Springer Series in Computational Mathematics

24

Hans-Görg Roos
Martin Stynes
Lutz Tobiska

Robust Numerical Methods for Singularly Perturbed Differential Equations

Convection-Diffusion-Reaction
and Flow Problems

Second Edition
With 41 Figures

Springer

Hans-Görg Roos
Technische Universität Dresden
Fakultät Mathematik und
Naturwissenschaften
Institut für Numerische Mathematik
01062 Dresden
Germany
hans-goerg.roos@tu-dresden.de

Martin Stynes
Department of Mathematics
University College Cork
Western Road
Cork
Ireland
m.stynes@ucc.ie

Lutz Tobiska
Otto-von-Guericke-Universität
Magdeburg
Fakultät für Mathematik
Institut für Analysis und Numerik
PF 4120
39016 Magdeburg
Germany
tobiska@ovgu.de

Preface

The analysis of singular perturbed differential equations began early in the twentieth century, when approximate solutions were constructed from asymptotic expansions. (Preliminary attempts appear in the nineteenth century – see [vD94].) This technique has flourished since the mid-1960s and its principal ideas and methods are described in several textbooks; nevertheless, asymptotic expansions may be impossible to construct or may fail to simplify the given problem and then *numerical approximations* are often the only option.

The systematic study of numerical methods for singular perturbation problems started somewhat later – in the 1970s. From this time onwards the research frontier has steadily expanded, but the exposition of new developments in the analysis of these numerical methods has not received its due attention. The first textbook that concentrated on this analysis was [DMS80], which collected various results for ordinary differential equations. But after 1980 no further textbook appeared until 1996, when *three* books were published: Miller et al. [MOS96], which specializes in upwind finite difference methods on Shishkin meshes, Morton's book [Mor96], which is a general introduction to numerical methods for convection-diffusion problems with an emphasis on the cell-vertex finite volume method, and [RST96], the first edition of the present book. Nevertheless many methods and techniques that are important today, especially for partial differential equations, were developed after 1996. To give some examples, layer-adapted special meshes are frequently used, new stabilization techniques (such as discontinuous Galerkin methods and local subspace projections) are prominent, and there is a growing interest in the use of adaptive methods. Consequently contemporary researchers must comb the literature to gain an overview of current developments in this active area. In this second edition we retain the exposition of basic material that underpinned the first edition while extending its coverage to significant new numerical methods for singularly perturbed differential equations.

Our purposes in writing this introductory book are twofold. First, we present a structured and comprehensive account of current ideas in the numerical analysis of singularly perturbed differential equations. Second, this

important area has many open problems and we hope that our book will stimulate their investigation. Our choice of topics is inevitably personal and reflects our own main interests.

We have learned a great deal about singularly perturbed problems from other researchers. We thank those colleagues who helped and influenced us; these include V.B. Andreev, A.E. Berger, P.A. Farrell, A. Felgenhauer, E.C. Gartland, Ch. Großmann, A.F. Hegarty, V. John, R.B. Kellogg, N. Kopteva, G. Lube, N. Madden, G. Matthies, J.J.H. Miller, K.W. Morton, F. Schieweck, G.I. Shishkin, E. Süli, and R. Vulanović; in particular Herbert Goering and Eugene O'Riordan guided our initial steps in the area. Our research colleague T. Linß deserves additional thanks for providing many of the figures in this book.

Our work was supported by the Deutsche Forschungsgemeinschaft and by the Boole Centre for Research in Informatics at the National University of Ireland, Cork. We are grateful to them, to the Mathematisches Forschungsinstitut in Oberwolfach for its hospitality, and to Springer-Verlag for its cooperation.

Contents

Part I Ordinary Differential Equations

Notation

I	identity
L	differential operator
L^*	adjoint operator
$a(\cdot,\cdot)$	bilinear form
g, G	Green's function
V, V^*	Banach space and the corresponding dual space
V_h	finite-dimensional subspace of V
$\|\cdot\|_V$	norm on the space V
$\|\cdot\|_{*,d}$	discrete version of the norm $\|\cdot\|_*$
$r \cdot s$	scalar product of vectors in $\mathbb{R}^d$
$(\cdot,\cdot)$	scalar product in Hilbert space
$f(v)$ or $\langle f, v \rangle$	functional f applied to v
$\|f\|_*$	norm of the functional f
$U \hookrightarrow V$	continuous embedding of U in V
Ω	given space variable(s) domain
$\partial\Omega = \Gamma$	boundary of Ω
$meas\,(\Omega)$	measure of Ω
n	outward-pointing unit vector normal to $\partial\Omega$
t, T	time with $t \in (0, T)$
$Q = \Omega \times (0, T)$	given domain for nonstationary problems
$C^l(\Omega),\, C^{l,\alpha}(\Omega)$	function spaces
$L_p(\Omega)$	function space, $1 \le p \le \infty$
$\|\cdot\|_{0,p}$	norm in $L_p(\Omega)$
$\|\cdot\|_{L_p,d}$	discrete norm in $L_p(\Omega)$
$W^{m,p}(\Omega), \|\cdot\|_{m,p,\Omega}$	Sobolev spaces and their norms
$H^l(\Omega),\, H_0^l(\Omega)$	Sobolev spaces $W^{1,2}(\Omega)$
$\|\cdot\|_l, |\cdot|_l$	norm and seminorm in $H^l(\Omega)$
$\|\cdot\|_{l,E}$	H^l-norm restricted to $E \subset \Omega$
ε	singular perturbation parameter
C	generic constant, independent of ε

$\lVert \cdot \rVert_\varepsilon$	ε-weighted $H^1(\Omega)$ norm
$\lVert \cdot \rVert_{gr}$	graph norm
∇ or $grad$	gradient
div, div $c = \nabla \cdot c$	divergence operator
$\mathcal{O}(\cdot),\ o(\cdot)$	Landau symbols
P_r	polynomials of degree at most r
P_r^{disc}	piecewise polynomials of degree at most r, discontinuous across element boundaries
Q_r	products of polynomials of degree at most r
Q_r^{disc}	products of polynomials of degree at most r, discontinuous across element boundaries
h, h_i	mesh parameter in space
τ, τ_j	mesh parameter in time
L_h	difference operator
D^+, D^-, D^0	difference quotients
$\triangle, \triangle_h$	Laplacian and its discretization
$\omega_h,\ \Omega_h$	set of meshpoints
$u, u_h, u_i, u_i^j, u_{ij}$	unknown(s)
u_0	reduced solution
I_h	interpolation operator
$u^I = I_h\, u$	nodal interpolant of u
$\pi_h u, \Pi_h u, i_h u$	quasi-interpolant of u, defined for non-smooth functions u
	mesh-dependent norms are written with three vertical lines: $\lvert\lvert\lvert \cdot \rvert\rvert\rvert$
$\lvert\lvert\lvert \cdot \rvert\rvert\rvert_{SD}$	norm used in streamline diffusion finite element method
$\lvert\lvert\lvert \cdot \rvert\rvert\rvert_{CIP}$	norm used in continuous interior penalty finite element method
$\lvert\lvert\lvert \cdot \rvert\rvert\rvert_{LPS}$	norm used in local projection stabilization finite element method
$\lvert\lvert\lvert \cdot \rvert\rvert\rvert_{dG}$	norm used in discontinuous Galerkin finite element method
$\lvert\lvert\lvert \cdot \rvert\rvert\rvert_{GLS}$	norm used in the Galerkin least-squares finite element method

Introduction

Imagine a river – a river flowing strongly and smoothly. Liquid pollution pours into the water at a certain point. What shape does the pollution stain form on the surface of the river?

Two physical processes operate here: the pollution *diffuses* slowly through the water, but the dominant mechanism is the swift movement of the river, which rapidly *convects* the pollution downstream. Convection alone would carry the pollution along a one-dimensional curve on the surface; diffusion gradually spreads that curve, resulting in a long thin curved wedge shape.

When convection and diffusion are both present in a linear differential equation and convection dominates, we have a *convection-diffusion problem.*

The simplest mathematical model of a convection-diffusion problem is a two-point boundary value problem of the form

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

with $u(0) = u(1) = 0$, where ε is a small positive parameter and a, b and f are some given functions. Here the term u'' corresponds to diffusion and its coefficient $-\varepsilon$ is small. The term u' represents convection, while u and f play the rôles of a source and driving term respectively. (Spriet and Vansteenkiste [SV82] explain why diffusion and convection should be modelled by second-order and first-order derivatives respectively.)

Example 0.1. Consider the problem

$$-\varepsilon u''(x) + u'(x) = 1 \qquad \text{for} \quad 0 < x < 1, \tag{0.1}$$

with $u(0) = u(1) = 0$ and $0 < \varepsilon \ll 1$.

Suppose that we set formally $\varepsilon = 0$ here. This yields

$$u'(x) = 1 \qquad \text{for} \quad 0 < x < 1, \tag{0.2}$$

with $u(0) = u(1) = 0$. Unlike (0.1) this problem has no solution in $C^1[0,1]$. We infer that when ε is near zero, the solution of (0.1) is badly behaved in some way. ♣

Problems like (0.1) form the subject matter of this book. They are differential equations (ordinary or partial) that depend on a small positive parameter ε and whose solutions (or their derivatives) approach a discontinuous limit as ε approaches zero. Such problems are said to be *singularly perturbed*, where we regard ε as a perturbation parameter. In more technical terms, one cannot represent the solution of a singularly perturbed differential equation as an asymptotic expansion in powers of ε.

The solutions of singular perturbation problems typically contain *layers*. Ludwig Prandtl introduced the terminology *boundary layer* at the Third International Congress of Mathematicians in Heidelberg in 1904. (Prandtl's paper, "Über Flüssigkeitsbewegung bei sehr kleiner Reibung", is one of the most influential applied mathematics papers of the 20th century.) To see how such layers arise, consider the following time-dependent Navier-Stokes problem in two space variables x and y:

$$\frac{\partial u}{\partial t} - \frac{1}{\mathrm{Re}} \triangle u + (u \cdot \nabla)u = -\nabla p \quad \text{in the upper half-plane } y > 0, \qquad (0.3a)$$

$$\nabla \cdot u = 0 \quad \text{in the same domain}, \qquad (0.3b)$$

$$u = 0 \quad \text{on the boundary } y = 0, \qquad (0.3c)$$

at large Reynolds number Re. One can regard the boundary $y = 0$ as a fixed plate, and we assume that the velocity u at $y = \infty$ is parallel to the x-axis with magnitude U. We seek a flow, at constant pressure p, whose velocity is parallel to the plate and independent of x. Then equation (0.3a) reduces to

$$\frac{\partial u}{\partial t} = \varepsilon \frac{\partial^2 u}{\partial y^2}, \quad \text{where} \quad \varepsilon = \frac{1}{\mathrm{Re}}.$$

Set $\eta = y/(2\sqrt{\varepsilon t})$ and let $u(y,t) = U f(\eta)$. A computation leads to

$$u = 2U \operatorname{erf}(\eta), \quad \text{where} \quad \operatorname{erf}(\eta) = \frac{1}{\sqrt{\pi}} \int_0^\eta e^{-s^2} ds. \qquad (0.4)$$

Equation (0.4) shows that there is a narrow region near $y = 0$ where u departs significantly from the constant flow U. We say that u has a *boundary layer* at $y = 0$. See [CM93] for a detailed discussion. Linearization of (0.3) yields an equation of the form

$$\frac{\partial u}{\partial t} - \varepsilon \triangle u + b \cdot \nabla u + cu = f,$$

where b is independent of u. Such convection-diffusion equations model many fluid flows [Hir88, KL04]; they appear in the well-known Oseen equations and in related subjects like water pollution problems [REI$^+$07], simulation of oil extraction from underground reservoirs [Ewi83], flows in chemical reactors [Alh07] and convective heat transport problems with large Péclet numbers [Jak59].

Of course, convection-diffusion equations do not arise only in fluid flows; the next illustration comes from semiconductor device simulation.

Example 0.2. The "continuity equation" for electrons [PHSM87] in a steady-state scaled model of a one-dimensional semiconductor – with several simplifying assumptions – is

$$\frac{d^2 n}{dx^2} - \frac{d}{dx}\left[n\frac{d}{dx}(\psi + \log n)\right] = 0, \qquad (0.5)$$

where the unknown function n is the electron concentration, and ψ (which is computed from another part of the model) is the electrostatic potential. Now $d\psi/dx$ is typically very large (perhaps 10^5) on part of its domain (see [PHSM87, Figure 2]), so the unit coefficient of the diffusion term $d^2 n/dx^2$ will be dominated there by the convection term coefficient. That is, equation (0.5) is a convection-diffusion problem. ♣

Singularly perturbed differential equations appear in several branches of applied mathematics. (We have seen only two examples, albeit significant ones.) The analysis and numerical solution of convection-diffusion problems deservedly attracts substantial attention.

In this book, we discuss the nature of solutions of various singularly perturbed differential equations before presenting methods for their numerical solution. Thus Part I begins with an exposition of the technique of matched asymptotic expansions, which is then used to examine various classes of two-point boundary value problems. In Part II we move on to time-dependent problems with one space dimension. Elliptic and parabolic problems in several space dimensions come in Part III. Finally, Part IV discusses finite element methods for a significant applied model: the Navier-Stokes equations.

If any discretization technique is applied to a parameter-dependent problem, then the behaviour of the discretization depends on the parameter. For singularly perturbed problems, conventional techniques often lead to discretizations that are worthless if the singular perturbation parameter is close to some critical value. We are interested in *robust* methods that work for all values of the singular perturbation parameter. We therefore track carefully the dependence on this parameter of those constants that arise in consistency, stability and error estimates. Thus the philosophy of this book emphasizes *realistic error estimates*. This contrasts sharply with much published research whose analysis ignores the effect of parameter dependence. There is a growing awareness of the dangers of this neglect; in the particular case of the incompressible Navier-Stokes equations, Johnson, Rannacher and Boman [JRB95a] observe that existing analyses often contain constants that depend on $\exp(\text{Re})$, where Re is the Reynolds number, and conclude that "in the majority of cases of interest, the existing error analysis has no meaning". We hope that the careful approach that is followed here will provide a serviceable foundation for future work.

Discretization leads to a linear or nonlinear system of equations with a large number of unknowns. Iterative methods are commonly used to solve

these systems. It is important to realize that these solvers, like the underlying discretization, should be robust with respect to the singular perturbation parameter. The discretization of a convection-diffusion problem usually produces a nonsymmetric system of equations and this asymmetry complicates the linear algebra analysis. No attempt is made in this book to discuss these issues; instead the recent textbook of Elman, Silvester and Wathen [ESW05] is recommended.

In general standard notation is used for function spaces, norms, etc. (see the notation list on page XIII), but two special conventions should be noted. First, the unknown u in a singular perturbation problem depends, of course, on the perturbation parameter ε. While one must always bear this dependence in mind, it is not included in our notation; that is, we write $u(x)$ instead of, for instance, $u(x, \varepsilon)$ or $u_\varepsilon(x)$. This simplifies the notation, especially when the discretization requires the use of some indices that depend on the mesh. On the other hand, an expression like $\lim_{\varepsilon \to 0} u(x)$ then looks odd, but one should remember that the unknown u does depend on ε. Every notation has its advantages and disadvantages! Second, in our analysis it is important to declare whether or not each constant depends on ε. Thus we denote by C (sometimes subscripted or superscripted) a *generic constant* that is always *independent of the perturbation parameter and of any mesh used*. Other letters are used to denote other "constants" when such a dependence is present.

The following example illustrates our system of numbering and internal cross-referencing. In Part I, Theorem 1.4 lies in Chapter 1 (hence the numbering "1.∗"). In Part I it is referred to as "Theorem 1.4", but we call it "Theorem I.1.4" when it's referred to from outside Part I. A similar convention is used for equations, Lemmas, etc.

We assume that the reader is familiar with the basic theory of ordinary and partial differential equations, and with the jargon and usage of finite difference and finite element methods.

Finally, despite our best efforts, mistakes are undoubtedly present in this book. We invite each reader to email us [rst-book@ovgu.de] any corrections that s/he notices, and this information will be made publicly available at the website [www.rst-book.ovgu.de].

Part I

Ordinary Differential Equations

Part I of this book deals with singularly perturbed two-point boundary value problems. This field of research is of interest in its own right and also serves as an introduction to the more complicated problems posed in higher dimensions that we shall meet later in Parts II, III and IV. An initial discussion of analytical techniques such as maximum principles, asymptotic expansions and stability estimates for the solution of the boundary value problem provides the background needed for the numerical analysis of these ordinary differential equations. Then finite difference, finite element and finite volume methods are presented and analysed, error estimates are derived in various norms, and the relevance of mesh selection is examined. The material here is explained in detail in order to lead the reader gently into this fascinating world.

1

The Analytical Behaviour of Solutions

We begin with a general form of the problem that will occupy our attention throughout most of Part I. Consider the linear two-point boundary value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{for } x \in (d, e),$$

with the boundary conditions

$$\alpha_d u(d) - \beta_d u'(d) = \gamma_d,$$
$$\alpha_e u(e) - \beta_e u'(e) = \gamma_e.$$

Assume that the functions b, c and f are continuous. The constants α_d, α_e, β_d, β_e, γ_d and γ_e are given, and the parameter ε satisfies $0 < \varepsilon \ll 1$.

In general, one can assume homogeneous boundary conditions $\gamma_d = \gamma_e = 0$ by subtracting from u a smooth function ψ that satisfies the original boundary conditions. For example, given Dirichlet boundary conditions $u(d) = \gamma_d$ and $u(e) = \gamma_e$, take

$$\psi(x) = \gamma_d \frac{x - e}{d - e} + \gamma_e \frac{x - d}{e - d}$$

and set $u^*(x) = u(x) - \psi(x)$. Then u^* is the solution of a differential equation of the same type but with homogeneous boundary conditions.

One can also assume without loss of generality that $x \in [0, 1]$ by means of the linear transformation

$$x \mapsto \frac{x - d}{e - d}.$$

The analytical behaviour of the solution of a singularly perturbed boundary value problem depends on the nature of the boundary conditions. From the numerical analyst's point of view, the most difficult case is when these conditions are Dirichlet. We consequently pay scant attention to other boundary conditions. Thus Sections 1.1 and 1.2 investigate the singularly perturbed problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{for } x \in (0,1), \qquad (1.1a)$$

$$u(0) = u(1) = 0, \quad \text{with } c(x) \geq 0 \quad \text{for } x \in [0,1], \qquad (1.1b)$$

under the conditions on ε, b, c and f stated earlier. This is a typical *convection-diffusion problem* (see the Introduction) because in general we assume that b is not identically zero.

We begin our study by stating three closely-related properties of differential operators $M : C^2(0,1) \to C(0,1)$. Let $w \in C^2(0,1) \cap C[0,1]$. Then M is said to be *inverse-monotone* if the inequalities

$$Mw(x) \geq 0 \quad \text{for all } x \in (0,1), \quad w(0) \geq 0, \quad w(1) \geq 0$$

together imply that $w(x) \geq 0$ for all $x \in [0,1]$. To see that the operator L of (1.1) is inverse-monotone, one argues by contradiction [GT83].

We say that M satisfies a *maximum principle* if $Mu(x) = 0$ for all $x \in (0,1)$ implies that

$$\min\{u(0), u(1), 0\} \leq u(x) \leq \max\{u(0), u(1), 0\} \quad \text{for all } x \in [0,1].$$

Inverse-monotonicity implies that L satisfies a maximum principle. It also implies that L satisfies the following *comparison principle* which for our purposes is the most useful of the three properties.

Lemma 1.1 (Comparison principle). *Let $v, w \in C^2(0,1) \cap C[0,1]$ satisfy*

$$Lw(x) \geq Lv(x) \quad \text{for all } x \in (0,1)$$

and $w(0) \geq v(0)$, $w(1) \geq v(1)$. Then

$$w(x) \geq v(x) \quad \text{for all } x \in [0,1].$$

We then say that w is a *barrier function* for v. A fairly complete discussion of maximum and comparison principles for second-order elliptic problems can be found in [GT83]. Unfortunately the terminology in the literature is inconsistent, in the sense that each of the three properties above is sometimes called a maximum principle.

Lemma 1.1 implies immediately the uniqueness of classical solutions of the boundary value problem (1.1). In this one-dimensional case, the existence of a classical solution follows. The condition $c \geq 0$ cannot in general be discarded, as is evident from the problem

$$-\varepsilon u'' + \lambda u = 0 \quad \text{on } (0,1), \quad u(0) = u(1) = 0,$$

which has multiple solutions when $\lambda = -\varepsilon k^2 \pi^2, \ k = 1, 2, \ldots$.

1.1 Linear Second-Order Problems Without Turning Points

Existence and uniqueness of the classical solution u of (1.1) are now guaranteed, but the behaviour of u when ε is small is still obscure. To gain an initial insight into the structure of u when ε is near zero, we study a simple example.

