

F. Brezzi and A. Russo (1994), 'Choosing bubbles for advection-diffusion problems', *Math. Models Methods Appl. Sci.* **4**, 571–587.

F. Brezzi, D. Marini and E. Süli (2000), 'Residual-free bubbles for advection-diffusion problems: the general error analysis', *Numer. Math.* **85**, 31–47.

B. Cockburn, G. E. Karniadakis and C.-W. Shu, eds (2000), *Discontinuous Galerkin Methods: Theory, Computation and Applications*, Vol. 11 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin. Papers from the 1st International Symposium held in Newport, RI, May 24–26, 1999.

M. Dobrowolski and H.-G. Roos (1997), '*A priori* estimates for the solution of convection-diffusion problems and interpolation on Shishkin meshes', *Z. Anal. Anwendungen* **16**, 1001–1012.

W. Dörfler (1999), 'Uniform *a priori* estimates for singularly perturbed elliptic equations in multidimensions', *SIAM J. Numer. Anal.* **36**, 1878–1900 (electronic).

K. Eriksson, D. Estep, P. Hansbo and C. Johnson (1996), *Computational Differential Equations*, Cambridge University Press, Cambridge.

R. E. Ewing and H. Wang (2001), 'A summary of numerical methods for time-dependent advection-dominated partial differential equations', *J. Comput. Appl. Math.* **128**, 423–445. Numerical analysis 2000, Vol. VII, Partial differential equations.

P. Farrell, A. Hegarty, J. Miller, E. O'Riordan and G. Shishkin (2000), *Robust Computational Techniques for Boundary Layers*, Chapman & Hall/CRC, Boca Raton.

B. Fischer, A. Ramage, D. J. Silvester and A. J. Wathen (1999), 'On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems', *Comput. Methods Appl. Mech. Engrg.* **179**, 179–195.

A. Fröhner, T. Linß and H.-G. Roos (2001), 'Defect correction on Shishkin-type meshes', *Numer. Algorithms* **26**, 281–299.

D. Gilbarg and N. S. Trudinger (2001), *Elliptic Partial Differential Equations of Second Order*, Classics in Mathematics, Springer, Berlin. Reprint of the 1998 edition.

M. B. Giles and E. Süli (2002), Adjoint methods for PDEs: *a posteriori* error analysis and postprocessing by duality, in *Acta Numerica*, Vol. 11, Cambridge University Press, pp. 145–236.

H. Goering, A. Felgenhauer, G. Lube, H.-G. Roos and L. Tobiska (1983), *Singularly Perturbed Differential Equations*, Vol. 13 of *Mathematical Research*, Akademie, Berlin.

J. Gopalakrishnan and G. Kanschat (2003), 'A multilevel discontinuous Galerkin method', *Numer. Math.* **95**, 527–550.

P. Grisvard (1985), *Elliptic Problems in Nonsmooth Domains*, Vol. 24 of *Monographs and Studies in Mathematics*, Pitman (Advanced Publishing Program), Boston, MA.

H. Han and R. B. Kellogg (1990), 'Differentiability properties of solutions of the equation $-\epsilon^2 \Delta u + ru = f(x, y)$ in a square', *SIAM J. Math. Anal.* **21**, 394–408.

P. Houston and E. Süli (2001), 'Stabilised *hp*-finite element approximation of partial differential equations with nonnegative characteristic form', *Computing* **66**, 99–119.

P. Houston, C. Schwab and E. Süli (2002), 'Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems', *SIAM J. Numer. Anal.* **39**, 2133–2163 (electronic).

T. J. R. Hughes and A. Brooks (1979), A multidimensional upwind scheme with no crosswind diffusion, in *Finite Element Methods for Convection Dominated Flows* (*Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979*), Vol. 34 of *AMD*, Amer. Soc. Mech. Engrs. (ASME), New York, pp. 19–35.

W. Hundsdorfer and J. Verwer (2003), *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Vol. 33 of *Springer Series in Computational Mathematics*, Springer, Berlin.

A. M. Il'in (1969), 'A difference scheme for a differential equation with a small parameter multiplying the highest derivative', *Mat. Zametki* **6**, 237–248.

A. M. Il'in (1992), *Matching of Asymptotic Expansions of Solutions of Boundary Value Problems*, Vol. 102 of *Translations of Mathematical Monographs*, AMS, Providence, RI. Translated from the Russian by V. Minachin [V. V. Minakhin].