Example 1.2. The boundary value problem

$$-\varepsilon u'' + u' = 1 \quad \text{on } (0,1), \quad u(0) = u(1) = 0,$$

has the solution

$$u(x) = x - \frac{\exp(-\frac{1-x}{\varepsilon}) - \exp(-\frac{1}{\varepsilon})}{1 - \exp(-\frac{1}{\varepsilon})}.$$

Hence, for $a \in [0,1)$,

$$\lim_{x \to a} \lim_{\varepsilon \to 0} u(x) = a = \lim_{\varepsilon \to 0} \lim_{x \to a} u(x),$$

but

$$1 = \lim_{x \to 1} \lim_{\varepsilon \to 0} u(x) \neq \lim_{\varepsilon \to 0} \lim_{x \to 1} u(x) = 0.$$

The presence of a point ($x = 1$ in this example) where such an inequality appears means that the problem is *singularly perturbed*. The inequality implies that the solution $u(x)$ changes abruptly as x approaches 1 – we say that there is a *boundary layer* at $x = 1$. See Figure 1.1. ♣
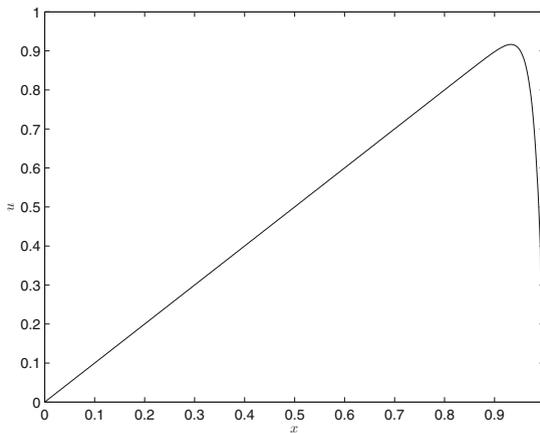
Fig. 1.1. Solution of Example 1.2 with a boundary layer at $x = 1$

1.1.1 Asymptotic Expansions

Can we approximate the solution u of (1.1) by a simple known function? Yes, by means of a standard technique in singular perturbation theory called the *method of matched asymptotic expansions*; see, for instance, [Eck73, O'M91]. The function u_{as} constructed by this technique is an *asymptotic expansion* of u; it illuminates the nature of u and thus is valuable information.

The function u_{as} is an *asymptotic expansion* of order m of u (in the maximum norm) if there is a constant C such that

$$|u(x) - u_{as}(x)| \leq C\varepsilon^{m+1} \quad \text{for all } x \in [0,1] \text{ and all } \varepsilon \text{ sufficiently small.}$$

Here we remind the reader that throughout the book C denotes a generic constant that is independent of ε. In the construction of u_{as} for (1.1), we assume that b, c and f are sufficiently smooth on [0,1].

The first step is to try to find a *global expansion* (or *regular expansion* or *outer expansion*) u_g. This function will be a good approximation of u away from any layer(s), i.e., on nearly all of the domain $[0,1]$. We set

$$u_g(x) = \sum_{\nu=0}^{m} \varepsilon^{\nu} u_{\nu}(x), \tag{1.2}$$

where the $u_{\nu}(x)$ are yet to be determined. (Here, as for regular perturbations, we try to expand the solution in a Taylor-type series.) Define the operator L_0 by formally setting $\varepsilon = 0$ in L, viz.,

$$L_0 v := bv' + cv.$$

Substituting u_g into (1.1) and equating coefficients of like powers of ε yields

$$L_0 u_0 = f,$$
$$L_0 u_{\nu} = u_{\nu-1}'' \quad \text{for} \quad \nu = 1, ..., m.$$

If $b(x)$ has any zero in the interval $[0,1]$, this causes difficulty in defining the coefficients u_{ν} of the global expansion because the operator L_0 then becomes singular. Zeros of b are called *turning points*. We exclude such phenomena here and defer their examination to Section 1.2.

Suppose that $b(x) \neq 0$ for all $x \in [0,1]$. Then in principle one can calculate $u_0, u_1, \ldots, u_m$ explicitly, provided that there is some additional condition on each of these functions that ensures its uniqueness. One of the boundary conditions in (1.1b) should be used to define u_0, and the crucial question is: which boundary condition should we discard? Guided by Example 1.2, we state the following *cancellation law*, which specifies the boundary condition to discard (see Section 1.4.1 for a more general formulation):

- If $b > 0$ then the boundary layer is located at $x = 1$ and to define u_0 one omits the boundary condition at $x = 1$. If $b < 0$ then the boundary layer is located at $x = 0$ and the boundary condition at $x = 0$ is dropped.

The transformation $x \mapsto 1 - x$ reduces the case $b < 0$ to $b > 0$; thus it suffices to study the case $b > 0$ in detail. The coefficients in the global expansion u_g are defined by

$$L_0 u_0 = f, \qquad u_0(0) = 0, \tag{1.3a}$$
$$L_0 u_\nu = u''_{\nu-1}, \qquad u_\nu(0) = 0 \quad \text{for } \nu = 1, ..., m. \tag{1.3b}$$

We call equation (1.3a) the *reduced problem* and u_0 is the *reduced solution*. The condition $u_0(0) = 0$ comes from (1.1b), while the conditions $u_\nu(0) = 0$ for $\nu \geq 1$ ensure that $u_g(0) = u(0)$.

The aim of the method of matched asymptotic expansions is to construct an approximation of u that is valid for all $x \in [0, 1]$. But u_g cannot be such an approximation since it fails to satisfy the boundary condition at $x = 1$. Therefore one adds a local correction to u_g near $x = 1$. First, observe that the difference $w := u - u_g$ satisfies

$$Lw = \varepsilon^{m+1} u''_m,$$

$$w(0) = 0, \quad w(1) = -\sum_{\nu=0}^{m} \varepsilon^\nu u_\nu(1).$$

Write $L = \varepsilon L_1 + L_0$. Recalling that a local correction is needed near $x = 1$, where the solution u has a boundary layer, we stretch the scale there in the x direction by introducing the local variable

$$\xi = \frac{1-x}{\delta}, \quad \text{where } \delta > 0 \text{ is small and yet to be specified.}$$

One chooses δ such that L_0 and εL_1 have formally the same order with respect to ε after the independent variable is transformed from x to ξ. That is, since $b \neq 0$, one sets

$$\varepsilon \delta^{-2} \approx \delta^{-1}.$$

This leads to the choice $\delta = \varepsilon$.

In terms of the new variable ξ, use Taylor expansions to write

$$b(1 - \varepsilon \xi) = \sum_{\nu=0}^{\infty} b_\nu \varepsilon^\nu \xi^\nu \quad \text{with } b_0 = b(1),$$

$$c(1 - \varepsilon \xi) = \sum_{\nu=0}^{\infty} c_\nu \varepsilon^\nu \xi^\nu \quad \text{with } c_0 = c(1).$$

Consequently, for any sufficiently differentiable function g, we can express L in terms of ξ as

$$\varepsilon L_1 g + L_0 g = \frac{1}{\varepsilon} \sum_{\nu=0}^{\infty} \varepsilon^\nu L_\nu^* g,$$

with

$$L_0^* := -\frac{d^2}{d\xi^2} - b_0\frac{d}{d\xi},$$

$$L_1^* := \qquad -b_1\xi\frac{d}{d\xi} + c_0,$$

etc. Now introduce the local expansion

$$v_{loc}(\xi) = \sum_{\mu=0}^{m+1} \varepsilon^\mu v_\mu(\xi). \qquad (1.4)$$

In order that v_{loc} approximates $w = u - u_g$, the local corrections v_μ should satisfy the *boundary layer equations*

$$L_0^* v_0 = 0, \qquad (1.5a)$$

$$L_0^* v_\mu = -\sum_{\kappa=1}^{\mu} L_\kappa^* v_{\mu-\kappa}, \quad \text{for } \mu = 1, ..., m+1. \qquad (1.5b)$$

To obtain the correct boundary condition at $x = 1$, one takes $v_\kappa(0) = -u_\kappa(1)$ for $\kappa = 0, 1, \ldots, m$. As the differential equations (1.5) are of second order, a further boundary condition is also needed. To ensure the local character of the local correction, one requires that $\lim_{\xi\to\infty} v_\mu(\xi) = 0$. With these two boundary conditions the problem (1.5) has a unique solution, because the characteristic equation corresponding to L_0^* (which is a differential operator with *constant* coefficients) is

$$-\lambda^2 - b(1)\lambda = 0,$$

which has exactly one negative root. For example, the first-order correction is

$$v_0(\xi) = -u_0(1)e^{-b(1)\xi}.$$

Remark 1.3. A critical question in this method is whether or not the equations (1.5) for the local correction possess a number of decaying solutions that is equal to the number of boundary conditions that are not satisfied by the global approximation. If one cancels the wrong boundary condition when defining the reduced problem, this can lead to boundary layer equations without decaying solutions and the method then fails. ♣

Boundary layers are classified according to the nature of the boundary layer equations. The simplest layers are *exponential boundary layers* (which are sometimes called *ordinary boundary layers*), where the solutions of the boundary layer equations are decaying exponential functions. The solution of (1.1) usually has a layer of this type at $x = 1$ when $b > 0$ on $[0, 1]$.

Theorem 1.4. *If the coefficients and the right-hand side of the boundary value problem (1.1) are sufficiently smooth and $b(x) > \beta > 0$ on $[0, 1]$, then its solution u has a matched asymptotic expansion of the form*

$$u_{as}(x) = \sum_{\nu=0}^{m} \varepsilon^{\nu} u_{\nu}(x) + \sum_{\mu=0}^{m} \varepsilon^{\mu} v_{\mu}\left(\frac{1-x}{\varepsilon}\right), \tag{1.6}$$

such that for any sufficiently small fixed constant ε_0 one has

$$|u(x) - u_{as}(x)| \leq C\varepsilon^{m+1} \quad \text{for } x \in [0,1] \text{ and } \varepsilon \leq \varepsilon_0.$$

Here C is independent of x and ε.

Proof. Consider

$$u_{as}^*(x) := \sum_{\nu=0}^{m} \varepsilon^{\nu} u_{\nu}(x) + \sum_{\mu=0}^{m+1} \varepsilon^{\mu} v_{\mu}\left(\frac{1-x}{\varepsilon}\right),$$

which has an additional term for $\mu = m+1$ compared with (1.6). (This is a standard trick: if the transformed problem in the local variables has a leading term that is $O(\varepsilon^{-l})$, one considers $\sum_{\mu=0}^{m+l}$.) Our construction of the u_{ν} and v_{μ} yields

$$L(u - u_{as}^*) = O(\varepsilon^{m+1}),$$
$$(u - u_{as}^*)(0) = O(\varepsilon^{\kappa}), \quad (u - u_{as}^*)(1) = O(\varepsilon^{m+1}),$$

where $\kappa > 0$ is arbitrary. Now apply the comparison principle of Lemma 1.1, with the barrier function $w(x) = C\varepsilon^{m+1}(1+x)$ – this choice of w exploits the property $b \geq b_0 > 0$. We get

$$|(u - u_{as}^*)(x)| \leq |w(x)| \leq C\varepsilon^{m+1} \quad \text{for all } x \in [0,1].$$

But $|u_{as}(x) - u_{as}^*(x)| = |\varepsilon^{m+1} v_{m+1}((1-x)/\varepsilon)| \leq C\varepsilon^{m+1}$, so a triangle inequality completes the argument. $\square$

A formal differentiation of (1.6) leads to the following conjecture:
If b, c and f are sufficiently smooth and $b > 0$ (so turning points are excluded), the solution u of the boundary value problem (1.1) satisfies

$$|u^{(i)}(x)| \leq C\left[1 + \varepsilon^{-i} \exp\left(-b(1)\frac{1-x}{\varepsilon}\right)\right].$$

A rigorous proof of the validity of this differentiation is possible [O'M91], but it is not simple. In Section 1.1.3 we shall prove a similar bound on $u^{(i)}(x)$ without using an asymptotic expansion.

Remark 1.5. (Effect of boundary conditions on the layer) In the case $b > 0$, suppose that the boundary conditions in (1.1b) are replaced by

$$u(0) = 0, \quad u'(1) = 0.$$

Then the method of matched asymptotic expansions yields a local correction of the type

$$v_{loc}(\xi) = \varepsilon \sum_{\mu=0}^{m} \varepsilon^{\mu} v_{\mu}(\xi)$$

because, for example,

$$-\frac{\varepsilon}{b(1)} u_0'(1) e^{-b(1)\xi}$$

corrects the boundary condition at $x = 1$. One can show that:
A Dirichlet boundary condition at $x = 1$ causes a boundary layer there with

$$u'(1) = O(\varepsilon^{-1}) \quad \text{as } \varepsilon \to 0,$$

but a Neumann boundary condition at $x = 1$ causes a less severe boundary layer, since then

$$u'(1) = O(1) \text{ and } u''(1) = O(\varepsilon^{-1}) \quad \text{as } \varepsilon \to 0.$$

For example, the exact solution of

$$-\varepsilon u'' + u' = 1, \quad u(0) = 0 \text{ and } u'(1) = 0$$

is $u(x) = x - \varepsilon[e^{-(1-x)/\varepsilon} - e^{-1/\varepsilon}]$.
   Under special circumstances, a different weakening of the boundary layer can occur. If, for example, the boundary condition at $x = 1$ were

$$b(1)u'(1) + c(1)u(1) = f(1)$$

– which is satisfied by the reduced solution u_0 of (1.3a) – then the asymptotic expansion of u starts with $u_0 + \varepsilon u_1 + \varepsilon^2 v_2$ because one can choose $v_0 \equiv v_1 \equiv 0$. In this particular case one has

$$u''(1) = 0 \text{ and } u'''(1) = O(\varepsilon^{-1}) \quad \text{as } \varepsilon \to 0,$$

while $u(x), u'(x)$ and $u''(x)$ are all bounded uniformly on $[0, 1]$ as $\varepsilon \to 0$.   ♣

1.1.2 The Green's Function and Stability Estimates

Assume that $b(x) \geq \beta > 0$ on $[0, 1]$. The comparison principle of Lemma 1.1 provides a simple proof of the *stability estimate*

$$\|v\|_{\infty} \leq C \|Lv\|_{\infty} \quad \text{for all } v \in C^2[0, 1] \text{ with } v(0) = v(1) = 0, \qquad (1.7)$$

where

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

To prove (1.7), use $w(x) = \|Lv\|_{\infty}(1 + x)/\beta$ as a barrier function for v.

Note that the stability constant C in (1.7) is independent of ε. When applied to the solution u of (1.1), inequality (1.7) yields

$$\|u\|_\infty \le C\|f\|_\infty.$$

This is typical: a stability inequality implies an *a priori estimate* for the exact solution. This a priori estimate tells us that u is bounded, uniformly with respect to ε, in the maximum norm.

For the analysis of numerical methods, especially on non-equidistant meshes and in the context of a posteriori error estimates, it is very useful to have stronger stability results that use other norms. Let $(A, \|\cdot\|_A)$ and $(B, \|\cdot\|_B)$ be normed linear spaces with $M : A \to B$. Then M is said to be *uniformly (A, B)-stable* if

$$\|v\|_A \le C\,\|Mv\|_B \quad \text{for all } v \in A \tag{1.8}$$

with a stability constant C that is independent of ε. If $A = B$, we say simply that M is A-stable.

In this section we shall derive stability results for the convection-diffusion problem (1.1) under the hypotheses that b is continuous and does not vanish in $[0, 1]$. The (L_∞, L_1) stability result (1.19) comes from [Gar89], while the negative norm stability estimate (1.20) is in [And01, Kop01b]. We follow the presentation of [Lin02a].

Consider the boundary value problem (1.1):

$$Lu := -\varepsilon u'' + bu' + cu = f,$$
$$u(0) = u(1) = 0,$$

where $b \ge \beta > 0$. Additionally, to simplify certain arguments, assume that

$$c \ge 0 \qquad \text{and} \quad c - b' \ge 0. \tag{1.9}$$

Remark 1.6. Because $b > 0$ the conditions (1.9) can always be guaranteed for ε smaller than some threshold value ε_0 by making a change of variable $u(x) = \hat{u}(x)\exp(kx)$ with the constant k chosen appropriately. ♣

The standard *Green's function* $G(x, \xi)$ associated with L and homogeneous Dirichlet boundary conditions is for each fixed $\xi \in [0, 1]$ the solution of

$$(L\,G(\cdot, \xi))(x) = \delta(x - \xi) \text{ for } x \in (0, 1), \quad G(0, \xi) = G(1, \xi) = 0, \tag{1.10}$$

where δ is the Dirac-δ distribution. Equivalently, to avoid introducing distributions, for fixed ξ one seeks a classical solution in $C^2((0, 1) \setminus \{\xi\}) \cap C[0, 1]$ that satisfies

$$(L\,G(\cdot, \xi))(x) = 0 \text{ for } x \in (0, 1) \setminus \{\xi\}, \quad G(0, \xi) = G(1, \xi) = 0, \tag{1.11}$$

and the jump condition

$$-\varepsilon[G(\cdot,\xi)'](\xi) = 1,$$

where the notation $[v](d) := v(d+0) - v(d-0)$ denotes the jump of a discontinuous function $v(x)$ at $x = d$.

In terms of the adjoint operator $L^*v := -\varepsilon v'' - (bv)' + cv$, for fixed x the Green's function $G(x,\xi)$ satisfies

$$(L^*G(x,\cdot))(\xi) = \delta(\xi - x) \text{ for } \xi \in (0,1), \quad G(x,0) = G(x,1) = 0. \quad (1.12)$$

To derive stability estimates we shall use the solution representation

$$v(x) = \int_0^1 G(x,\xi)(L\,v)(\xi)\,d\xi \qquad (1.13)$$

which is valid for all v satisfying $v(0) = v(1) = 0$. Thus some bounds on G are needed.

Similarly to the classical comparison principle of Lemma 1.1, one has: if the functions v and w in $C^2((0,1) \setminus \{\xi\}) \cap C[0,1]$ satisfy

$$v(0) \le w(0),$$
$$v(1) \le w(1),$$
$$\mathcal{L}v(x) \le \mathcal{L}w(x) \quad \text{in } (0,1) \setminus \{\xi\},$$
$$-\varepsilon[v'](\xi) \le -\varepsilon[w'](\xi),$$

then $v(x) \le w(x)$ for all $x \in [0,1]$. This piecewise comparison principle can be found in [Mey98]; it is well known in the field of enclosing discretization methods but is rarely stated explicitly in the literature. Using the comparison principle with the barrier functions $\hat{G}_1 \equiv 0$ and

$$\hat{G}_2 = \begin{cases} (1/\beta)\exp(-\beta(\xi - x)/\varepsilon) & \text{for } 0 \le x \le \xi, \\ 1/\beta & \text{for } \xi \le x \le 1, \end{cases}$$

we get the following bounds for the Green's function:

$$0 \le G(x,\xi) \le \frac{1}{\beta} \quad \text{for } (x,\xi) \in [0,1] \times [0,1]. \qquad (1.14)$$

The representation (1.13) then implies that for any function $v \in W^{2,1}(0,1)$ with $v(0) = v(1) = 0$, the stability estimate (1.7) has been sharpened to the (L_∞, L_1) estimate

$$\|v\|_\infty \le \frac{1}{\beta}\|L\,v\|_{L_1} \text{ for } v \in W_0^{1,1}(0,1) \cap W^{2,1}(0,1).$$

Here we used the notation $W^{m,p}(0,1)$ for the Sobolev space of functions defined on $[0,1]$ whose derivatives of order m are in L_p. Functions in $W_0^{1,p}(0,1)$ vanish at $x = 0$ and $x = 1$. See [Ada78] for a thorough discussion of Sobolev spaces.

We want to go one step further. For each $v \in W_0^{1,\infty}(0,1)$ let the auxiliary function $V \in L_\infty(0,1)$ satisfy $V' = Lv$. Then an integration by parts gives

$$v(x) = -\int_0^1 G_\xi(x,\xi)V(\xi)d\xi \qquad (1.15)$$

and

$$v'(x) = -\int_0^1 G_{x\xi}(x,\xi)V(\xi)d\xi. \qquad (1.16)$$

These formulas are well defined: piecewise existence of $G_{x\xi}$ follows from explicit representations of G in [And01] or, alternatively, from the piecewise existence of G_{xx} and $G_{\xi\xi}$.

To extract the desired stability estimates from these representations, we need more information about the Green's function.

Since $G \geq 0$ and G satisfies the boundary conditions of (1.12), one has $G_\xi(x,0) \geq 0$ and $G_\xi(x,1) \leq 0$. Rearranging (1.12) shows that $v(\cdot) := G_\xi(x,\cdot)$ satisfies

$$\varepsilon v_\xi + bv = (c - b_\xi)G \geq 0 \quad \text{for } \xi \in (0,x) \qquad (1.17)$$

where we used (1.9). As $v(0) \geq 0$, an integration of (1.17) yields $v \geq 0$ on $[0,x]$, so $G(x,\cdot)$ increases monotonically on $[0,x]$. Integrating (1.12) over $[\xi,1]$ with $\xi > x$ gives

$$\varepsilon G_\xi(x,\xi) - \varepsilon G_\xi(x,1) + b(\xi)G(x,\xi) - b(1)G(x,1) = -\int_\xi^1 c(s)G(x,s)ds.$$

Hence, using $c \geq 0$ from (1.9),

$$\varepsilon G_\xi(x,\xi) \leq \varepsilon G_\xi(x,1) - b(\xi)G(x,\xi) \leq 0.$$

Thus $G(x,\cdot)$ decreases monotonically on $[x,1]$.

One can prove similarly that $G_x(x,\xi) \geq 0$ for $0 \leq x < \xi \leq 1$ and $G_x(x,\xi) \leq 0$ for $0 \leq \xi < x \leq 1$. Consequently

$$G_{x\xi}(x,0) \leq 0 \quad \text{and} \quad G_{x\xi}(x,1) \leq 0 \quad \text{for } x \in (0,1).$$

For $\xi < x$ we see that $w = G_{x\xi}(x,\cdot)$ satisfies

$$\varepsilon w_\xi + bw = (c - b_\xi)G_x \leq 0.$$

It now follows from $w(0) \leq 0$ that $G_{x\xi} \leq 0$ for $\xi < x$. For $\xi > x$, differentiate the above identity:

$$\varepsilon G_{x\xi}(x,\xi) - \varepsilon G_{x\xi}(x,1) + b(\xi)G_x(x,\xi) - b(1)G_x(x,1) = -\int_\xi^1 c(s)G_x(x,s)ds.$$

This gives $G_{x\xi} \leq 0$ for $\xi > x$.

The next step is to bound the L_1 norms of G_ξ and $G_{x\xi}$ using the above monotonicity properties and the L_∞ bound (1.14). First, we get

$$\|G_\xi(x,\cdot)\|_{L_1} = \int_0^x G_\xi(x,\xi)d\xi - \int_x^1 G_\xi(x,\xi)d\xi = 2G(x,x) \leq \frac{2}{\beta}. \qquad (1.18)$$

A related argument shows that

$$\|G_{x\xi}(x,\cdot)\|_{L_1} = \frac{2}{\varepsilon},$$

on taking account of the singularity caused by $G_x(x,x+0)-G_x(x,x-0) = 1/\varepsilon$. These bounds can be combined with (1.15) and (1.16) to produce new stability estimates. In summary, introducing the norm

$$\|v\|_* := \inf_{V:V'=v} \|V\|_\infty,$$

the stability results we have proved in this section are the following:

Theorem 1.7. *The operator L satisfies the stability estimates*

$$\|v\|_\infty \leq \frac{1}{\beta}\|L\,v\|_{L_1} \text{ for } v \in W_0^{1,1}(0,1) \cap W^{2,1}(0,1) \qquad (1.19)$$

and

$$\frac{\beta}{2}\|v\|_\infty + \frac{\varepsilon}{2}\|v'\|_\infty \leq \|L\,v\|_* \quad \text{ for } v \in W_0^{1,\infty}(0,1). \qquad (1.20)$$

The space $W^{-1,\infty} = (W_0^{1,1})'$ is isometrically isomorphic to the space of distributions generated by integrals of L_∞ functions and equipped with the norm $\|\cdot\|_*$; see [Ada78, Theorem 3.10]. In this sense, the norm $\|\cdot\|_*$ is the $W^{-1,\infty}$-norm and we say that (1.20) is a negative-norm stability estimate.

Now $L_1[0,1] \subset W^{-1,\infty} = (W_0^{1,1})'$. Andreev [And01, Lemma 2.6] observed that

$$\|f\|_{-1,\infty} = \sup_{1=\|v\|_{W_0^{1,1}}} \left| \int_0^1 fv\,dx \right| = \inf_C \left\| \int^1 f(s)\,ds + C \right\|_\infty = \|f\|_*.$$

Note that since

$$\|v\|_* \leq \|v\|_{L_1} \leq \|v\|_\infty,$$

the negative-norm bound is the strongest of our stability results.

In [And01] an assumption of the type (1.9) was not used, which makes the analysis more difficult; this paper begins with a differential equation in conservation form (assuming a different sign for the convective term)

$$\mathcal{L}v := -\varepsilon v'' - (bv)' + cv,$$

then goes on to the more complicated case where the equation is not in conservation form.

1.1.3 A Priori Estimates for Derivatives and Solution Decomposition

The numerical analysis of discretization methods requires information about higher-order derivatives of u, the solution of (1.1). Theorem 1.7 tells us that

$$|u^{(k)}(x)| \le C\varepsilon^{-k} \quad \text{for } x \in [0,1], \ k = 0,1.$$

Hence, by repeated differentiation of the differential equation (1.1a), we obtain

$$|u^{(k)}(x)| \le C\varepsilon^{-k} \quad \text{for } x \in [0,1], \ k = 0,1,\dots,q,$$

where q depends on the smoothness of the data.

In general, crude bounds like these are inadequate for the job of analysing discretization methods. We now use the argument of [KT78, Lemma 2.3] to deduce a sharper estimate directly from (1.1); no asymptotic expansion is used.

Lemma 1.8. *Assume that $b(x) > \beta > 0$ and b, c, f are sufficiently smooth. Then for $i = 1, 2, \dots, q$, the solution u of (1.1) satisfies*

$$|u^{(i)}(x)| \le C \left[1 + \varepsilon^{-i} \exp\left(-\beta \frac{1-x}{\varepsilon} \right) \right] \quad \text{for } 0 \le x \le 1,$$

where the maximal order q depends on the smoothness of the data.

Proof. Set $h = f - cu$. Using an integrating factor we integrate $-\varepsilon u'' + bu' = h$ twice, obtaining

$$u(x) = u_p(x) + K_1 + K_2 \int_x^1 \exp[-\varepsilon^{-1}(B(1) - B(t))]\, dt,$$

where

$$u_p(x) := -\int_x^1 z(t)\, dt, \quad z(x) := \int_x^1 \varepsilon^{-1} h(t) \exp[-\varepsilon^{-1}(B(t) - B(x))]\, dt,$$

$$B(x) := \int_0^x b(t)\, dt;$$

here the constants of integration (K_1 and K_2) may depend on ε.

The boundary condition $u(1) = 0$ implies that $K_1 = 0$. One can also see that $u'(1) = -K_2$. Now $u(0) = 0$ gives

$$K_2 \int_0^1 \exp[-\varepsilon^{-1}(B(1) - B(t))]\, dt = -u_p(0). \tag{1.21}$$

The bound $\|u\|_\infty \le C$ implied by (1.7) leads to

$$|z(x)| \leq C\varepsilon^{-1} \int_x^1 \exp[-\varepsilon^{-1}(B(t) - B(x))]\, dt.$$

Applying the inequality

$$\exp[-\varepsilon^{-1}(B(t) - B(x)] \leq \exp[-\beta\varepsilon^{-1}(t - x)] \quad \text{for} \quad x \leq t,$$

we obtain

$$|z(x)| \leq C\varepsilon^{-1} \int_x^1 \exp[-\beta\varepsilon^{-1}(t - x)]\, dt \leq C.$$

Hence $|u_p(0)| \leq C$. Set $\|b\|_\infty = \max_{x \in [0,1]} b(x)$. Then

$$\int_0^1 \exp[-\varepsilon^{-1}(B(1) - B(t)]\, dt) \geq \int_0^1 \exp[-\|b\|_\infty \varepsilon^{-1}(1 - t)]\, dt \geq C\varepsilon.$$

It then follows from (1.21) that $|K_2| \leq C\varepsilon^{-1}$.

Now

$$u'(x) = z(x) - K_2 \exp[-\varepsilon^{-1}(B(1) - B(x))]$$

implies that

$$|u'(x)| \leq C \left[1 + \varepsilon^{-1} \exp\left(-\frac{\beta(1 - x)}{\varepsilon}\right)\right].$$

The bound on $u^{(i)}(x)$ for $i > 1$ follows by induction on i and repeated differentiation of (1.1a). □

A classical asymptotic expansion like that of Theorem 1.4 decomposes the solution u into a smooth part (i.e., a function for which certain low-order derivatives are bounded uniformly in ε), a layer part and a remainder. We now construct a decomposition of u into a sum of a smooth part and a layer part, with no remainder. This type of decomposition is helpful in the analysis of certain numerical methods.

The standard asymptotic expansion of Theorem 1.4 gives

$$u = u_0 + \varepsilon u_1 + \ldots + \varepsilon^k u_k + v_0 + \varepsilon v_1 + \ldots + \varepsilon^k v_k + \varepsilon^{k+1} R,$$

where R satisfies a boundary value problem similar to (1.1). Set

$$S^* := u_0 + \varepsilon u_1 + \ldots + \varepsilon^k u_k + \varepsilon^{k+1} R,$$
$$E^* := v_0 + \varepsilon v_1 + \ldots + \varepsilon^k v_k,$$

The crude estimate $\|R^{(m)}\|_\infty \leq C\varepsilon^{-m}$ yields

$$|S^{*(l)}(x)| \leq C \quad \text{for } l \leq k + 1. \tag{1.22}$$

For the boundary layer functions, the construction of Section 1.1 leads to

$$|E^{*(l)}(x)| \leq C\varepsilon^{-l} \exp\left(-\frac{\beta(1 - x)}{\varepsilon}\right). \tag{1.23}$$

We call a decomposition $u = S^* + E^*$ with the properties (1.22) and (1.23) an *S-type decomposition*.

A minor modification of this construction yields an *S-decomposition*; this splitting of u enjoys the extra property that the layer part lies in the null space of L. Decompositions of this type were introduced by Shishkin in the analysis of difference schemes on piecewise equidistant meshes; see Section 2.4.2. Write

$$u = u_0 + \varepsilon u_1 + \ldots + \varepsilon^k u_k + \varepsilon^{k+1} u_{k+1}^* + v_0 + \varepsilon v_1 + \ldots + \varepsilon^k v_k + \varepsilon^{k+1} v_{k+1}^*,$$

where $u_0, \ldots, u_k, v_0, \ldots, v_k$ are the standard terms of the asymptotic expansion whereas u_{k+1}^* and v_{k+1}^* are defined by

$$Lu_{k+1}^* = u_k'', \quad u_{k+1}^*(0) = u_{k+1}^*(1) = 0$$

and

$$Lv_{k+1}^* = -\varepsilon^{-(k+1)} L(v_0 + \varepsilon v_1 + \ldots + \varepsilon^k v_k),$$
$$v_{k+1}^*(0) = 0, \quad v_{k+1}^*(1) = -(v_0 + \varepsilon v_1 + \ldots + \varepsilon^k v_k)(1).$$

Now set

$$S := u_0 + \varepsilon u_1 + \ldots + \varepsilon^k u_k + \varepsilon^{k+1} u_{k+1}^*,$$
$$E := v_0 + \varepsilon v_1 + \ldots + \varepsilon^k v_k + \varepsilon^{k+1} v_{k+1}^*,$$

and putting $q = k + 1$ we obtain

Lemma 1.9. (S-decomposition) *Let q be some positive integer. Consider the boundary value problem (1.1) with $b(x) > \beta > 0$ and sufficiently smooth data. Its solution u can be decomposed as $u = S + E$, where the smooth part S satisfies $LS = f$ and*

$$|S^{(l)}(x)| \leq C \quad \text{for } 0 \leq l \leq q,$$

while the layer part E satisfies $LE = 0$ and

$$|E^{(l)}(x)| \leq C\varepsilon^{-l} \exp\left(-\frac{\beta(1-x)}{\varepsilon}\right) \quad \text{for } 0 \leq l \leq q.$$

Clearly Lemma 1.9 implies the bounds of Lemma 1.8. Conversely, the S-decomposition of Lemma 1.9 can in fact be deduced from Lemma 1.8, as we now show. Assume the bounds of Lemma 1.8. Let $x^* = 1 - (q\varepsilon/\beta)\ln 1/\varepsilon$. Set $S(x) = u(x)$ in $[0, x^*]$. Then Lemma 1.8 implies that

$$|S^{(l)}(x)| \leq C \quad \text{on } [0, x^*] \text{ for } 0 \leq l \leq q$$

since $e^{-\beta(1-x^*)/\varepsilon} = \varepsilon^q$. Thus one can extend the definition of S to all of $[0, 1]$ with $|S^{(l)}(x)| \leq 2C$ on $[0, 1]$ for $0 \leq l \leq q$.

Now consider $E := u - S$. Then $E \equiv 0$ in $[0, x^*]$, while in $(x^*, 1]$ one has

$$|E^{(q)}(x)| \leq |u^{(q)}(x)| + |S^{(q)}(x)| \leq C\left(1 + \varepsilon^{-q}e^{-\beta(1-x)/\varepsilon}\right) \leq C\varepsilon^{-q}e^{-\beta(1-x)/\varepsilon}.$$

Integrating $E^{(k)}$ for $k = q, q-1, \ldots, 1$, we get inductively

$$|E^{(k-1)}(x)| = \left|\int_{x^*}^{x} E^{(k)}(s)ds\right|$$

$$\leq C\int_{x^*}^{x} \varepsilon^{-k}e^{-\beta(1-s)/\varepsilon}ds \leq C\varepsilon^{-(k-1)}e^{-\beta(1-x)/\varepsilon}.$$

Thus $S + E$ is an S-decomposition of u.

In [Lin02b] Linß shows how to construct an S-decomposition under minimal regularity hypotheses.

Remark 1.10. (Reaction-Diffusion Problems) Consider the reaction-diffusion problem

$$-\varepsilon u'' + c(x)u = f(x) \quad \text{on } (0,1)$$

with Dirichlet boundary conditions. Assume that $c > \gamma > 0$ on $[0,1]$. Then in general the solution u contains exponential boundary layers of the form $\exp(-\sqrt{\gamma}x/\sqrt{\varepsilon})$ and $\exp(-\sqrt{\gamma}(1-x)/\sqrt{\varepsilon})$; note that these layers depend on $\sqrt{\varepsilon}$ and are present at both $x = 0$ and $x = 1$. An S-decomposition of u can be found in [MOS96, Chapter 6].

The stability properties of the reaction-diffusion operator are very different from those of the convection-diffusion operator. For instance, the Green's function of the reaction-diffusion problem with homogeneous Dirichlet conditions satisfies

$$\|G\|_\infty \leq \frac{C}{\sqrt{\varepsilon}}$$

and is not bounded as $\varepsilon \to 0$.   ♣

Remark 1.11. (Two-parameter convection-diffusion-reaction problems) Consider the two-parameter problem

$$-\varepsilon_1 u'' + \varepsilon_2 b(x)u' + c(x)u = f(x)$$

where ε_1 and ε_2 are small positive parameters, $b > 0$ and $c > 0$. It is shown in [LR04] that the nature of the solution decomposition depends on the relative sizes of ε_1 and ε_2. The associated Green's function satisfies

$$\|G\|_\infty \leq \frac{C}{\sqrt{\varepsilon_1 + \varepsilon_2^2}};$$

see [RU03].   ♣

1.2 Linear Second-Order Turning-Point Problems

In second-order singularly perturbed differential equations, isolated points where the coefficient of u' vanishes are called *turning points*. We first look at the case of a single turning point in the interior of the domain. For convenience, the differential equation is posed on $(-1, 1)$ with its turning point placed at $x = 0$. That is, we consider

$$Lu := -\varepsilon u'' + xb(x)u' + c(x)u = f(x) \quad \text{in} \quad (-1, 1), \tag{1.24a}$$
$$u(-1) = u(1) = 0, \tag{1.24b}$$

under the following hypotheses:

$$(i) \quad b(x) \neq 0 \quad \text{on} \quad [-1, 1], \tag{1.25a}$$
$$(ii) \quad c(x) \geq 0, \quad c(0) > 0. \tag{1.25b}$$

The assumption $c(0) > 0$ simplifies the problem, as will be seen later. As in the cancellation law of page 12, the location of any boundary layer(s) depends on the sign of the convection term. From our previous experience, we expect a boundary layer at $x = -1$ if the coefficient $xb(x)$ of the convection term is negative at $x = -1$, and a boundary layer at $x = 1$ if the same coefficient is positive at $x = 1$.

If $b(x)$ is positive on $[-1, 1]$, we have $xb(x)|_{x=-1} < 0$ and $xb(x)|_{x=1} > 0$. Consequently, *if b is positive on $[-1, 1]$, then the solution u has two boundary layers*. In this case, the reduced solution is the smooth solution of

$$L_0 u_0 := xb(x)u_0' + c(x)u_0 = f(x) \quad \text{for} \ -1 < x < 1,$$

with no additional boundary condition! The function u_0 is well defined: use $c(0) > 0$ and a Taylor expansion about the singular point $x = 0$. Combining u_0 with two boundary layer corrections, we obtain a first-order asymptotic expansion of u, and it is straightforward to prove a result analogous to Theorem 1.4.

If the condition $c(0) > 0$ is removed, this changes the nature of the problem. In the example

$$-\varepsilon u'' + xu' = x, \quad u(-1) = u(1) = 0,$$

one finds that

$$u_0(x) = x + A,$$

with a constant A that is not determined by the method of matched asymptotic expansions. This is called a *resonance case*. The difficulty arises because $\mu_1 \to 0$ as $\varepsilon \to 0$, where μ_1 is an eigenvalue of

$$-\varepsilon w'' + xw' + \mu w = 0, \quad w(-1) = w(1) = 0.$$

See [dG76] for details of the asymptotic behaviour in this situation.

We return to the case $c(0) > 0$. Our experience in Section 1.1 leads us to expect that *if b is negative on $[-1, 1]$, then boundary layers will not occur*. In this case the reduced solution u_0 satisfies

$$L_0 u_0 = f \quad \text{in} \quad (-1, 0), \quad u_0(-1) = 0,$$

and

$$L_0 u_0 = f \quad \text{in} \quad (0, 1), \quad u_0(1) = 0.$$

The behaviour of u_0 near the turning point $x = 0$ depends strongly on the parameter $\lambda := -c(0)/b(0) > 0$. This is clearly demonstrated by the example

$$x b u_0' + c u_0 = b x^k \quad \text{(constants } b < 0 < c, \text{ integer } k > 0\text{)},$$

whose solution is

$$u_0(x) = \begin{cases} (|x|^k - |x|^\lambda)/(k - \lambda), & \text{if } \lambda \neq k, \\ x^k \ln |x|, & \text{if } \lambda = k. \end{cases}$$

At $x = 0$ the solution has an *interior layer*.

Once more, we digress to the case where $c(0) > 0$ does not hold. If $\lambda = 0$, then an interior *shock layer* in u exists, i.e., u_0 is *discontinuous*. For example, the solution of

$$-x u_0' = x$$

that satisfies $u_0(-1) = u_0(1) = 0$ is

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

Returning to the case $\lambda > 0$, we state without proof a result of Berger et al. [BHK84] on the behaviour of the derivatives of u (see [CL93] for a simpler argument in the case $0 < \lambda < 1$).

Lemma 1.12. *In the turning-point problem (1.24), assume that $b(x)$ is negative and λ is not an integer. Assume also that b, c and f are sufficiently smooth. Write $\lambda = m + \beta$, where m is a non-negative integer and $0 < \beta < 1$. Then the solution u of (1.24) satisfies*

$$|u^{(l)}(x)| \leq C \quad \text{on } (-1, 1) \quad \text{for } l \leq m, \tag{1.26}$$

and for $-1 < x < 1$ and $l = m + 1, m + 2, \dots, q$,

$$|u^{(l)}(x)| \leq C \left(1 + |x| + \varepsilon^{1/2} \right)^{\lambda - l} \quad \text{on } (-1, 1). \tag{1.27}$$

Here the value of q depends on the smoothness of b, c and f.

The interior layer in u is called a *cusp layer* because it can be modelled approximately by the cusp-like function $(x^2 + \varepsilon)^{\lambda/2}$. If one defines the local variable ξ in the layer by $\xi := x/\varepsilon^{1/2}$, one obtains the interior layer equation

$$-\frac{d^2v}{d\xi^2} + b(0)\xi\frac{dv}{d\xi} + c(0)v = 0.$$

The solution of this equation can be expressed in terms of parabolic cylinder functions; see [BHK84].

The problem analysed in Lemma 1.12, where the coefficient of u' has a simple zero, has a *simple turning point* at $x = 0$. If the problem has a finite number of simple turning points in $(-1, 1)$, then the result of this lemma is valid in a neighbourhood of each of these turning points. There are few stability estimates for turning-point problems in the literature; see [Doe98] for some (L_∞, L_∞) and (L_1, L_1) estimates in certain situations for simple turning points. For *multiple turning-point* problems, where the coefficient of u' has a multiple zero, less is known; see [VF93], where such a problem is discussed.

We close this section with a general L_1-norm bound on the derivative of the solution u of (1.1). No assumption is made on the sign of b so this result applies also to solutions of (1.24a).