V. John (2000), 'A numerical study of *a posteriori* error estimators for convection-diffusion equations', *Comput. Methods Appl. Mech. Engrg.* **190**, 757–781.

R. B. Kellogg and M. Stynes (2005), 'Corner singularities and boundary layers in a simple convection-diffusion problem', *J. Differential Equations*, to appear.

R. B. Kellogg and M. Stynes (2004), A singularly perturbed convection-diffusion problem in a half-plane, Technical Report 13, Industrial Mathematics Institute, University of South Carolina.

R. B. Kellogg and A. Tsan (1978), 'Analysis of some difference approximations for a singular perturbation problem without turning points', *Math. Comp.* **32**, 1025–1039.

J. Kevorkian and J. D. Cole (1996), *Multiple Scale and Singular Perturbation Methods*, Vol. 114 of *Applied Mathematical Sciences*, Springer, New York.

T. Knopp, G. Lube and G. Rapin (2002), 'Stabilized finite element methods with shock capturing for advection-diffusion problems', *Comput. Methods Appl. Mech. Engrg.* **191**, 2997–3013.

N. Kopteva (2001), 'Maximum norm *a posteriori* error estimates for a one-dimensional convection-diffusion problem', *SIAM J. Numer. Anal.* **39**, 423–441 (electronic).

N. Kopteva (2003), 'Error expansion for an upwind scheme applied to a two-dimensional convection-diffusion problem', *SIAM J. Numer. Anal.* **41**, 1851–1869 (electronic).

N. Kopteva (2004), 'How accurate is the streamline-diffusion FEM inside characteristic (boundary and interior) layers?', *Comput. Methods Appl. Mech. Engrg.* **193**, 4875–4889.

N. Kopteva and M. Stynes (2001), 'A robust adaptive method for a quasi-linear one-dimensional convection-diffusion problem', *SIAM J. Numer. Anal.* **39**, 1446–1467 (electronic).

O. A. Ladyzhenskaya and N. N. Ural'tseva (1968), *Linear and Quasilinear Elliptic Equations*, Translated from the Russian by Scripta Technica, Inc. (translation editor, Leon Ehrenpreis), Academic Press, New York.

Q. Lin (1991), A rectangle test for finite element analysis, in *Proc. Syst. Sci. Eng.*, Great Wall (H.K.) Culture Publish Co., pp. 213–216.

T. Linß (2001), 'The necessity of Shishkin decompositions', *Appl. Math. Lett.* **14**, 891–896.

T. Linß (2003), 'Layer-adapted meshes for convection-diffusion problems', *Comput. Methods Appl. Mech. Engrg.* **192**, 1061–1105.

T. Linß (2005), 'On a convection-diffusion problem with a weak layer', *Appl. Math. Comput.* **160**, 791–795.

T. Linß and M. Stynes (1999), 'A hybrid difference scheme on a Shishkin mesh for linear convection-diffusion problems', *Appl. Numer. Math.* **31**, 255–270.

T. Linß and M. Stynes (2001*a*), 'Asymptotic analysis and Shishkin-type decomposition for an elliptic convection-diffusion problem', *J. Math. Anal. Appl.* **261**, 604–632.

T. Linß and M. Stynes (2001*b*), 'Numerical methods on Shishkin meshes for linear convection-diffusion problems', *Comput. Methods Appl. Mech. Engrg.* **190**, 3527–3542.

N. Madden and M. Stynes (1996), 'Linear enhancements of the streamline diffusion method for convection-diffusion problems', *Comput. Math. Appl.* **32**, 29–42.

N. Madden and M. Stynes (1997), 'Efficient generation of oriented meshes for solving convection-diffusion problems', *Int. J. Numer. Methods Engrg.* **40**, 565–576.

J. M. Melenk (2002), *hp-Finite Element Methods for Singular Perturbations*, Vol. 1796 of *Lecture Notes in Mathematics*, Springer, Berlin.

J. Miller, E. O'Riordan and G. Shishkin (1996), *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, Singapore.

K. W. Morton (1996), *Numerical Solution of Convection-Diffusion Problems*, Vol. 12 of *Applied Mathematics and Mathematical Computation*, Chapman & Hall, London.

J. T. Oden, I. Babuška and C. E. Baumann (1998), 'A discontinuous *hp* finite element method for diffusion problems', *J. Comput. Phys* **146**, 491–519.