Theorem 1.13. *For the boundary value problem (1.1), assume that b, c and f are smooth and $c(x) \geq c_0 > 0$ for $0 \leq x \leq 1$. Then there exists a constant C such that*

$$\int_0^1 |u'(x)|\, dx \leq C. \tag{1.28}$$

Proof. The argument uses Lorenz's technique [Lor82, Nii84]. First, write (1.1) in the form

$$-\varepsilon u'' + (bu)' + (c - b')u = f$$

and differentiate, to get

$$(c - b')u' = \varepsilon u''' - (bu)'' + f' - (c' - b'')u. \tag{1.29}$$

An integration by parts then yields

$$\int_0^1 (c - b')u'\, dx = [\varepsilon u'' - (bu)']_0^1 + \int_0^1 [f' - (c' - b'')u]\, dx$$

$$= [(c - b')u - f]_0^1 + \int_0^1 [f' - (c' - b'')u]\, dx.$$

Since $c(x) \geq c_0 > 0$, a comparison principle and barrier function argument gives $\|u\|_\infty \leq \|f\|_\infty/c_0 = C$. Hence

$$\left|\int_0^1 (c - b')u'\, dx\right| \leq C.$$

Unfortunately, this is not exactly the desired estimate and we have to modify the simple argument presented above. Thus, before integrating (1.29), multiply by $\operatorname{sgn}(u')$, where

$$\operatorname{sgn}(z) := \begin{cases} -1 & \text{if } z < 0, \\ 0 & \text{if } z = 0, \\ 1 & \text{if } z > 0. \end{cases}$$

This gives

$$\int_0^1 (c - b')|u'|\, dx = \varepsilon \int_0^1 u'''\operatorname{sgn}(u')\, dx - \int_0^1 (bu)''\operatorname{sgn}(u')\, dx$$
$$+ \int_0^1 [f' - (c' - b'')u]\operatorname{sgn}(u')\, dx.$$

We would like to integrate by parts as before, but this is impossible because the function sgn is not differentiable. Thus replace sgn by a differentiable approximation s_μ, where μ is a positive parameter and $s_\mu \to \operatorname{sgn}$ as $\mu \to 0^+$. This is done by defining

$$s_\mu(z) = \begin{cases} -1 & \text{for} & z \le -\mu, \\ -1 + (z/\mu + 1)^2 & \text{for} & -\mu < z \le 0, \\ 1 - (z/\mu - 1)^2 & \text{for} & 0 < z < \mu, \\ 1 & \text{for} & z \ge \mu, \end{cases}$$

for each $\mu > 0$. For later use, observe that

$$\left| \frac{ds_\mu(z)}{dz} \right| \le \frac{C^*}{\mu} \quad \text{for all } z \in (-1, 1).$$

Replacing s above by s_μ, one obtains

$$\int_0^1 (c - b')u' s_\mu(u')\, dx = \varepsilon \int_0^1 u''' s_\mu(u')\, dx - \int_0^1 (bu)'' s_\mu(u')\, dx$$
$$+ \int_0^1 [f' - (c' - b'')u]s_\mu(u')\, dx.$$

Since

$$\int_0^1 u''' s_\mu(u')\, dx = u'' s_\mu(u')|_0^1 - \int_0^1 (u'')^2 \frac{ds_\mu(z)}{dz}|_{z=u'}\, dx$$

and $ds_\mu(z)/dz \ge 0$, it follows that

$$\int_0^1 u''' s_\mu(u')\, dx \le u'' s_\mu(u')|_0^1.$$

Now letting $\mu \to 0^+$ gives

$$\int_0^1 (c - b')|u'|\, dx \le \varepsilon u'' s(u')|_0^1 + \lim_{\mu \to 0+} E + C \qquad (1.30)$$

with

$$E = -\int_0^1 (bu)'' s_\mu(u')\, dx.$$

Integrating by parts, write

$$E = -(bu)' s_\mu(u')|_0^1 + E_1 + E_2, \quad \text{with} \quad E_2 = \int_0^1 b' u (s_\mu(u'))'\, dx.$$

By Lebesgue's dominated convergence theorem one has

$$\lim_{\mu \to 0+} E_2 = b' u s(u')|_0^1 - \int_0^1 b'|u'|\, dx - \int_0^1 b'' u s(u')\, dx.$$

We will show below that $\lim_{\mu \to 0+} E_1 = 0$. Assuming this for the moment, it follows from (1.30) that

$$\int_0^1 (c - b')|u'|\, dx \le (\varepsilon u'' - bu') s(u')|_0^1 - \int_0^1 b'|u'|\, dx + C,$$

whence

$$\int_0^1 |u'|\, dx \le C,$$

since $c(x) \ge c_0 > 0$ and $\varepsilon u'' - bu' = cu - f$.

To complete the proof, consider $\lim_{\mu \to 0+} E_1$. Now $|bu''| \le K$, where K may depend on ε, and $|(d/dz)(s_\mu(z))| \le C^*/\mu$. Hence

$$|E_1| = \left| \int_{|u'| < \mu} b u' u'' \frac{d}{dz} s_\mu(z)|_{z=u'}\, dx \right|$$

$$\le C^* K(\varepsilon) \operatorname{meas}\{x \in [0,1] \ : \ 0 < |u'(x)| < \mu\},$$

which implies that $\lim_{\mu \to 0+} E_1 = 0$.   □

Theorem 1.13 is quite powerful because it makes no assumption regarding the location or multiplicity of turning points.

1.3 Quasilinear Problems

We now move on to the more general quasilinear boundary value problem

$$-\varepsilon u''(x) + b(x, u(x))u'(x) + c(x, u(x)) = 0, \quad \text{for } x \in (0,1), \qquad (1.31a)$$

$$u(0) = A, \quad u(1) = B. \qquad (1.31b)$$

Unlike the previous sections, inhomogeneous boundary conditions are assumed here since a transformation to homogeneous boundary conditions would alter slightly the nonlinear differential operator. In the *semilinear* case, i.e., when $b(x, u) = b(x)$, results similar to those of Sections 1.1 and 1.2 are valid.

Assume that

$$\frac{\partial c}{\partial s}(x, s) \geq \mu > 0 \quad \text{for all } x \in (0, 1) \text{ and all } s \in R. \tag{1.32}$$

Then Nagumo's theory of upper and lower solutions [CH84] yields existence of a solution u of (1.31) with

$$|u(x)| \leq \max \left\{ \frac{1}{\mu} \max_{x \in [0,1]} |c(x, 0)|, |A|, |B| \right\} \quad \text{for all } x \in [0, 1].$$

This solution is unique [O'M91].

If $b(\cdot, \cdot)$ has constant sign – say $b < 0$ – then, as in Section 1.1, we expect a boundary layer at $x = 0$. The theory is more complicated than in the linear case: one must include a pertinent *boundary layer stability assumption*, as we describe below. For the moment assume that u has a boundary layer at $x = 0$. Then the reduced solution u_R is defined by

$$b(x, u_R)u'_R + c(x, u_R) = 0 \quad \text{on } (0, 1) \quad \text{with } u_R(1) = B,$$

where we assume that

$$b(x, u_R(x)) \leq -\kappa < 0 \quad \text{for all } x \in [0, 1] \text{ and some } \kappa > 0.$$

With the aim of finding a boundary layer correction v_0 at $x = 0$, set $\xi = x/\varepsilon$. Then v_0 should satisfy

$$-\frac{d^2 v_0}{d\xi^2} + b(0, u_R(0) + v_0)\frac{dv_0}{d\xi} = 0, \quad v_0(0) = A - u_R(0).$$

In the linear case, one can compute v_0 explicitly and see that it is exponentially decaying. But in the nonlinear case, the existence of exponentially boundary layers v_0 depends on $|A - u_R(0)|$. One needs the following additional boundary layer stability assumption [CH84, VBK95], which guarantees that the boundary layer jump $|A - u_R(0)|$ belongs to the domain of influence of the asymptotically stable solution $v_0 \equiv 0$:

$$\int_\eta^{u_R(0)} b(0, s) \, ds < 0 \quad \text{if } A < \eta < u_R(0) \tag{1.33a}$$

and

$$\int_{u_R(0)}^\eta b(0, s) \, ds < 0 \quad \text{if} \quad u_R(0) < \eta < A. \tag{1.33b}$$

The necessity of the inequalities (1.33) can be deduced from the implicit representation

$$\xi = \int_{v_0}^{A-u_R(0)} \frac{ds}{q(s)} \quad \text{where } q(s) = -\int_0^s b(0, u_R(0)+t)dt.$$

These conditions say essentially that the jump $|A - u_R(0)|$ should not be too large; if they are violated, then we cannot construct a boundary layer correction at $x = 0$.

A rigorous analysis leads to the following classical result [O'M91], which is due to Coddington and Levinson.

Theorem 1.14. *Assume that b and c are sufficiently smooth. Define the reduced solution u_R by*

$$b(x, u_R)u_R' + c(x, u_R) = 0 \text{ on } (0, 1) \text{ with } u_R(1) = B.$$

Assume that $b(x, u_R(x)) \leq -\kappa < 0$ and that the boundary layer stability conditions (1.33) are satisfied. Then for $0 < x < 1$ one has

$$u(x) = u_R(x) + O\left(|A - u_R(0)|\exp(-\kappa x/\varepsilon)\right) + O(\varepsilon),$$
$$u'(x) = u_R'(x) + O\left(\varepsilon^{-1}\exp(-\kappa x/\varepsilon)\right) + O(\varepsilon).$$

The hypotheses of Theorem 1.14 can be weakened. In particular, one can replace the condition $b(x, u_R(x)) \leq -\kappa < 0$ by the hypothesis that u_R is *globally stable*, viz., that $b(x, u_R(x)) < 0$ for $0 < x \leq 1$; see [How78, Theorem 5.5]. Analogously, if u_L is defined by

$$b(x, u_L)u_L' + c(x, u_L) = 0 \quad \text{with } u_L(0) = A,$$

we say that u_L is globally stable if $b(x, u_L(x)) > 0$ for $0 \leq x < 1$.

One can verify that the conditions of Theorem 1.14 are satisfied in the example

$$-\varepsilon u'' - e^u u' + \frac{\pi}{2}\sin\frac{\pi x}{2}e^{2u} = 0, \quad u(0) = A, \ u(1) = 0,$$

without any restriction on the boundary layer jump.

In the example

$$-\varepsilon u'' - uu' + u = 0, \quad u(0) = -2, \ u(1) = 1.5, \tag{1.34}$$

both $u_R(x) = x + 0.5$ and $u_L(x) = -2 + x$ are globally stable, but neither boundary layer stability condition (the condition for u_L is analogous to (1.33)) is satisfied:

$$\int_{A=-2}^{u_R(0)=0.5} (-s)\, ds \not< 0 \quad \text{and} \quad \int_{u_L(1)=-1}^{B=1.5} (-s)\, ds \not> 0.$$

Thus a boundary layer cannot exist at $x = 0$ nor at $x = 1$ because the boundary layer jump is too large! That is, the solution u has an interior layer but no boundary layer.

As with the linear turning-point problems of Section 1.2, we expect interior layers if no boundary layer is present. In the nonlinear case the analysis can be much more complicated than before. It is not easy to find the location(s) of possible interior layers, and the reduced equation may have more than one solution – then it is not clear which of these is the correct limit (as $\varepsilon \to 0$) of the exact solution u in a given subinterval and where a transition from one reduced solution to another takes place. A discontinuous transition will cause a *shock* layer in the solution u, and a continuous transition a *corner* layer.

We sketch the situation for the problem

$$-\varepsilon u'' + b(u)u' + c(x, u) = 0 \quad \text{for } x \in (0, 1), \tag{1.35a}$$
$$u(0) = A, \quad u(1) = B, \tag{1.35b}$$

under the hypothesis (1.32). It is easier to handle (1.35) than (1.31) because the convection term can be written in the conservation form

$$b(u)u' = (e(u))', \quad \text{with } e(u) := \int^u b(s)\,ds.$$

The principal approach used to find the reduced solution $u_0(x) := \lim_{\varepsilon \to 0} u(x)$ is a standard technique in the theory of conservation laws (see [LeV90]); these are equations of the form $u_t + (e(u))' = 0$, where t is a time variable.

Introduce the *entropy flux* $E(\cdot)$ and the convex *entropy function* $U(\cdot)$, which depend on $e(\cdot)$ above. These functions are related by

$$\frac{dE}{dz} = \frac{dU}{dz}\frac{de}{dz}.$$

A simple example is $U(z) = z^2/2$, $E(z) = \int^z se'(s)\,ds$. Another important choice is due to Kruzkov [LeV90]: set

$$U(z) = |z - k| \quad \text{and} \quad E(z) = [e(z) - e(k)]\operatorname{sgn}(u - k),$$

where k is an arbitrary constant. Multiplying the differential equation (1.35a) by $U'(u)$, one writes it in the form

$$\frac{d}{dx}E(u) + U'(u)c(x, u) = \varepsilon\frac{d^2}{dx^2}U(u) - \varepsilon U''(u)\left(\frac{du}{dx}\right)^2.$$

Now multiply by a smooth function φ, integrate by parts, and take the limit as $\varepsilon \to 0$. This steers us to the inequality

$$\int_0^1 [-E(u_0)\varphi' + U'(u_0)c(x, u_0)\varphi]\,dx \le -E(u_0)\varphi|_0^1.$$

That is, Kruzkov's choice yields

$$\int_0^1 \text{sgn}(u_0 - k) \left[(e(u_0) - e(k))\varphi' - c(x, u_0)\varphi \right] dx$$

$$\geq \sum_{i=0,1} (-1)^i \text{sgn}(u_0(i) - k)(e(u_0(i)) - e(k))\varphi(i).$$

If one chooses special test functions φ, this yields [Lor84] the following convenient characterization of the reduced solution u_0 :

Theorem 1.15. *For $0 \leq x \leq 1$, set $u_0(x) = \lim_{\varepsilon \to 0} u(x)$, where u is the solution of (1.35). Then*
(i)   If u_0 is smooth in a subinterval, it satisfies the reduced equation

$$b(u_0)u_0' + c(x, u_0) = 0.$$

(ii) At the boundaries $x = 0$ and $x = 1$, u_0 satisfies

$$\text{sgn}(u_0(0) - A) \int_k^{u_0(0)} b(s)\, ds \leq 0 \text{ for all } k \text{ between } A \text{ and } u_0(0),$$

$$\text{sgn}(u_0(1) - B) \int_k^{u_0(1)} b(s)\, ds \geq 0 \text{ for all } k \text{ between } B \text{ and } u_0(1).$$

(iii) At a discontinuity $x_* \in (0,1)$ of u_0, the following jump condition is satisfied:

$$\text{sgn}(u_0(x_*^+) - u_0(x_*^-)) \int_k^{u_0(x_*)} b(s)\, ds \geq 0$$

$$\text{for all } k \text{ between } u_0(x_*^+) \text{ and } u_0(x_*^-).$$

Part (ii) of Theorem 1.15 is closely related to the boundary layer stability conditions (1.33), and the characterization (iii) allows us to find the position of interior layers.

For example, consider the case where u_L and u_R are globally stable but no boundary layer exists. For convenience we assume that $u_L < 0 < u_R$. We expect that

$$u_0(x) = \begin{cases} u_L(x) & \text{for } 0 \leq x < x_*, \\ u_R(x) & \text{for } x_* < x \leq 1, \end{cases}$$

but x_* is unknown. Theorem 1.15 (iii) tells us that

$$J(x_*) = 0, \quad \text{where } J(x) := \int_{u_L(x)}^{u_R(x)} b(s)\, ds. \tag{1.36}$$

Because no boundary layer is present,

$$J(0) = \int_A^{u_R(0)} b(s)ds > 0 \quad \text{and} \quad J(1) = \int_{u_L(1)}^B b(s)ds < 0.$$

Furthermore, for some $\zeta \in (u_L, u_R)$,

$$J'(x) = b(u_R)u'_R - b(u_L)u'_L = c(x, u_L) - c(x, u_R) = c_u(x, \zeta)(u_L - u_R)$$
$$< 0.$$

Hence x_* is uniquely determined by (1.36). In example (1.34),

$$J(x) = \int_{-2+x}^{x+0.5} (-s)ds = -\frac{1}{2}(5x - 3.75),$$

which delivers the value $x_* = 0.75$.

Suppose now that we know only that $b(x, u_L(x)) > 0$ on $[0, x_L)$ for some $x_L \in (0, 1)$ (i.e., u_L is stable only on $[0, x_L)$), and $b(x, u_R(x)) < 0$ on $(x_R, 1]$ for some $x_R \in (x_L, 1)$. Then one expects that

$$u_0(x) = \begin{cases} u_L(x) \text{ for } 0 \le x \le x_L, \\ u_s(x) \text{ for } x_L \le x \le x_R, \\ u_R(x) \text{ for } x_R \le x \le 1, \end{cases}$$

with u_s a smooth solution of the reduced equation and corner layers at x_L and x_R. If example (1.34) is modified to

$$-\varepsilon u'' - uu' + u = 0, \quad u(0) = -\frac{1}{2}, \; u(1) = \frac{1}{3},$$

then one gets $u_L(x) = -1/2 + x$ with $x_L = 1/2$, and $u_R(x) = x - 2/3$ with $x_R = 2/3$. In this example, $u_s \equiv 0$ and

$$u_0(x) = \begin{cases} x - \frac{1}{2} & \text{for} & 0 \le x \le \frac{1}{2}, \\ 0 & \text{for} & \frac{1}{2} \le x \le \frac{2}{3}, \\ x - \frac{2}{3} & \text{for} & \frac{2}{3} \le x \le 1. \end{cases}$$

We end with a stability result from [Lor82] and an *a priori* bound on the first-order derivative of the exact solution of the quasilinear problem (1.35). Define the operator T by

$$Tv := -\varepsilon v'' + b(v)v' + c(x, v).$$

Theorem 1.16. *In the boundary value problem (1.31) assume that*

$$\frac{\partial c}{\partial s}(x, s) \ge \mu > 0 \quad \text{for all } x \in (0, 1) \text{ and all } s \in R.$$

Then for all v and w in $C^2(0, 1)$ that satisfy $v(0) = w(0)$ and $v(1) = w(1)$, one has

$$\|v - w\|_{L_1} \le \frac{1}{\mu}\|Tv - Tw\|_{L_1}.$$

Furthermore,

$$\int_0^1 |u'(x)|dx \le C.$$

The proof of the stability result uses the Green's function of the linearized problem, while the proof of the *a priori* bound for u' resembles the proof of Theorem 1.13.

1.4 Linear Higher-Order Problems and Systems

1.4.1 Asymptotic Expansions for Higher-Order Problems

Consider the linear differential equation

$$Lu := \varepsilon^{m-n} u^{(m)} + \sum_{\nu=0}^{n} a_\nu(x) u^{(\nu)} = f(x), \quad \text{for } 0 < x < 1, \tag{1.37}$$

subject to the boundary conditions

$$u^{(\mu_i)}(0) = 0, \quad \text{for } i = 1, ..., r, \tag{1.38a}$$

$$u^{(\mu_i)}(1) = 0, \quad \text{for } i = r+1, ..., m. \tag{1.38b}$$

Here m and n are positive integers with $m > n$, so the order of the differential equation decreases if one sets $\varepsilon = 0$. The boundary conditions are ordered so that $m > \mu_1 > \mu_2 > ... > \mu_r \geq 0$ and $m > \mu_{r+1} > \mu_{r+2} > ... > \mu_m \geq 0$. Furthermore, we exclude turning points by assuming that

$$a_n(x) \neq 0 \quad \text{for all } x \in [0, 1]. \tag{1.39}$$

Applying the method of matched asymptotic expansions, the leading part u_0 of the global expansion satisfies the nth-order equation

$$L_0 u_0 := \sum_{\nu=0}^{n} a_\nu(x) u_0^{(\nu)} = f.$$

It is natural to attach n boundary conditions to this differential equation. That is, $m - n$ of the original m boundary conditions will be discarded and we must decide which conditions to retain.

Introduce the local variable $\xi = x/\varepsilon$ to investigate possible boundary layers at $x = 0$ (one could similarly explore the behaviour of u near $x = 1$). The leading term in the local correction is a differential equation with constant coefficients. Its characteristic equation is

$$\lambda^n \left(\lambda^{m-n} + a_n(0) \right) = 0.$$

Suppose that σ roots of this equation have negative real part and τ roots have positive real part. Two possible situations can occur [O'M91]: in the *nonexceptional* case, $\sigma + \tau = m - n$, while in the *exceptional* case there are two pure imaginary roots so $\sigma + \tau = m - n - 2$. The corresponding *cancellation law* is:

- Cancel σ boundary conditions at $x = 0$ and τ boundary conditions at $x = 1$, choosing those with the highest-order derivatives.
- In the exceptional case, also cancel from the remaining boundary conditions those two with the highest-order derivatives, provided that they belong to the same endpoint and that the selection is without ambiguity.

After the application of the cancellation law, the reduced solution is required to satisfy the remaining n boundary conditions; this defines the *reduced problem*. If the cancellation law and reduced problem are well defined then the method of matched asymptotic expansions works, but the cancellation law is not well defined in all cases.

For example, consider the boundary value problem

$$\varepsilon^2 u^{(4)} - u'' = f(x) \quad \text{for } x \in (0,1),$$

subject to the boundary conditions

$$u'''(0) = u(0) = u'(1) = u(1) = 0.$$

Here we have $\sigma = \tau = 1$ and the cancellation law is well defined. The reduced problem is

$$-u_0'' = f \quad \text{with} \quad u_0(0) = u_0(1) = 0.$$

This has a unique solution. We find that $u(x)$ has an asymptotic expansion of the form

$$u_{as}(x) = \sum_{\nu=0}^{m} u_\nu(x)\varepsilon^\nu + \varepsilon^3 \left(\sum_{\mu=0}^{m} v_\mu(\xi)\varepsilon^\mu \right) e^{-x/\varepsilon}$$

$$+ \varepsilon \left(\sum_{\mu=0}^{m} w_\mu(\zeta)\varepsilon^\mu \right) e^{-(1-x)/\varepsilon},$$

for arbitrary m, with $\xi = x/\varepsilon$ and $\zeta = (1-x)/\varepsilon$. This expansion can be formally differentiated to get information about derivatives of u; see [O'M91].

Little is known about higher-order problems with turning points.

1.4.2 A Stability Result

Stability is an essential property of every discretization method and to get some insight into this property one must study the stability properties of the given continuous problem. Furthermore, asymptotic expansions require high smoothness of the coefficients of the problem; consequently, they may fail to provide sufficient information about derivatives of the exact solution for the analysis of discretization methods.

We consider the boundary value problem (1.37)–(1.38), under the assumption (1.39), for the case $n = m-1$. That is, the order of the differential equation decreases by one if $\varepsilon = 0$. We introduce the abbreviation

$$Bu = (B_1 u, B_2 u, ..., B_m u) = 0$$

for the m boundary conditions (1.38) and define the norm

$$\|v\|_{\varepsilon,m-1,\infty} := \max \left\{ \|v\|_\infty, \|v'\|_\infty, ..., \|v^{(m-2)}\|_\infty, \varepsilon\|v^{(m-1)}\|_\infty \right\}.$$

Remark 1.17. It is possible to replace $\varepsilon\|v^{(m-1)}\|_\infty$ by $\|v^{(m-1)}\|_{L_1}$.   ♣

Niederdrenk and Yserentant [NY83] prove the following stability estimate for continuous coefficients, and Gartland [Gar91] extends it to the case

$$a_{m-1} \in L_\infty, \quad a_0, a_1, ..., a_{m-2} \in L_1. \tag{1.40}$$

Theorem 1.18. *Assume that the boundary conditions are bounded with respect to the norm $\|\cdot\|_{\varepsilon,m-1,\infty}$, in the sense that*

$$|B_\nu(v)| \le C\|v\|_{\varepsilon,m-1,\infty} \text{ for } \nu = 1, ..., m.$$

Suppose that (1.40) is satisfied. If there exists a fundamental system $\{\phi_\nu\}$ for $L\phi = 0$ that satisfies

$$\|\phi_\nu\|_{\varepsilon,m-1,\infty} \le C,$$

and the $m \times m$ matrix $[B_\mu(\phi_\nu)]$ has an inverse whose norm (induced by the discrete L_1 norm) can be bounded independently of ε, then we have the stability inequality

$$\|v\|_{\varepsilon,m-1,\infty} \le C\left(\|Lv\|_{L_1} + |Bv|\right).$$

The theorem is also valid for more general boundary condition functionals. Note that for (1.38), the boundedness of the boundary conditions with respect to the norm $\|\cdot\|_{\varepsilon,m-1,\infty}$ requires that

$$\mu_1 \le m - 2 \quad \text{and} \quad \mu_{r+1} \le m - 2;$$

thus the boundary conditions cannot contain the $(m-1)^{\text{th}}$ derivative.

The conditions on the fundamental system and on the inverse of the matrix $[B_\mu(\phi_\nu)]$ are opposing constraints, as can be seen from a careful study of the following example.

Example 1.19. Consider the differential operator and boundary conditions

$$Lu := \varepsilon u^{(4)} + u''', \quad Bu := (u(0), u''(0), u(1), u''(1)).$$

Then the fundamental system $\{1, x, x^2, \varepsilon^2 e^{-x/\varepsilon}\}$ satisfies the conditions of Theorem 1.18. With homogeneous boundary data, the theorem gives not only stability but also the *a priori* estimate

$$\|u\|_\infty + \|u'\|_\infty + \|u''\|_\infty + \varepsilon\|u'''\|_\infty + \|u'''\|_{L_1} \le C\|f\|_{L_1}.$$

If, however, the boundary conditions are

$$Bu := (u(0), u'(0), u(1), u'(1)),$$

then Theorem 1.18 does *not* apply and stability holds only in some weaker norm.   ♣

Little attention has been paid in the literature to the case $n \le m - 2$ for $m > 2$. See [SS95a] for some results when $n = m - 2$.

1.4.3 Systems of Ordinary Differential Equations

Systems of ordinary differential equations are often discussed in books on asymptotic expansions for singularly perturbed problems: see, e.g., [O'M91, Chapter 3], [VB90, Chapter 2] or [Was65, Chapter 7]. Nevertheless in the past relatively little attention was paid to their numerical solution, although the papers [Bak69] (reaction-diffusion systems) and [AKK74] (convection-diffusion systems) are worth noting. In recent years interest in this area has grown, as we now describe.

Consider a general system of M equations:

$$L\mathbf{u} := -\varepsilon\mathbf{u}'' + B\mathbf{u}' + A\mathbf{u} = \mathbf{f} \quad \text{on } \Omega := (0,1), \qquad (1.41\text{a})$$

$$\mathbf{u}(0) = \mathbf{g}_0, \ \mathbf{u}(1) = \mathbf{g}_1, \qquad (1.41\text{b})$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_M)^T$ is the unknown solution while $\mathbf{f} = (f_1, \ldots, f_M)^T$, $\mathbf{g}_0$ and $\mathbf{g}_1$ are constant column vectors, and $A = (a_{ij})$ and $B = (b_{ij})$ are $M \times M$ matrices.

The system (1.41) is said to be *weakly coupled* if the convection coupling matrix B is diagonal, i.e., the ith equation of the system is

$$-\varepsilon u_i'' + b_{ii} u_i' + \sum_{j=1}^{M} a_{ij} u_j = f_i, \qquad (1.42)$$

so the system is coupled only through the lower-order reaction terms.

Linß [Lin07b] allows different diffusion coefficients in different equations: $\varepsilon = \varepsilon_i$ in the ith equation for $i = 1, \ldots, M$. Assume that $b_{ii}(x) \geq \beta_i > 0$ and $a_{ii}(x) \geq \alpha > 0$ on $[0,1]$ for each i. (In [Lin07b] the weaker hypothesis $|b_{ii}(x)| \geq \beta_i > 0$ is used, which permits layers in $\mathbf{u}$ at both ends of $[0,1]$, but for brevity we won't consider this here.) Rewrite (1.42) as

$$-\varepsilon_i u_i'' + b_{ii} u_i' + a_{ii} u_i = -\sum_{j \neq i} a_{ij} u_j + f_i, \qquad (1.43)$$

Then $\|u_i\|_\infty \leq \|(-\sum_{j \neq i} a_{ij} u_j + f_i)/a_{ii}\|_\infty$ by a standard maximum principle argument. Rearranging, one gets

$$\|u_i\|_\infty - \sum_{j \neq i} \left\|\frac{a_{ij}}{a_{ii}}\right\|_\infty \|u_j\|_\infty \leq \left\|\frac{f_i}{a_{ii}}\right\|_\infty \quad \text{for } i = 1, \ldots, M.$$

Define the $M \times M$ matrix $\Gamma = (\gamma_{ij})$ by $\gamma_{ii} = 1$, $\gamma_{ij} = -\|a_{ij}/a_{ii}\|_\infty$ for $i \neq j$. Assume that Γ is inverse-monotone, i.e., that $\Gamma^{-1} \geq 0$. It follows that $\|\mathbf{u}\|_\infty \leq C\|\mathbf{f}\|_\infty$ for some constant C, where $\|\mathbf{v}\|_\infty = \max_i \|v_i\|_\infty$ for $\mathbf{v} = (v_1, \ldots, v_M)^T$. One can now apply the scalar-equation analysis of Lemma 1.8 to (1.43) for each i and get

$$|u_i^{(k)}(x)| \leq C\left[1 + \varepsilon_i^{-k} e^{-\beta_i(1-x)/\varepsilon_i}\right] \quad \text{for } x \in [0,1] \text{ and } k = 0, 1.$$

Thus there is no strong interaction between the layers in the first-order derivatives of different components u_i; nevertheless the domains of these layers can overlap and this influences the construction of numerical methods for (1.41).

The system (1.41) is said to be *strongly coupled* if for some $i \in \{1, \dots, M\}$ one has $b_{ij} \neq 0$ for some $j \neq i$. Such systems do not satisfy a maximum principle of the usual type. One now gets stronger interactions between layers; see [AKK74, Lin07a, OS, OSS]. For each i assume $b_{ii}(x) \geq \beta_i > 0$ and $a_{ii}(x) \geq 0$ on $[0, 1]$. Rewrite the ith equation as

$$L_i u := -\varepsilon u_i'' + b_{ii} u_i' + a_{ii} u_i = f_i + \sum_{\substack{j=1 \\ j \neq i}}^{m} \left[(b_{ij} u_j)' - (b_{ij}' + a_{ij}) u_j \right], \quad (1.44a)$$

$$u_i(0) = u_i(1) = 0. \quad (1.44b)$$

For the scalar problem $L_i v = \phi$ and $v(0) = v(1) = 0$, one has by (1.20) – see [AK98, And02] for the case where (1.9) is not satisfied – the stability result $\|v\|_\infty \leq C_i \|\phi\|_{W^{-1,\infty}}$ for a certain constant C_i that depends only on b_{ii} and a_{ii}. Apply this result to (1.44) then, similarly to the analysis of (1.43), gather the $\|u_j\|_\infty$ terms to the left-hand side. Define the $M \times M$ matrix $\Upsilon = (\gamma_{ij})$ by $\gamma_{ii} = 1$, $\gamma_{ij} = -C_i[\|b_{ij}' + a_{ij}\|_{L_1} + \|b_{ij}\|_\infty]$ for $i \neq j$. Assuming that Υ is inverse monotone, we get an a priori bound on $\|\mathbf{u}\|_\infty$. Using this bound, it is shown in [OSS] that one can decompose each component of $\mathbf{u}$ similarly to (1.22) and (1.23).

For the analysis of systems of reaction-diffusion equations (i.e., $B \equiv 0$ in (1.41)), see [Bak69, LM, MS03].

2

Numerical Methods for Second-Order Boundary Value Problems

2.1 Finite Difference Methods on Equidistant Meshes

2.1.1 Classical Convergence Theory for Central Differencing

This section examines linear two-point boundary value problems that are not singularly perturbed, in order to introduce the classical terminology of finite difference methods. Thus consider the problem

$$Lu := -u'' + b(x)u' + c(x)u = f(x), \quad u(0) = u(1) = 0, \qquad (2.1)$$

under the assumptions that b, c, f are smooth and $c(x) \geq 0$.

Finite difference methods will be studied on an *equidistant* grid with *mesh size* $h = 1/N$; that is, set

$$x_i = ih \quad \text{for } i = 0, 1, ..., N, \quad \text{with } x_0 = 0 \text{ and } x_N = 1.$$

(We could work equally well with almost-equidistant meshes, but for simplicity restrict ourselves to the equidistant case. See Section 2.4 for a classification of meshes and for extensions of the theory to meshes that are not almost equidistant.)

A finite difference method is a discretization of the differential equation using the *grid points* x_i, where the unknowns u_i (for $i = 0, \ldots, N$) are approximations of the values $u(x_i)$. It is natural to approximate $u'(x)$ by the *central difference*

$$(D^0 u)(x) := [u(x + h) - u(x - h)]/(2h).$$

Composing the *forward* and *backward* differences

$$(D^+ u)(x) := [u(x + h) - u(x)]/h \text{ and } (D^- u)(x) := [u(x) - u(x - h)]/h,$$

yields the following central approximation for $u''(x)$:

$$(D^+ D^- u)(x) := [u(x + h) - 2u(x) + u(x - h)]/h^2.$$

The *order of accuracy* of every finite difference approximation depends on the smoothness of u. For instance, Taylor's formula yields

$$u(x \pm h) = u(x) \pm hu'(x) + h^2\frac{u''(x)}{2} \pm h^3\frac{u'''(x)}{6} + R_4,$$

with

$$R_4 = \int_x^{x \pm h} [u'''(\xi) - u'''(x)]\frac{(x \pm h - \xi)^2}{2} \, d\xi.$$

Hence

$$\left|(D^+D^-u)(x) - u''(x)\right| \leq Kh^2 \quad \text{if } u \in C^4, \tag{2.2}$$

– this condition can be weakened to the Lipschitz continuity of u''' – and we say that D^+D^- is second-order accurate, which is sometimes written as $O(h^2)$ accurate. The order decreases if u is less smooth; for example, if one only has $u \in C^3$, then D^+D^- is first-order accurate. Using the notation

$$g_i = g(x_i), \quad \text{where } g \text{ can be } b, c \text{ or } f,$$

the classical *central difference scheme* for the boundary value problem (2.1) is

$$-D^+D^-u_i + b_iD^0u_i + c_iu_i = f_i \quad \text{for } i = 1, ..., N-1, \tag{2.3a}$$
$$u_0 = u_N = 0. \tag{2.3b}$$

This is a tridiagonal system of linear equations:

$$r_iu_{i-1} + s_iu_i + t_iu_{i+1} = f_i \quad \text{for } i = 1, ..., N-1, \quad \text{with } u_0 = u_N = 0, \tag{2.4}$$

where

$$r_i = -\frac{1}{h^2} - \frac{1}{2h}b_i, \quad s_i = c_i + \frac{2}{h^2}, \quad t_i = -\frac{1}{h^2} + \frac{1}{2h}b_i. \tag{2.5}$$

Two questions must now be tackled: what properties does the discrete problem (2.3) enjoy? What can we say about the errors $|u(x_i) - u_i|$?

Classical convergence theory for finite difference methods is based on the complementary concepts of *consistency* and *stability*. First, formally write (2.3) (or any difference scheme) as

$$L_hu_h = f_h, \tag{2.6}$$

where L_h is a matrix,

$$u_h := (u_h(x_0), u_h(x_1), ..., u_h(x_N))^T := (u_0, u_1, \ldots, u_N)^T,$$

and $f_h := (f(x_0), f(x_1), \ldots, f(x_N))^T$. Functions defined on the grid, such as u_h and f_h, are called *grid functions*. The restriction of a function $v \in C[0,1]$ to a grid function is denoted by R_hv, viz., $R_hv = (v(x_0), v(x_1), ..., v(x_N))$. We sometimes omit R_h when the meaning is clear. The discrete maximum norm on the space of grid functions is

$$\|v_h\|_{\infty,d} := \max_i |v_h(x_i)|.$$

Definition 2.1. *Consider a difference scheme of the form $L_h u_h = R_h(Lu)$, where we incorporate the boundary conditions into the scheme by taking the first and last rows of L_h to be identical to the first and last rows respectively of the identity matrix, with $(R_h Lu)_0 = u_0$ and $(R_h Lu)_N = u_N$. This scheme is consistent of order k in the discrete maximum norm if*

$$\|L_h R_h u - R_h Lu\|_{\infty,d} \leq Kh^k,$$

where the positive constants K and k are independent of h.

One could define consistency analogously with respect to an arbitrary norm. As in (2.2), one can apply Taylor's formula to prove

Lemma 2.2. *Under the assumption $u \in C^4[0,1]$, the central difference scheme (2.3) is consistent of order two.*

Applying the discrete operator L_h to the error at the interior grid points yields

$$L_h(R_h u - u_h) = L_h R_h u - f_h = L_h R_h u - R_h Lu. \tag{2.7}$$

In order to estimate $R_h u - u_h$ from (2.7) and the consistency order, it is natural to introduce the concept of *stability*.

Definition 2.3. *A discrete problem $L_h u_h = f_h$ is stable in the discrete maximum norm, if there exists a constant K (the stability constant) that is independent of h, such that*

$$\|u_h\|_{\infty,d} \leq K\|L_h u_h\|_{\infty,d} \tag{2.8}$$

for all mesh functions u_h.

Note that, analogously to the continuous case, one could generalize this to (A, B) stability which is particularly important for non-equidistant meshes. Thus, to be precise, Definition 2.3 deals with (L_∞, L_∞) stability.

Our final ingredient is

Definition 2.4. *A difference method for (2.1) is convergent (of order k) in the discrete maximum norm if there exist positive constants K and k that are independent of h for which*

$$\|u_h - R_h u\|_{\infty,d} \leq Kh^k.$$

The main result of classical convergence theory for finite difference methods now follows immediately:

$$Consistency + Stability \Longrightarrow Convergence.$$

The investigation of the order of consistency is usually based on Taylor's formula and is straightforward. But to prove stability one needs some new tools.

In classical finite difference analyses, it is standard to use the theory of *M-matrices*, which is now described; see [Boh81, OR70] for further information.

The material that follows uses the natural ordering of vectors, viz., $x \leq y$ if and only if $x_i \leq y_i$ for all i. Sometimes we simply write $z \geq 1$ when we mean that $z_i \geq 1$ for all i. For each matrix $A = (a_{ij})$, the inequality $A \geq 0$ means that $a_{ij} \geq 0$ for all i and j.

A matrix A for which A^{-1} exists with $A^{-1} \geq 0$ is called *inverse-monotone*.

Lemma 2.5 (Discrete comparison principle). *Let A be inverse-monotone. Then $Av \leq Aw$ implies that $v \leq w$.*

Proof. The argument is simple: multiply $A(v - w) = b \leq 0$ by A^{-1} and use $A^{-1} \geq 0$. □

The class of M-matrices is an important subset of the class of inverse-monotone matrices.

Definition 2.6. *A matrix A is an M-matrix if its entries a_{ij} satisfy $a_{ij} \leq 0$ for $i \neq j$ and its inverse A^{-1} exists with $A^{-1} \geq 0$.*

The diagonal entries of an M-matrix satisfy $a_{ii} > 0$.

While the condition $a_{ij} \leq 0$ is easy to check, it may be difficult to verify directly the inequality $A^{-1} \geq 0$. Fortunately, several equivalent but more tractable characterizations of M-matrices are known. The following result is frequently used in the context of discretization methods (see [Boh81] or [AK90] for a proof).

Theorem 2.7 (M-criterion). *Let the matrix A satisfy $a_{ij} \leq 0$ for $i \neq j$. Then A is an M-matrix if and only if there exists a vector $e > 0$ such that $Ae > 0$. Furthermore, we have*

$$\|A^{-1}\|_{\infty,d} \leq \frac{\|e\|_{\infty,d}}{\min_k (Ae)_k} . \tag{2.9}$$

Here the matrix norm is the norm induced by the corresponding vector norm.

In Theorem 2.7 the vector e is called a *majorizing element* for the matrix A. This theorem allows us to verify that the coefficient matrix of a given discretization is an M-matrix while simultaneously estimating the stability constant from (2.9) — provided that we are able to find a majorizing element. The following recipe for construction of this element is often successful:

- Find a function $e > 0$ such that $Le(x) > 0$ for $x \in (0,1)$ – this is a majorizing element for the differential operator L.
- Restrict e to a grid function e_h.

In general, if the first step in this method is feasible then the method will work (at least for sufficiently small h) provided the discretization is consistent to some positive order.

For homogeneous boundary conditions one usually eliminates the variables u_0 and u_N before applying Theorem 2.7.

Example 2.8. Consider the special case where $b(x) \equiv 0$ in the differential operator L of (2.1). Choose $e(x) := x(1-x)/2$. Then

$$Le(x) = 1 + c(x)e(x) \geq 1.$$

On setting $e_h := R_h e$ one obtains

$$L_h e_h \geq (1, ..., 1)^T.$$

since $D^+ D^-$ discretizes quadratic functions exactly at the interior grid points. Now inequality (2.9) provides a stability constant of $1/8$.   ♣

In the general case of (2.1), the construction of a majorizing element is slightly more complicated. Define $e(x)$ to be the solution of the boundary value problem

$$-w'' + b(x)w' = 1, \quad w(0) = w(1) = 0.$$

Then $e(x) > 0$ for $x \in (0,1)$ and $e(x)$ is bounded. The inequality $c(x) \geq 0$ and the consistency of the discretization imply that at the interior grid points one has

$$L_h e_h = R_h Le + (L_h e_h - R_h Le) \geq 1/2$$

for all sufficiently small h, because $R_h Le = 1$. This proves

Lemma 2.9. *For all sufficiently small h, the central difference scheme for the boundary value problem (2.1) is stable in the discrete maximum norm; moreover, the corresponding coefficient matrix is then an M-matrix.*

One can clearly combine Lemmas 2.2 (consistency) and 2.9 (stability) to obtain a second-order convergence result.

Remark 2.10. In general, the proof of stability via M-matrices is inapplicable to higher-order difference schemes that are based on stencils with more than three points. It may nevertheless be possible to use the property of strong diagonal dominance (see, e.g., [Her90]) or to factor a matrix as a product of M-matrices [Lor75] or to use special splittings [AK90]. For a general stability theory of difference schemes see [Gri85a].   ♣

2.1.2 Upwind Schemes

This subsection and its two successors study difference schemes for the singularly perturbed boundary value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{on } (0,1), \quad u(0) = u(1) = 0, \quad (2.10)$$

when turning points are excluded, i.e., when $b(x) \neq 0$ for all $x \in [0,1]$. We also assume that $c \geq 0$ on [0,1] and that the functions b, c and f are smooth. Recall that for $b > 0$ there is an exponential boundary layer at $x = 1$, and for

$b < 0$ the boundary layer is at $x = 0$. The conditions "$b < 0$" and "$b > 0$" are equivalent: the change of variable $x \mapsto 1 - x$ transforms the problem from one formulation to the other.