M. H. Protter and H. F. Weinberger (1984), *Maximum Principles in Differential Equations*, Springer, New York. Corrected reprint of the 1967 original.

A. Quarteroni and A. Valli (1994), *Numerical Approximation of Partial Differential Equations*, Springer, Berlin.

B. Rivière, M. F. Wheeler and V. Girault (2001), '*A priori* error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems', *SIAM J. Numer. Anal.* **39**, 902–931 (electronic).

H.-G. Roos (1994), 'Ten ways to generate the Il'in and related schemes', *J. Comput. Appl. Math.* **53**, 43–59.

H.-G. Roos (1996), 'A note on the conditioning of upwind schemes on Shishkin meshes', *IMA J. Numer. Anal.* **16**, 529–538.

H.-G. Roos and H. Zarin (2003), The discontinuous Galerkin finite element method for singularly perturbed problems, in *Challenges in Scientific Computing: CISC 2002* (E. Bänsch, ed.), Vol. 35 of *Lecture Notes in Computational Science and Engineering*, Springer, Berlin, pp. 246–267.

H.-G. Roos, M. Stynes and L. Tobiska (1996), *Numerical Methods for Singularly Perturbed Differential Equations*, Springer, Berlin/Heidelberg/New York.

H.-G. Roos, M. Stynes and L. Tobiska (2005), *Numerical Methods for Singularly Perturbed Differential Equations*, 2nd edn, Springer, Berlin/Heidelberg/New York, in preparation.

R. Sacco and M. Stynes (1998), 'Finite element methods for convection-diffusion problems using exponential splines on triangles', *Comput. Math. Appl.* **35**, 35–45.

R. Sacco, E. Gatti and L. Gotusso (1999), 'A nonconforming exponentially fitted finite element method for two-dimensional drift-diffusion models in semiconductors', *Numer. Methods Partial Diff. Equations* **15**, 133–150.

G. Sangalli (2001), 'A robust *a posteriori* estimator for the residual-free bubbles method applied to advection-diffusion problems', *Numer. Math.* **89**, 379–399.

G. Sangalli (2003), 'Quasi optimality of the SUPG method for the one-dimensional advection-diffusion problem', *SIAM J. Numer. Anal.* **41**, 1528–1542 (electronic).

F. Schieweck (1986), Eine asymptotische angepaßte Finite-Element-Methode für singulär gestörte elliptische Randwertaufgaben, PhD thesis, Technische Hochschule Magdeburg, GDR.

C. Schwab (1998), *p- and hp-Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*, Numerical Mathematics and Scientific Computation, The Clarendon Press (Oxford University Press), New York.

Y.-T. Shih and H. C. Elman (2000), 'Iterative methods for stabilized discrete convection-diffusion problems', *IMA J. Numer. Anal.* **20**, 333–358.

G. I. Shishkin (1989), 'Approximation of solutions of singularly perturbed boundary value problems with a parabolic boundary layer', *Zh. Vychisl. Mat. i Mat. Fiz.* **29**, 963–977, 1102.

G. I. Shishkin (1990), Grid approximation of singularly perturbed elliptic and parabolic equations, second doctoral thesis, Keldysh Institute, Moscow.

M. Stynes (2003), Numerical methods for convection-diffusion problems or the 30 years war, in *20th Biennial Conf. on Numerical Analysis* (D. F. Griffiths and G. A. Watson, eds), Numerical Analysis Report NA/217, University of Dundee, UK, pp. 95–103.

M. Stynes and L. Tobiska (1998), 'A finite difference analysis of a streamline diffusion method on a Shishkin mesh', *Numer. Algorithms* **18**, 337–360.

M. Stynes and L. Tobiska (2003), ‘The SDFEM for a convection-diffusion problem with a boundary layer: optimal error analysis and enhancement of accuracy’, *SIAM J. Numer. Anal.* **41**, 1620–1642.

R. Verfürth (2004), Robust *a posteriori* error estimates for stationary convection-diffusion equations, Technical report, University of Bochum.

H. Zarin and H.-G. Roos (2005), ‘Interior penalty discontinuous approximations of convection-diffusion problems with parabolic layers’, *Numer. Math.*, to appear.

G. Zhou (1997), ‘How accurate is the streamline diffusion finite element method?’, *Math. Comp.* **66**, 31–44.