Suppose that $\varepsilon > 0$ is small. If u exhibits a boundary layer, this adversely affects both consistency and stability. If instead the boundary conditions are such that u has no layer, then the consistency error improves but stability may still be a problem.

To begin, the central difference scheme is applied to the example

$$-\varepsilon u'' + u' = 0 \text{ on } (0, 1), \quad u(0) = 0, \ u(1) = 1.$$

A transformation $u(x) = x + v(x)$ would give homogeneous boundary conditions, but one can use the scheme directly with inhomogeneous conditions. The discrete problem is

$$-\varepsilon D^+ D^- u_i + D^0 u_i = 0, \quad u_0 = 0, \ u_N = 1.$$

It is easy to solve this exactly:

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

If $h \gg 2\varepsilon$, then $r \approx -1$ so this computed solution oscillates badly and is not close to the true solution

$$u(x) = \frac{e^{-(1-x)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}.$$

Figure 2.1 shows the oscillations of the central scheme on an uniform mesh if ε is small compared with h. On the other hand if $h < 2\varepsilon$, then the central difference scheme works — but from the practical point of view this assumption is unsatisfactory when, for instance, $\varepsilon = 10^{-5}$. A fortiori, in two or three dimensions such a mesh restriction would lead to unacceptably large numbers of mesh points, as for small ε the dimension of the algebraic system generated would be too large for computer solution.

Returning to the general problem (2.10), write the central difference scheme in the form of (2.5), viz.,

$$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{2h}b_i, \quad s_i = c_i + \frac{2\varepsilon}{h^2}, \quad t_i = -\frac{\varepsilon}{h^2} + \frac{1}{2h}b_i.$$

This gives an M-matrix and hence stability if we assume that

$$h \le h_0(\varepsilon) = \frac{2\varepsilon}{\|b\|_\infty},$$

which generalizes the observation of the example above. Note that $h_0(\varepsilon) \to 0$ if $\varepsilon \to 0$. This conclusion is not confined to the central difference scheme: *Classical numerical methods on equidistant grids yield satisfactory numerical*
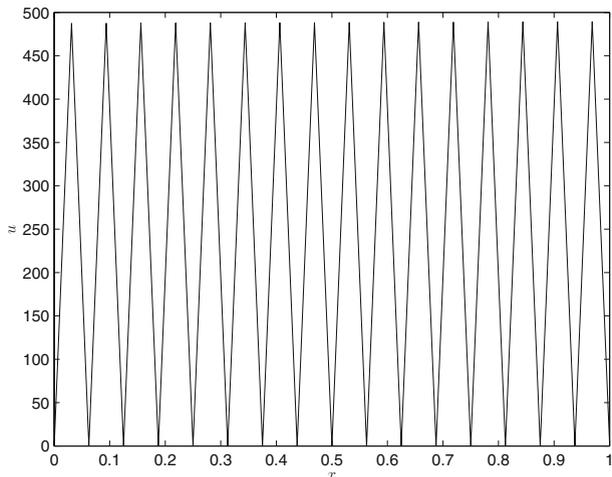
Fig. 2.1. Oscillations of the central difference scheme

solutions for singularly perturbed boundary value problems only if one uses an unacceptably large number of grid points. In this sense, *classical methods fail.*

An alternative heuristic explanation for the failure of central differencing in the above example is that when $\varepsilon \ll h$ the scheme is essentially $D^0 u_i = 0$, which implies in particular that $u_{N-2} \approx u_N = 1$, so u_{N-2} is a poor approximation to $u(x_{N-2}) \approx 0$.

This argument also shows that we would do well to avoid any difference approximation of $u'(x_{N-1})$ that uses u_N. The simplest candidate meeting this requirement is the approximation

$$u'(x_i) \approx \frac{u_i - u_{i-1}}{h} \ . \tag{2.11}$$

An inspection of the signs of the matrix entries of the earlier discrete problem, with the aim of modifying the difference scheme in order to generate an M-matrix, also motivates (2.11).

Thus for the general case where the sign of b may be positive or negative, consider the scheme

$$-\varepsilon D^+ D^- u_i + b_i D^\aleph u_i + c_i u_i = f_i \quad \text{for } i = 1, ..., N-1, \tag{2.12a}$$

$$u_0 = u_N = 0, \tag{2.12b}$$

with

$$D^\aleph = \begin{cases} D^+ & \text{if} \quad b < 0, \\ D^- & \text{if} \quad b > 0. \end{cases} \tag{2.12c}$$

This is the *simple upwind scheme.* (We saw in the Introduction that convection dominates the problem and assigns a direction to the flow; *upwind* means that the finite difference approximation of the convection term is taken on the upstream side of each mesh point.) The numerical behaviour of the upwind scheme is much better than the central scheme: see Figure 2.2.
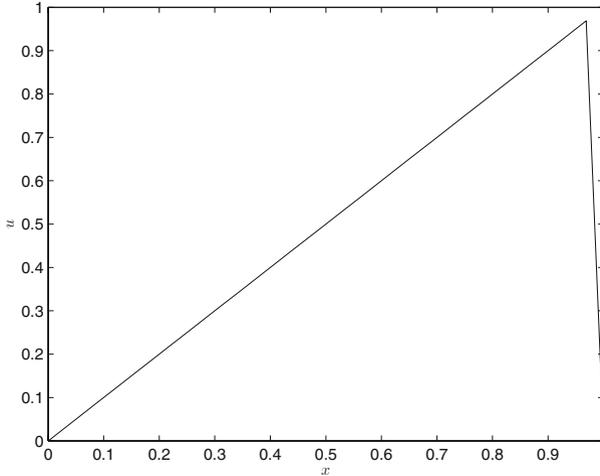
Fig. 2.2. Solution of the upwind scheme on an equidistant mesh

We now begin our analysis of the upwind scheme. Write L_h for the matrix of the scheme after eliminating u_0 and u_N. In the form (2.4), the coefficients of the discrete problem are

$$r_i = -\frac{\varepsilon}{h^2} - \frac{1}{h}\max\{0, b_i\}, \quad s_i = c_i + \frac{2\varepsilon}{h^2} + \frac{1}{h}|b_i|,$$
$$t_i = -\frac{\varepsilon}{h^2} + \frac{1}{h}\min\{0, b_i\}.$$

Now the off-diagonal matrix entries are non-positive, irrespective of the relative sizes of h and ε.

Lemma 2.11. *Assume that $b(x) \neq 0$ for all $x \in [0, 1]$. Then the coefficient matrix L_h for the upwind scheme (2.12) is an M-matrix and the upwind scheme is uniformly stable with respect to the perturbation parameter:*

$$\|u_h\|_{\infty,d} \leq C\|L_h u_h\|_{\infty,d},$$

with a stability constant C that is independent of ε and h.

Proof. For definiteness assume that $b(x) \geq \beta > 0$. We construct a suitable majorizing vector. Choose $e(x) := x$, so $Le(x) \geq \beta$. A direct computation yields $L_h e_h \geq \beta$. By Theorem 2.7 the matrix is an M-matrix and one gets the desired stability bound with stability constant $C = 1/\beta$. $\quad\square$

This stability result for the upwind scheme remains valid on arbitrary meshes. Moreover, introducing mesh analogues of the norms previously seen, one can also prove $(L_{\infty,d}, L_{1,d})$ and $(L_{\infty,d}, W_{-1,\infty,d})$ stability results which are useful when analysing the scheme on layer-adapted meshes, as will be seen later.

In ensuring the stability of the upwind scheme, we have paid a certain price in accuracy: D^+ and D^- are only $O(h)$ approximations of the first-order derivative whereas the central difference D^0 is an $O(h^2)$ approximation. The precise analysis of the consistency error and convergence behaviour of the upwind scheme that now follows is based on [KT78] and draws on the bounds of Lemma 1.8 on derivatives of the exact solution u.

Theorem 2.12. *Assume that $b > \beta > 0$ and $c \geq 0$. Then there exists a positive constant β^*, which depends only on β, such that the error of the simple upwind scheme (2.12) at the inner grid points $\{x_i : i = 1, \ldots, N-1\}$ satisfies*

$$|u(x_i) - u_i| \leq \begin{cases} Ch\left[1 + \varepsilon^{-1}\exp(-\beta^*(1-x_i)/\varepsilon)\right] & \text{if } h \leq \varepsilon, \\ Ch + C\exp(-\beta^*(1-x_{i+1})/\varepsilon) & \text{if } h \geq \varepsilon. \end{cases}$$

Proof. As for the central scheme in Section 2.1.1, the consistency error is estimated using Taylor's formula. At each grid point x_i one obtains

$$|\tau_i| := |L_h u(x_i) - f(x_i)| \leq C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon|u^{(3)}(t)| + |u^{(2)}(t)|\right) dt. \qquad (2.13)$$

The crude bound $|u^{(k)}| \leq C\varepsilon^{-k}$ combined with the stability result of Lemma 2.11 yields only $|u(x_i) - u_i| \leq Ch/\varepsilon^2$, so a more precise bound on $|u^{(k)}|$ is needed. Invoking Lemma 1.8 yields the inequality

$$|\tau_i| \leq Ch + C\varepsilon^{-2} \int_{x_{i-1}}^{x_{i+1}} \exp\left(-\beta(1-t)/\varepsilon\right) dt$$

$$\leq Ch + C\varepsilon^{-1} \sinh\left(\frac{\beta h}{\varepsilon}\right) \exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right).$$

Consider first the case when $h \leq \varepsilon$. Then $\beta h/\varepsilon$ is bounded. Now $\sinh t \leq Ct$ when t is bounded, so

$$|\tau_i| \leq Ch\left[1 + \varepsilon^{-2}\exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right)\right].$$

At first sight, this inequality seems unable to deliver the desired power of ε (viz., ε^{-1} instead of ε^{-2}) when Lemma 2.11 is applied. But if one considers the boundary value problem

$$-\varepsilon w'' + bw' + cw = C\varepsilon^{-1}\exp\left(-\frac{\beta(1-x)}{\varepsilon}\right), \quad w(0) = w(1) = 0,$$

then using the barrier function

$$w^*(x) = C\exp\left(-\frac{\beta^*(1-x)}{\varepsilon}\right)$$

where $\beta^* > \beta$, the comparison principle of Lemma 1.1 yields the estimate

$$|w(x)| \le C\exp\left(-\frac{\beta^*(1-x)}{\varepsilon}\right)$$

– where w has gained a power of ε compared with Lw! The same calculation at the discrete level, using the discrete comparison principle of Lemma 2.5, completes the proof of the theorem when $h \le \varepsilon$.

In the more difficult case $h \ge \varepsilon$, we decompose the solution as

$$u(x) = -u_0(1)\exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right) + z(x).$$

By imitating the proof of Lemma 1.8 one finds that

$$|z^{(i)}(x)| \le C\left[1 + \varepsilon^{1-i}\exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right)\right].$$

Set

$$v(x) = -u_0(1)\exp\left(-\frac{b(1)(1-x)}{\varepsilon}\right)$$

and define v_h and z_h by

$$L_h v_h = Lv \quad \text{and} \quad L_h z_h = Lz,$$

where v_h and z_h agree with v and z, respectively, at x_0 and x_N. Then

$$|u(x_i) - u_i| = |v(x_i) + z(x_i) - (v_i + z_i)| \le |v(x_i) - v_i| + |z(x_i) - z_i|.$$

For the consistency error associated with z, similarly to before one gets

$$|\tau_i(z)| \le Ch + C\sinh\left(\frac{\beta h}{\varepsilon}\right)\exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right).$$

As now $h \ge \varepsilon$, we use the inequality $\sinh t \le Ce^t$. Hence

$$|\tau_i(z)| \le Ch + C\exp\left(-\frac{\beta(1-x_{i+1})}{\varepsilon}\right).$$

The consistency error due to v must still be bounded. The definition of v gives

$$|Lv(x)| \le C\varepsilon^{-1}|v(x)|.$$

Thus

$$|(L_h v_h)_i| = |Lv(x_i)| \le C\varepsilon^{-1} \exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right).$$

Appealing again to the discrete comparison principle, one obtains

$$|v(x_i) - v_i| \le |v(x_i)| + |v_i| \le C \exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right).$$

Combining the various estimates proves the result for the case $h \ge \varepsilon$. $\square$

Remark 2.13. If the boundary layer is weaker, for instance if there is a Neumann condition at $x = 1$, then a factor ε is gained in the analysis and the conclusion is that

$$\|u - u_h\|_{\infty,d} \equiv \max_i |u(x_i) - u_i| \le Ch.$$

To get $O(h^2)$ uniformly with respect to ε for some method such as central differencing, the layer must be weaker still: not only the first-order derivative but also the second-order derivative should be bounded uniformly with respect to ε.   ♣

Theorem 2.12 shows that *outside* the boundary layer (i.e., in the interval $[0, 1 - \delta]$ for any fixed $\delta > 0$) simple upwinding gives *first-order convergence with a convergence constant independent of ε.* But inside the layer the theorem does not prove convergence, and indeed the story here is disappointing: take the example

$$-\varepsilon u'' - u' = 0, \quad u(0) = 0, \, u(1) = 1,$$

which has a boundary layer at $x = 0$. Then the simple upwind scheme yields

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

Thus for $h = \varepsilon$ one gets

$$u_1 = \frac{1/2}{1 - (1/2)^N} \quad \text{but} \quad u(x_1) = \frac{1 - e^{-1}}{1 - e^{-1/\varepsilon}},$$

so the error at this mesh point is $O(1)$. Thus one cannot expect to sharpen significantly Theorem 2.12 at the layer. Figure 2.3 on page 58 shows the typical error behaviour of the upwind scheme at the first interior grid point close to the layer, as h varies while ε is fixed; as h decreases, the error increases (because the grid point is moving from outside into the layer) and begins to decrease only when h is sufficiently small.

Several options are available for the construction of upwind schemes that achieve higher-order convergence outside the layer. (Here "upwind" means that the first-order derivative in the differential equation is approximated by a non-centred difference approximation.)

First, taking $b > 0$ for convenience, the simple upwind scheme (2.12) can be rearranged as

$$-\left(\varepsilon + \frac{b_i h}{2}\right) D^+ D^- u_i + b_i D^0 u_i + c_i u_i = f_i, \quad u_0 = u_N = 0. \qquad (2.14)$$

This resembles the central difference scheme, but the diffusion coefficient has been modified from ε to $\varepsilon + b_i h/2$. That is, simple upwinding applied to (2.10) is the same as central differencing applied to a modified version of (2.10). For $\varepsilon > b_i h/2$ the dominant diffusion is $O(\varepsilon)$, but in the more interesting case $\varepsilon < b_i h/2$ it becomes $O(b_i h/2)$. The scheme (2.14) is said to have *artificial diffusion* or *artificial viscosity*. It is the simplest example of a general strategy: add artificial diffusion to the given differential equation to *stabilize* a standard discretization method.

Too much artificial viscosity will "smear" the computed solution (that is, the computed layers are too wide – an unsurprising consequence since the layer width in the original differential equation depends on the diffusion coefficient); see also Remark 2.19. In two dimensions this effect is particularly important, so we shall continue this discussion in Part III.

Artificial diffusion can be introduced directly by means of a *fitting factor* σ, as in the following fitted upwind scheme, which generalizes (2.14):

$$-\varepsilon\sigma(q(x_i))D^+ D^- u_i + b_i D^0 u_i + c_i u_i = f_i \quad \text{for } i = 1, \ldots, N-1, \qquad (2.15\text{a})$$
$$u_0 = u_N = 0, \qquad (2.15\text{b})$$
$$\text{with} \quad q(x) := \frac{b(x)h}{2\varepsilon}. \qquad (2.15\text{c})$$

If $\sigma(q) = 1 + q$, this becomes the simple upwind scheme (2.14).

Which choices of σ will generate good upwind schemes? As part of the answer to this question, it's easy to generalize Lemma 2.11 to the following stability result.

Lemma 2.14. *Assume that $b(x) > \beta > 0$, $c \geq 0$, and $\sigma(q) > q$. Then the coefficient matrix of the fitted upwind scheme (2.15) is an M-matrix and the method is stable in the discrete maximum norm, uniformly in ε.*

The next step is to investigate the consistency error τ_i. Now

$$\tau_i = \varepsilon\sigma\left(u''(x_i) - D^+ D^- u(x_i)\right) + \varepsilon(1-\sigma)u''(x_i) + b_i\left(D^0 u(x_i) - u'(x_i)\right),$$

which leads to

$$|\tau_i| \leq C\left\{\varepsilon|\sigma(q_i)|\,h^2\,\|u^{(4)}\|_\infty + \varepsilon|1 - \sigma(q_i)|\,\|u^{(2)}\|_\infty + h^2\|u^{(3)}\|_\infty\right\}.$$

Assume that
$$|\sigma(q) - 1| \leq \min\{q, \, Mq^2\}.$$

Then
$$|\tau_i| \leq C \left\{ \left(1 + \frac{h}{\varepsilon}\right) \varepsilon h^2 ||u^{(4)}||_\infty + \varepsilon M \left(\frac{h}{\varepsilon}\right)^2 ||u^{(2)}||_\infty + h^2 ||u^{(3)}||_\infty \right\},$$

whence
$$|\tau_i| \leq C \frac{h^2}{\varepsilon^3} \left(1 + \frac{h}{\varepsilon}\right) \max_{k=2,3,4} \left\{ \varepsilon^k ||u^{(k)}||_\infty \right\}.$$

This implies

Lemma 2.15. *Suppose that*

$$|\sigma(q) - 1| \leq \min\{q, \, Mq^2\}.$$

Then for fixed ε, the consistency error of the generalized upwind scheme (2.15) is second order.

We emphasize that in the statement that the consistency error τ_i is second order, the "constant" factor depends on ε and, moreover, tends to infinity if ε tends to zero. When dealing with singularly perturbed problems, consistency error for fixed ε is sometimes called *formal consistency* or *formal accuracy*.

Examples of polynomial-type fitting factors that satisfy the assumptions of Lemmas 2.14 and 2.15 are

$$\sigma(q) = \max\{1, q\}, \quad \sigma(q) = \sqrt{1 + q^2},$$
$$\sigma(q) = 1 + q^2/(1 + q) \quad \text{(which generates Samarskiĭ's upwind scheme).}$$

A more careful analysis shows that *for fitted upwind schemes of the form (2.15), when the conditions of Lemma 2.14 and 2.15 are satisfied, then the order of convergence is two for fixed ε but in general is only one, uniformly in ε, in the region outside the layer;* see [KT78, Tob83]. That is, one observes a kind of *order reduction* that is well known in stiff initial-value problems.

To obtain *second*-order convergence that is uniform in ε outside the layer, Stoyan [Sto79] devised the scheme

$$-\varepsilon \sigma(q(x_{i-\alpha})) D^+ D^- u_i + b(x_{i-\alpha}) D^0 u_i + c(x_{i-\alpha}) u_i = f(x_{i-\alpha}),$$

with the shifted evaluation

$$x_{i-\alpha} := (i - \alpha)h \quad \text{and} \quad \alpha = \alpha(q) := (\sigma(q) - 1)/(2q),$$

where q satisfies the nonlinear equation

$$q = \frac{h}{2\varepsilon} b\big(x_i - \alpha(q)h\big).$$

Stoyan's scheme generalizes an idea of Abrahamsson, Keller and Kreiss [AKK74], who proposed the midpoint upwind scheme

$$-\varepsilon D^+ D^- u_i + b_{i-1/2} D^- u_i + c_{i-1/2} \frac{u_i + u_{i-1}}{2} = f_{i-1/2}.$$

For $\varepsilon = 0$ this is a second-order consistent approximation of the reduced problem at $x_{i-1/2}$, while the simple upwind scheme is only first-order consistent. See also [BSC+80] for schemes that are higher-order outside the layer.

So far we have examined three-point schemes. For these schemes, M-matrices are a powerful stability analysis tool. Schemes with more than three points, however, rarely yield M-matrices. This makes their stability analysis much more difficult; cf. Remark 2.10. Furthermore, schemes with more than three points are not in general inverse-monotone, which is sometimes more important in practice than higher-order accuracy.

Gushchin and Shchennikov [GS74] combine the central difference scheme with a midpoint scheme that is inverse-monotone when the central difference scheme loses this property. When $b(x) > \beta > 0$, the *Gushchin-Shchennikov scheme* for the boundary value problem (2.10) is

$$-\varepsilon \frac{u_{i+1} - u_i - u_{i-1} + u_{i-2}}{2h^2} + b_{i-1/2} D^- u_i$$

$$+ c_{i-1/2} \frac{u_i + u_{i-1}}{2} = f_{i-1/2}. \qquad (2.16)$$

The approximation used for the second-order derivative is well known (see, e.g., [For88] for half-point approximations of different orders for $u, u', \dots, u^{(4)}$) and is second order at $x_{i-1/2}$. The consistency error at $x_{i-1/2}$ is therefore second-order. This scheme is stable when $\varepsilon \leq 2b_0 h$ because the coefficient matrix is then an M-matrix.

This scheme and those below must be modified near the endpoints of the interval.

An alternative approach is to use the central scheme for the second-order derivative but higher-order one-sided approximations for the first-order derivative. For instance, the following scheme seems natural when $b > 0$:

$$-\varepsilon D^+ D^- u_i + \frac{b_i}{2h} (3u_i - 4u_{i-1} + u_{i-2}) + c_i u_i = f_i, \ u_0 = u_N = 0.$$

See [For88] for higher-order one-sided approximations.

A general scheme with a formally second-order consistent four-point approximation for u' is

$$-\varepsilon D^+ D^- u_i + \frac{b_i}{h} \left[(-\lambda + 1/2)u_{i+1} + 3\lambda u_i - (3\lambda + 1/2)u_{i-1} + \lambda u_{i-2} \right]$$

$$+ c_i u_i = f_i, \quad u_0 = u_N = 0, \qquad (2.17)$$

where λ is a free parameter. The choice $\lambda = 0$ reduces to the central scheme, while $\lambda = 1/2$ uses grid points on only one side of x_i. In fact, (2.17) is a special case of the more general five-point scheme

$$-\varepsilon D^+ D^- u_i + \frac{b_i}{h} \sum_{k=1}^{5} \alpha_k u_{i+k-3} + c_i u_i = f_i, \quad u_0 = u_N = 0, \qquad (2.18)$$

with the following conditions enforced for formal second-order consistency:

$$\alpha_1 = -\frac{1}{4} - \frac{1}{8}\alpha + \frac{3}{8}\beta, \qquad \alpha_2 = -\beta, \qquad \alpha_3 = \frac{3}{4}(\alpha + \beta)$$

$$\alpha_4 = -\alpha, \qquad \alpha_5 = \frac{1}{4} + \frac{3}{8}\alpha - \frac{1}{8}\beta.$$

One obtains (2.17) from (2.18) by taking $\alpha = \lambda - 1/2$ and $\beta = 3\lambda + 1/2$; then $\alpha_5 = 0$.

Some particular cases of (2.17) are associated by name in the engineering literature with *Atia* ($\lambda = 1/2$), *Agarwal* ($\lambda = 1/6$) and *Leonard* ($\lambda = 1/8$); see also the *LECUSSO* [Leo79b, Leo79a] and *LUDS* schemes [Gün88].

Little attention has been paid to a five-point scheme that was introduced in [GF88] and worked well — even for the Navier-Stokes equations at high Reynolds numbers. This *Goncharov-Fryazinov scheme* is related to our earlier observation that the simple upwind scheme can be regarded as the central difference scheme applied to an $O(h)$ regularization of the second-order derivative:

$$\frac{u_{i+1} - u_i}{h} = \frac{u_{i+1} - u_{i-1}}{2h} + \frac{h}{2}\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}.$$

In [GF88] an $O(h^3)$ regularization of the fourth-order derivative is used, which leads to a five-point scheme.

Unfortunately *no general mathematical theory that yields pointwise error bounds is available for higher-order upwind schemes with more than three points*. Some proofs of stability based on inverse-monotonicity are known. In some cases, the coefficient matrix of (2.17) or (2.18) is the product of two M-matrices and therefore inverse-monotone. For the scheme (2.17) we have uniform stability with respect to ε and second-order convergence in the domain outside the layer under the assumptions $\varepsilon \leq Ch$ and $1/2 \leq \lambda \leq 1/2 + \sqrt{1/3}$. See [Roo86b] for the more general scheme (2.18). In [Gün88], several related schemes (but not Goncharov-Fryazinov) are tested numerically.

In this context, it is interesting to note that recently-developed stabilized finite element methods for convection-diffusion problems – such as edge stabilization or local projection methods (see Part III) – generate five-point difference schemes for (2.10). These methods often contain user-chosen parameters. One guide to determining the values of these free parameters is to use upwinding instead of central differencing at the boundary where the layer is located, i.e., choose the parameters in such a way that the scheme avoids using the corresponding boundary value. See [RV07] for a detailed discussion.

Schemes like *LECUSSO-C*, *LSUDS-C* and *QUICK-PLUS* from [Gün88, Leo79b, Leo79a] use some form of exponential fitting, whose basic theory will be discussed in subsequent sections.

Remark 2.16. (An upwind scheme for a higher-order problem) As in Section 1.4, we consider the singularly perturbed higher-order problem

$$Lu := \varepsilon u^{(m)} + \sum_{\nu=0}^{m-1} a_\nu(x) u^{(\nu)} = f(x) \quad \text{for} \quad 0 < x < 1,$$

subject to the m homogeneous boundary conditions

$$Bu := (B_1 u, B_2 u, ..., B_m u) = 0.$$

Assume that the functions f and a_ν are sufficiently smooth and exclude turning points by the assumption that $a_{m-1}(x) \geq \alpha > 0$. Finally, assume that the hypotheses of the basic stability result of Niederdrenk and Yserentant (Theorem 1.18) are satisfied.

Let us introduce a (possibly nonequidistant) mesh

$$0 = x_0 < x_1 < ... < x_N = 1$$

and the notation

$$h_i := x_{i+1} - x_i, \quad h := \max h_i, \quad h_k(x_i) := \frac{x_{i+k} - x_i}{k}.$$

Define the difference operators

$$D_0 u_i := u_i, \quad D^\nu u_i := \frac{D^{\nu-1} u_{i+1} - D^{\nu-1} u_i}{h_\nu(x_i)} \quad \text{for } \nu = 1, \ldots, m-1.$$

Niederdrenk and Yserentant [NY83] consider the scheme

$$\varepsilon D^m u_i + a_{m-1}^h(x_i)[\theta_i D^{m-1} u_i + (1 - \theta_i) D^{m-1} u_{i+1}] + \sum_{\nu=0}^{m-2} a_\nu^h(x_i) D^\nu u_{i+1}$$
$$= f_h(x_i),$$

where the a_ν^h are approximations of the coefficients a_ν, and f_h approximates f. If $m = 2$ and $\theta = 0$, this scheme collapses to the simple upwind scheme. It is thus a natural upwind approximation of the given problem, but *on equidistant grids no conditions are known that guarantee the stability of the scheme, uniformly with respect to ε, in some appropriate norm.*

Niederdrenk and Yserentant derive conditions equivalent to stability of the discrete problem (cf. Theorem 1.18 in the continuous case) under the assumption that

$$0 \leq \theta_i \leq \min\left\{\frac{1}{\rho_i}, 1\right\}, \quad \text{where} \quad \rho_i := \frac{h_m(x_i) a_{m-1}^h(x_i)}{\varepsilon}.$$

This condition is more restrictive than is needed in practice. Gartland [Ga88] shows that uniform stability of the discrete boundary value problem follows from uniform stability of an associated discrete initial-value problem and uniform consistency of the scheme. Uniform consistency does however require exponential fitting or a special mesh or both.   ♣

2.1.3 The Concept of Uniform Convergence

We continue our study of the singularly perturbed boundary value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x), \quad u(0) = u(1) = 0, \qquad (2.19)$$

under the assumptions that $b(x) > \beta > 0$ (i.e., no turning points) and $c \geq 0$. In Section 2.1.2 we considered convergence, uniformly in ε, outside the boundary layer; we now examine convergence on the whole interval [0,1].

A difference method is called *uniformly convergent* (with respect to ε) of order $\gamma > 0$ in the discrete maximum norm $\|\cdot\|_{\infty,d}$, if there exists a constant C that is independent of ε and of the mesh, such that

$$\|u - u_h\|_{\infty,d} \leq Ch^{\gamma} \qquad (2.20)$$

for all sufficiently small h (independently of ε). Uniform consistency is defined analogously; we have already discussed finite difference stability that is uniform in ε.

More generally, a discretization method is called *uniformly convergent* (with respect to ε) of order $\gamma > 0$ in the norm $\|\cdot\|$, if there exists a constant C that is independent of ε and of the mesh, such that for all sufficiently small h (independently of ε), one has

$$\|u - u_h\| \leq Ch^{\gamma}, \qquad (2.21)$$

where u_h is the solution computed by the method.

The simple upwind scheme is not uniformly convergent in the discrete maximum norm because of its behaviour in the layer (see Figure 2.3 and Theorem 2.12). The same observation holds for most of the schemes discussed in Section 2.1.2 if the given problem exhibits a typical exponential boundary layer; on the other hand, see Remark 2.13 concerning upwinding if the layer is weaker.

Uniformly convergent schemes are interesting not just from a theoretical viewpoint. Consider upwind schemes that are not uniformly convergent: a careful examination of numerical results shows that for fixed ε, the maximum pointwise error may initially decrease, but then usually *increases* as the mesh is refined because of the boundary layer – on a coarse mesh all interior mesh points are outside the layer, but as the mesh is refined the closest interior point approaches the layer, which makes the maximum pointwise error increase – until the mesh parameter and the perturbation parameter have the same order of magnitude, when the error again begins to decrease. This runs contrary to the reasonable expectation that the error of an acceptable numerical method should *decrease* when the mesh is refined; furthermore, for problems posed in more than one dimension it is often too expensive to use an equidistant mesh whose diameter has the same order of magnitude as the perturbation parameter. Figure 2.3 exhibits this undesirable error behaviour at the first interior grid point for the simple upwind scheme applied to the example

$$-\varepsilon u'' - u' = 2x, \quad u(0) = u(1) = 0,$$

with $\varepsilon = 10^{-6}$, where the mesh width $h = 1/N$. This figure also shows the error behaviour of the uniformly convergent Il'in-Allen-Southwell scheme that we shall meet shortly.
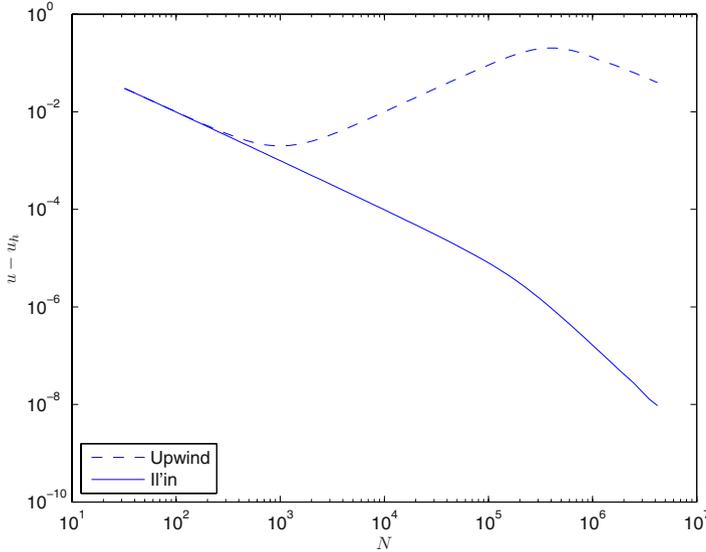
Fig. 2.3. The error at the layer for the upwind and Il'in-Allen-Southwell methods

For uniformly convergent methods on an equidistant mesh, the error bound for $\|u - u_h\|_{\infty,d}$ decreases as the mesh is refined for all $h \leq h_0$ (where the constant h_0 is independent of ε), regardless of the ratio of the parameters h and ε.

We now look for fitted upwind schemes that are uniformly convergent with respect to the discrete maximum norm $\|\cdot\|_{\infty,d}$. These schemes take the form of (2.15):

$$-\varepsilon\sigma\big(q(x_i)\big)D^+D^-u_i + b_iD^0u_i + c_iu_i = f_i, \quad u_0 = u_N = 0,$$

$$\text{where} \quad q(x) = \frac{b(x)h}{2\varepsilon} \,.$$

Miller [Mil76] derives *necessary conditions for uniform convergence.*

Theorem 2.17. *Assume that the scheme (2.15) is uniformly convergent with respect to $\|\cdot\|_{\infty,d}$. If n is a fixed positive integer and $\rho = h/\varepsilon$ is fixed, then*

$$\lim_{h \to 0} \sigma(q(x_{N-n})) = q(1) \coth q(1). \tag{2.22}$$

Proof. By virtue of Theorem 1.4 with $m = 0$, for $x \in [0,1]$ one has

$$\left| u(x) - \left[u_0(x) - u_0(1) \exp\left(-\frac{b(1)(1-x)}{\varepsilon} \right) \right] \right| \le C\varepsilon.$$

As $\rho = h/\varepsilon$ is fixed, it follows that for each fixed i one gets

$$\lim_{h \to 0} u((N-i)h) = u_0(1) \left[1 - \exp(-2q(1)i) \right].$$

Now the uniform convergence of the scheme implies that

$$\lim_{h \to 0} u_{N-i} = u_0(1) \left[1 - \exp(-2q(1)i) \right].$$

On the other hand, (2.15) yields

$$\lim_{h \to 0} | -\sigma(q(x_{N-1}))(u_N - 2u_{N-1} + u_{N-2}) + q(x_{N-1})(u_N - u_{N-2})| = 0.$$

Combining the last two equations gives (one can assume without loss of generality that $u_0(1) \ne 0$)

$$\sigma(q(1)) \left[1 - 2\exp(-2q(1)) + \exp(-4q(1)) \right] = q(1) \left[1 - 4\exp(-4q(1)) \right].$$

Finally, use

$$\frac{1 - e^{-4q}}{1 - 2e^{-2q} + e^{-4q}} = \frac{e^{2q} - e^{-2q}}{e^{2q} - 2 + e^{-2q}} = \frac{(e^q - e^{-q})(e^q + e^{-q})}{(e^q - e^{-q})^2} = \coth q$$

to obtain (2.22).   □

While Theorem 2.17 states a necessary condition for schemes of the form (2.15), the same argument applies to other three-point schemes. In [Sty03] Theorem 2.17 is generalized to other schemes and other meshes.

The obvious choice

$$\sigma(q(x)) = q(x) \coth q(x)$$

satisfies the conditions of Lemmas 2.14 and 2.15 and generates the *Il'in-Allen-Southwell scheme* [AS55, Il'69]. This scheme is uniformly stable and second-order consistent for fixed ε. Since

$$\coth z \to \begin{cases} 1 & \text{as } z \to \infty, \\ -1 & \text{as } z \to -\infty, \end{cases}$$

the scheme shifts automatically to the simple upwind scheme as $h/\varepsilon \to \infty$. Its consistency error is

$$\tau_i = -\varepsilon[q(x_i)\coth q(x_i) - 1]D^+D^-u(x_i) - \varepsilon[D^+D^-u(x_i) - u''(x_i)] +$$
$$+ b_i \left(D^0u(x_i) - u'(x_i)\right).$$

Now

$$c_1 \frac{z^2}{z+1} \le z\coth z - 1 \le c_2 \frac{z^2}{z+1} \quad \text{and} \quad \varepsilon \frac{(h/\varepsilon)^2}{h/\varepsilon + 1} = \frac{h^2}{h + \varepsilon},$$

so we expect to lose an order of convergence for small values of ε, and in fact it turns out that the scheme is first-order uniformly convergent.

Theorem 2.18. *Assume that $b(x) > \beta > 0$. Then the Il'in-Allen-Southwell scheme is first-order uniformly convergent in the discrete maximum norm:*

$$\|u - u_h\|_{\infty,d} \le Ch.$$

Proof. The argument is like the one used for Theorem 2.12. In particular one relies on the splitting $u = v + z$, where v is a boundary layer function and the bound on $|z^{(j)}|$ has a factor ε^{1-j} (which is better than the factor ε^{-j} that appears if we bound $|u^{(j)}|$).

First consider $|z(x_i) - z_i|$. The corresponding consistency error satisfies

$$|\tau_i| \le C \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon|z^{(3)}| + |z^{(2)}|\right) dt$$

$$\le Ch + C\varepsilon^{-1} \int_{x_{i-1}}^{x_{i+1}} \exp(-\beta(1-t)/\varepsilon)dt$$

$$\le Ch + C\sinh\frac{\beta h}{\varepsilon} \exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right).$$

An application of the discrete comparison principle gains a power of ε, as in the proof of Theorem 2.12. We now have

$$|z(x_i) - z_i| \le Ch + C\varepsilon \sinh\frac{\beta h}{\varepsilon} \exp\left(-\frac{\beta(1-x_i)}{\varepsilon}\right) \quad \text{for } i = 1, \ldots, N-1.$$

If $\varepsilon \le h$, this gives immediately $|z(x_i) - z_i| \le Ch$; if $h \le \varepsilon$, use the inequality $\sinh t \le Ct$ for bounded $t > 0$ to get the desired estimate.

It is more difficult to bound $|v(x_i) - v_i|$. A direct computation gives

$$Lv = -\frac{b(1)}{\varepsilon}[b(1) - b(x)]v + c(x)v,$$

and at the grid points

$$L_h v = -\frac{2b(x)\sinh q(1)\sinh[q(1) - q(x)]}{h\sinh q(x)}v + c(x)v.$$

Using the consistency error and a barrier function, some careful manipulations [KT78, pp. 1034–1035] yield

$$|v(x_i) - v_i| \leq C\frac{h^2}{h + \varepsilon} \leq Ch.$$

This completes the proof of Theorem 2.18. $\square$

Remark 2.19. If $c(x) \equiv 0$ with $b(x)$ and $f(x)$ both constant, the Il'in-Allen-Southwell scheme is exact (i.e., $u_i = u(x_i)$ for all i). But even in this case the scheme has some artificial viscosity because $\sigma(q(x)) > 1$. It is therefore false to assert (as is sometimes claimed) that minimal artificial viscosity leads to minimal numerical error. See Remark III.2.2 and [Tob95] for further discussion of this point.   ♣

Remark 2.20. An examination of the behaviour of the Il'in-Allen-Southwell scheme scheme when applied to the example

$$-\varepsilon u'' + u' = x, \quad u(0) = u(1) = 0,$$

shows that in the region outside the layer the order of uniform convergence is only one.   ♣

Remark 2.21. If instead of from (2.15) one starts from the fitted scheme

$$-\varepsilon \sigma_i^* D^+ D^- u_i + b_i D^- u_i + c_i u_i = f_i, \tag{2.23}$$

then, analogously to Theorem 2.17, one obtains

$$\lim_{h \to 0} \sigma_n^* = B(2q(1)) \quad \text{for fixed } h/\varepsilon$$

as a necessary condition for uniform convergence, where $B(z) := z/(e^z - 1)$ is the *Bernoulli function*.

Farrell [Far83, Far88] derives *sufficient conditions* for uniform convergence of schemes written in the form (2.23). They show that schemes whose coefficients are close to the coefficients of the Il'in-Allen-Southwell scheme are also uniformly convergent. Furthermore, these sufficient conditions imply that for uniform convergence, exponential coefficients are needed only in the boundary layer; compare Theorem 2.12 and the results on exponentially fitted finite element methods in Section 2.2.5.   ♣

Remark 2.22. (The Scharfetter-Gummel scheme) The drift-diffusion equations, which are used to model currents in semiconductor devices, comprise a coupled system of three partial differential equations. In the easiest case two of these simplify to

$$-(e^{-\psi} u')' = f$$

with some boundary conditions. The potential ψ can have interior layers where its gradient is extremely large.

Scharfetter and Gummel developed a special difference scheme for this problem which is widely used; this special scheme turns out to be merely a natural reformulation of the Il'in-Allen-Southwell scheme for equations written in the form above [Roo86a, Gar93].   ♣

We now describe some alternative ways of constructing uniformly convergent difference schemes.

Consider first the standard derivation (see, e.g., [Mar77, Chapter 2.1]) of an *exact scheme* for the boundary value problem (2.19). Introduce the formal adjoint operator M^* of $Mw := -\varepsilon w'' + bw'$, viz.,

$$M^*v := -\varepsilon v'' - (bv)'.$$

Then for smooth v and w with $v(0) = w(0) = v(1) = w(1) = 0$, one has the identity

$$\int_0^1 (Mv)w\,dx = \int_0^1 v(M^*w)\,dx.$$

Let g_i be the *local Green's function* of M^* with respect to the point x_i, i.e.,

$$M^*g_i = 0 \quad \text{in } (x_{i-1}, x_i) \cup (x_i, x_{i+1}),$$
$$g_i(x_{i-1}) = g(x_{i+1}) = 0,$$
$$\varepsilon\left[g_i'(x_i^-) - g_i'(x_i^+)\right] = 1.$$

Now

$$\int_{x_{i-1}}^{x_{i+1}} (Mu)g_i\,dx = \int_{x_{i-1}}^{x_{i+1}} (f - cu)g_i\,dx,$$

and an integration by parts yields the identity

$$-\varepsilon g_i'(x_{i-1})u_{i-1} + u_i + \varepsilon g_i'(x_{i+1})u_{i+1} = \int_{x_{i-1}}^{x_{i+1}} (f - cu)g_i\,dx. \qquad (2.24)$$

The difference scheme whose i^{th} equation is (2.24) is exact if $c(x) \equiv 0$ (or if from the beginning we replace M by the full operator L).

In general one cannot evaluate each g_i' exactly, so a further approximation is needed to convert (2.24) to a serviceable scheme. The simplest approach is to consider b and f as constants b_i and f_i on the interval (x_{i-1}, x_{i+1}), which allows the explicit computation of g_i, and to use the quadrature rule

$$\int_{x_{i-1}}^{x_{i+1}} (f - cu)g_i\,dx \approx (f_i - c_iu_i)\int_{x_{i-1}}^{x_{i+1}} g_i\,dx.$$

This generates the scheme

$$-\frac{e^{\rho_i} - 1}{e^{\rho_i} - e^{-\rho_i}}u_{i-1} + u_i - \frac{1 - e^{-\rho_i}}{e^{\rho_i} - e^{-\rho_i}}u_{i+1} = (f_i - c_iu_i)\frac{h}{b_i}\frac{e^{\rho_i} - 1}{e^{\rho_i} + 1},$$

where $\rho_i := b_ih/\varepsilon$. One can write this in the form

$$\alpha_iu_{i-1} + \beta_iu_i + \gamma_iu_{i+1} = f_i,$$

with

$$\alpha_i = -2\varepsilon q_i \frac{e^{2q_i}}{e^{2q_i} - 1}, \quad \beta_i = -(\alpha_i + \gamma_i) \quad \text{and} \quad \gamma_i = -2\varepsilon q_i \frac{1}{e^{2q_i} - 1}.$$

This is again the Il'in-Allen-Southwell scheme. The coefficients could also be expressed in terms of the Bernoulli function mentioned in Remark 2.21.

Remark 2.23. The above derivation of the Il'in-Allen-Southwell scheme inspires two modifications that might reasonably be expected to yield superior schemes.

First, it seems better to use separate constants on (x_{i-1}, x_i) and (x_i, x_{i+1}) to approximate b and f. We thus define

$$b_j := [b(x_{j-1}) + b(x_j)]/2 \quad \text{on } (x_{j-1}, x_j) \text{ for each } j.$$

Using this approximation and the quadrature rule

$$\int_{x_{i-1}}^{x_{i+1}} a g_i \, dx \approx \frac{a(x_{i-1}) + a(x_i)}{2} \int_{x_{i-1}}^{x_i} g_i \, dx + \frac{a(x_i) + a(x_{i+1})}{2} \int_{x_i}^{x_{i+1}} g_i \, dx,$$

generates the *El-Mistikawy-Werle scheme:*

$$-\frac{\varepsilon}{h^2}(r_i^- u_{i-1} + r_i^c u_i + r_i^+ u_{i+1}) + q_i^- c_{i-1} u_{i-1} + q_i^c c_i u_i + q_i^+ c_{i+1} u_{i+1}$$
$$= q_i^- f_{i-1} + q_i^c f_i + q_i^+ f_{i+1},$$

where

$$
\begin{aligned}
r_i^- &= \rho_i^- \exp(-\rho_i^-)/[1 - \exp(-\rho_i^-)], & r_i^+ &= \rho_i^+/[1 - \exp(-\rho_i^+)], \\
r_i^c &= -(r_i^- + r_i^+), & q_i^- &= (1 - r_i^-)/(2\rho_i^-), \\
q_i^+ &= (r_i^+ - 1)/(2\rho_i^+), & q_i^c &= q_i^- + q_i^+, \\
\rho_i^- &= -(b_i + b_{i-1})h/(2\varepsilon), & \rho_i^+ &= -(b_i + b_{i+1})h/(2\varepsilon).
\end{aligned}
$$

Second, one could start from the original differential operator L instead of the simplified operator M, then introduce the corresponding local Green's function and apply a quadrature rule as before. This yields a variant of the El-Mistikawy-Werle scheme that is a scheme with *complete exponential fitting*, because the exponentials used depend on *all* terms of the differential operator L. When only some terms of the differential operator are used — as in the derivation of (2.24) — this is called *partial exponential fitting*.   ♣

An alternative way of deriving uniformly convergent schemes is the *exact solution of comparison problems with frozen coefficients*. Define the piecewise constant approximation of a given continuous function d on a given grid by

$$\bar{d}(x) = \frac{d(x_{i-1}) + d(x_i)}{2} \quad \text{for} \quad x \in (x_{i-1}, x_i).$$

Later d may be b, c or f. Consider the comparison problem

$$\bar{L}w := -\varepsilon w'' + \bar{b}w' + \bar{c}w = \bar{f}, \quad w(0) = w(1) = 0, \tag{2.25}$$

for the boundary value problem (2.19). Using Green's functions, one can see that the solution of (2.25) is differentiable and piecewise twice differentiable.

Lemma 2.24. *For all sufficiently small h (independently of ε), the boundary value problem (2.25) has a unique solution w. Furthermore,*

$$\|u - w\|_\infty + \varepsilon\|(u - w)'\|_\infty \le Ch. \tag{2.26}$$

Proof. For sufficiently small h, one can show that Theorems 1.7 and 1.13 hold true for (2.25), as does Theorem 1.18. Consequently a unique solution w exists. For the difference $u - w$, we have

$$\bar{L}(u - w) = \bar{L}u - Lu + f - \bar{f} = (\bar{b} - b)u' + (\bar{c} - c)u + f - \bar{f},$$
$$(u - w)(0) = (u - w)(1) = 0.$$

An application of the stability estimate of Theorem 1.7 yields

$$\|u - w\|_\infty + \varepsilon\|(u - w)'\|_\infty \le Ch\left\{\|u'\|_{L_1} + \|u\|_{L_1} + 1\right\}.$$

Now (2.26) follows from the *a priori* bounds of Theorems 1.7 and 1.13. □

The bound of Lemma 2.24 is valid not only at the grid points but on all of $[0, 1]$.

The comparison problem (2.25) is equivalent to a difference scheme, as we now show. Define adapted spline functions (*L-splines*) ϕ_i and ψ_i for each i by

$$\bar{L}\varphi_i = 0 \quad \text{on each mesh subinterval, with } \varphi_i(x_j) = \delta_{ij}, \tag{2.27}$$
$$\bar{L}\psi_i = 1 \quad \text{on } (x_{i-1}, x_i), \ \psi_i(x_{i-1}) = \psi_i(x_i) = 0, \ \psi_i \equiv 0 \text{ off } [x_{i-1}, x_i].$$

Thus supp $\varphi_i = [x_{i-1}, x_{i+1}]$ and supp $\psi_i = [x_{i-1}, x_i]$. The solution w of (2.25) can be represented as

$$w(x) = \sum_{i=1}^{N-1} u_i\varphi_i(x) + \sum_{i=1}^{N} f_i\psi_i(x).$$

Here $u_i := w(x_i)$ and f_i is the restriction of $\bar{f}$ to the interval (x_{i-1}, x_i). The property $w'(x_i^-) = w'(x_i^+)$ for $i = 1, \ldots, N - 1$, expressed in terms of the L-splines, gives the three-point difference scheme

$$\varphi_{i-1}'(x_i^-)u_{i-1} + [\varphi_i'(x_i^-) - \varphi_i'(x_i^+)]u_i - \varphi_{i+1}'(x_i^+)u_{i+1} = \tag{2.28}$$
$$- \psi_i'(x_i^-)f_i + \psi_{i+1}'(x_i^+)f_{i+1}.$$

If the splines φ_i are known, then it is possible to compute the ψ_i from

$$\psi_i(x) = [1 - \varphi_{i-1}(x) - \varphi_i(x)]/c_i \quad \text{for} \quad x \in [x_{i-1}, x_i].$$

An explicit computation of the L-splines and of their derivatives produces the El-Mistikawy-Werle scheme from (2.28).

Remark 2.25. (i) Two independent proofs in [HMO80] and [BSC81] show that the El-Mistikawy-Werle scheme is *uniformly second-order convergent* in the discrete maximum norm when $c(x) \equiv 0$. Both proofs use finite-difference techniques like those in Theorem 2.18, with large amounts of detailed estimation. Simpler methods are used to prove the same result, without the restriction that $c(x) \equiv 0$, in [OS86] (by means of finite elements) and [Gar87] (in a HODIE framework). We present the HODIE technique in Section 2.1.4, while the finite element approach will be discussed in Section 2.2. A slightly different second-order scheme is presented in [HMMR95].

(ii) One might think that the second-order convergence in (i) contradicts the well-known fact that uniform consistency of order $\alpha > 1$ forbids the difference operator to be of positive type (i.e., non-positive offdiagonal entries) – see [KT78] for the one-dimensional case, [Lu95] for the general case. But uniform consistency in $\| \cdot \|_{\infty,d}$ is not necessary for uniform convergence.

(iii) Starting from the comparison problem

$$-\varepsilon w'' + \bar{b}w' = \bar{f} - \bar{c}\bar{w}, \quad w(0) = w(1) = 0,$$

it is possible to derive a scheme with partial exponential fitting.   ♣

Remark 2.26. (Collocation with exponential splines) Exponential splines are also useful in finite element and collocation methods. Unlike finite element methods, collocation methods with exponential splines are typically used to generate finite difference schemes (see [SU91], for instance). This collocation technique seems to be restricted to problems in one-dimensional domains. ♣

Remark 2.27. (Approximation of derivatives) It is important to note that (2.26) gives also the opportunity, by computing $\varepsilon w'$, of obtaining a uniformly accurate approximation of the ε-weighted derivative. In this context, observe that *on a uniform mesh the weighted derivative cannot be approximated accurately using standard difference approximations based on nodal values, even if the nodal values are exact.* This is so because a linear interpolant is a poor approximation of the layer on the interval $[0, h]$ when $\varepsilon = h$. One can see this explicitly in the example

$$-\varepsilon u'' - u' = 1, \quad u(0) = u(1) = 0,$$

where a direct calculation for $\varepsilon = h$ yields

$$\lim_{h \to 0} \varepsilon \left| \frac{u(h) - u(0)}{h} - u'(0) \right| = 1/e.$$

Later we shall discuss alternative approaches based on finite elements and the use of layer-adapted meshes.   ♣

Uniformly convergent schemes on special meshes will be examined in Section 2.4.

2.1.4 Uniformly Convergent Schemes of Higher Order

In the early 1980s some Russian authors [Ale81, Eme82] constructed uniformly convergent schemes of arbitrarily high order using the exact difference scheme of (2.24). We describe in this section an alternative way of generating high-order schemes for the problem (2.19): the HODIE (High Order Differences with Identity Expansion) framework of Doedel [Doe78] and Lynch and Rice [LR80].

To begin with, we state a generalization of Lemma 2.24. Let k be a fixed non-negative integer. Given a smooth function g, let $\bar{g}$ denote a piecewise polynomial approximation of g that satisfies

$$\|g - \bar{g}\|_\infty \le Ch^{k+1}. \tag{2.29}$$

Define the two comparison problems

$$\bar{L}w =: -\varepsilon w'' + \bar{b}w' + \bar{c}w = \bar{f}, \quad w(0) = w(1) = 0, \tag{2.30}$$

and

$$\hat{L}w =: -\varepsilon w'' + \bar{b}w' + cw = f, \quad w(0) = w(1) = 0. \tag{2.31}$$

By imitating the proof of Lemma 2.24, one can prove the following result:

Lemma 2.28. *For all sufficiently small h (independently of ε), the boundary value problems (2.30) and (2.31) each have a unique solution. Furthermore, for both problems one has*

$$\|u - w\|_\infty + \varepsilon\|(u - w)'\|_\infty \le Ch^{k+1},$$

where u denotes the solution of (2.19).

Doedel, Lynch and Rice construct difference approximations of the form

$$\alpha_{i,-1}u_{i-1} + \alpha_{i,0}u_i + \alpha_{i,1}u_{i+1} = \sum_{j=1}^{J} \beta_{ij}f(\xi_{ij}), \quad u_0 = u_N = 0, \tag{2.32}$$

for second-order boundary value problems. Such schemes are called *compact* because they use three discretization points for a second-order problem (more generally, $2m+1$ points for a problem of order $2m$). The points ξ_{ij} are auxiliary evaluation or *HODIE points* that lie between x_{i-1} and x_{i+1}. In the special case that the ξ_{ij} are mesh points, such schemes are known as OCI (operator compact implicit) schemes.

The coefficients $\alpha_{i,-1}, \alpha_i, \alpha_{i,1}$ and β_{ij} are chosen in non-singularly perturbed problems to make the scheme locally exact on (x_{i-1}, x_{i+1}) for polynomials of degree at most n (say). That is,

$$\alpha_{i,-1}s_l(x_{i-1}) + \alpha_{i,0}s_l(x_i) + \alpha_{i,1}s_l(x_{i+1}) = \sum_{j=1}^{J} \beta_{ij}(Ls_l)(\xi_{ij}) \tag{2.33}$$

for a basis $\{s_l\}$ of the space of polynomials of degree at most n. Together with the normalization condition

$$\sum_{j=1}^{J} \beta_{ij} = 1,$$

this leads to a local linear system that determines the α and β. One can show that the consistency order is $O(h^{n-1})$ and that all $2J+3$ free parameters can be determined in such a way that one obtains an $O(h^{2J})$ scheme.

For a singularly perturbed boundary value problem such as (2.19), however, it is not enough to require exactness of the scheme only for certain polynomials. Gartland [Gar87] introduces *exponentially fitted HODIE schemes*, which are locally exact on a collection of functions of the type

$$\left\{ 1, x, \ldots, x^p, \exp\left(\frac{1}{\varepsilon}\int_x^1 b\right), \ x\exp\left(\frac{1}{\varepsilon}\int_x^1 b\right), \ldots, x^{p-1}\exp\left(\frac{1}{\varepsilon}\int_x^1 b\right) \right\}. \tag{2.34}$$

Remark 2.29. The integrals in (2.34) cannot always be evaluated exactly, but by using the comparison problem (2.31) and applying Lemma 2.28, we can first simplify b by approximation and then apply the HODIE technique.   ♣

If one chooses $J = 1$ and $\xi_{i1} = x_i$, and requires exactness on the family $\{1, x, \exp(b(x_i)x/\varepsilon)\}$, this generates the Il'in-Allen-Southwell scheme.

Remark 2.30. This idea of exactness on polynomials and exponentials can be applied also to non-compact schemes. For instance, the schemes LECUSSO-C and QUICK-PLUS are derived in this way [Gün88] from the four-point scheme (2.17). But no convergence theory is available for non-compact schemes.   ♣

Gartland chooses $J = 2p - 1$, where p is any positive integer, and works with equally spaced auxiliary points:

$$\xi_{i1} = x_i \quad \text{for } p = 1,$$
$$\xi_{ij} = x_{i-1} + \frac{j-1}{p-1}h, \ j = 1, \ldots, 2p - 1 \quad \text{for } p = 2, 3, \ldots \tag{2.35}$$

One must then show that the remaining $2p + 1$ parameters are uniquely determined by the condition of exactness on the family (2.34). We state without proof the main result [Gar87]:

Theorem 2.31. *Let the positive integer p be given. Construct an exponentially fitted HODIE scheme based on (2.32), (2.35) and exactness on the family (2.34), where the coefficients in (2.19) are approximated by piecewise polynomials of degree at most $p - 1$. If b, c and f are sufficiently smooth, then for h sufficiently small (independently of ε), the finite difference scheme generated is well defined and uniformly stable, and is uniformly convergent of order $O(h^p)$ in the discrete maximum norm.*

Example 2.32. Take $p = 2$. Then to invoke Theorem 2.31, one must approximate b by piecewise linears. This generates a scheme of the form

$$\alpha_{i,-1}u_{i-1} + \alpha_{i,0}u_i + \alpha_{i,1}u_{i+1} = \beta_{i,1}f(x_{i-1}) + \beta_{i,2}f(x_i) + \beta_{i,3}f(x_{i+1})$$

that is exact on the family

$$\left\{ 1, x, x^2, \exp\left(\frac{1}{\varepsilon}\int_x^1 \bar{b}\right), x\exp\left(\frac{1}{\varepsilon}\int_x^1 \bar{b}\right) \right\},$$

and is related to the El-Mistikawy-Werle scheme.   ♣

In [CLM95] the HODIE approach is applied to generate high-order methods for the reaction-diffusion problem

$$-\varepsilon u'' + c(x)u = f(x), \quad u(0) = u(1) = 0.$$

2.1.5 Linear Turning-Point Problems

As in Section 1.2, we begin with the case of an isolated first-order turning point:

$$Lu := -\varepsilon u'' + xb(x)u' + c(x)u = f(x) \quad \text{in } (-1,1), \qquad (2.36a)$$
$$u(-1) = u(1) = 0, \qquad (2.36b)$$

under the assumptions that

$$b(x) \neq 0 \quad \text{on } [-1,1] \quad \text{and} \quad c(x) \geq c_0 > 0. \qquad (2.36c)$$

If b is positive, then u has two boundary layers and the El-Mistikawy-Werle scheme for (2.36) is first-order uniformly convergent [BHK84, Theorem 3.2]. The proof exploits the fact that the scheme can be generated by a comparison problem, as described in Section 2.1.3; with a suitable barrier function one can bound the difference between u and the solution of this comparison problem by Ch.

If b is negative, then u has an interior layer and we need a more specialized approach. Set $b^*(x) = xb(x)$ for all x. Let $\bar{L}$ and w be as in (2.25), with $\bar{b}$ replaced by $\bar{b}^*$. The comparison principle yields the stability estimate

$$\|v\|_\infty \leq C\|\bar{L}v\|_\infty$$

for all functions v that satisfy $v(-1) = v(1) = 0$. Then (2.25), (2.36a) and the stability estimate imply that

$$\|u - w\|_\infty \leq C\left\{ \|(b^* - \bar{b}^*)u'\|_\infty + \|c - \bar{c}\|_\infty\|u\|_\infty + \|f - \bar{f}\|_\infty \right\}.$$

Now the *a priori* bound of Lemma 1.12 gives

$$\|u'\|_\infty = O\left(\varepsilon^{(\lambda-1)/2}\right), \quad \text{where } \lambda = -b(0)/c(0).$$

Thus when $\varepsilon = h^2$,

$$\|(b^* - \bar{b}^*)u'\|_\infty = O(h^\lambda).$$

In numerical experiments, precisely this rate of convergence is observed if the El-Mistikawy-Werle scheme is used to solve (2.36a) with $\varepsilon = h^2$ when an interior cusp layer is present. If λ is close to zero, the convergence is very slow.

Remark 2.33. Farrell [Far88] studies sufficient conditions for the uniform convergence of difference schemes applied to a turning-point problem with an interior cusp layer. For schemes of the type

$$-\varepsilon_i^\pm D^+ D^- u_i + \beta_i b_i^* D^\pm u_i + \gamma_i c_i u_i = f_i,$$

he proves that

$$|u(x_i) - u_i| \leq Ch^{\min\{\lambda,1\}}$$

when

$$|\beta_i b_i^* - b^*(x_i)| \leq Ch, \quad |\gamma_i c_i - c(x_i)| \leq Ch \quad \text{and } |\varepsilon_i^\pm - \varepsilon| \leq Ch.$$

Many schemes, including simple upwinding, Samarskiĭ's scheme, and the Il'in-Allen-Southwell scheme satisfy these conditions. See also [CL93].   ♣

When $0 < \lambda < 1$, the deterioration of the uniform convergence rate for the El-Mistikawy-Werle scheme can be circumvented in the following way. Approximate b (not b^*) using piecewise constants. Then

$$|(b^* - \bar{b}^*)(x)| = |x(b - \bar{b})(x)| \leq Ch|x|,$$

and it follows that

$$\|(b^* - \bar{b}^*)u'\|_\infty \leq Ch \max_{x\in[-1,1]} \left\{|x|\,(x^2 + \varepsilon)^{\frac{\lambda-1}{2}}\right\} \leq Ch.$$

This proves

Lemma 2.34. *Let the comparison problem for the turning-point problem (2.36) be constructed by replacing $b, c,$ and f by piecewise $O(h)$ approximations. Then in the case of an interior cusp layer, the error between the solution u of (2.36) and the solution w of the comparison problem satisfies*

$$\|u - w\|_\infty \leq Ch.$$

To solve the problem

$$-\varepsilon w'' + x\bar{b}w' + \bar{c}w = \bar{f}, \quad w(-1) = w(1) = 0,$$

one could in theory again use L-splines, but then parabolic cylinder functions are needed to represent the basic splines; see [FG88] for details. Alternatively, it is possible to combine the approximation idea of Lemma 2.34 with an iterative process to achieve higher-order convergence [RV93].

Example 2.35. We have discussed b positive and b negative with the general assumption that $c(x) \geq c_0 > 0$. If $c \equiv 0$, then some eigenvalue of the differential operator may tend to zero as $\varepsilon \to 0$: the operator is unstable [dG76]. Consequently any standard discretization will fail. If one takes, for instance, the boundary value problem

$$-\varepsilon u'' + (x - 1/2)u' = x - 1/2, \quad u(0) = -1/2, \, u(1) = 1/2,$$

whose exact solution is $u(x) = x - 1/2$, then applying the simple upwind scheme with $N = 128$ and $\varepsilon = 10^{-3}, 10^{-4}$ yields two different discrete solutions that are both wrong; the error at the interior meshpoints is of order $O(1)$ in both cases.   ♣

Some authors have considered the construction of uniformly convergent schemes for the singularly perturbed boundary value problem

$$-\varepsilon u'' + bu' + cu = f, \quad u(0) = u(1) = 0,$$

where $c(x) \geq c_0 > 0$, but *without any assumption on b,* so arbitrary turning points are allowed. To ensure some kind of stability, further condition(s) are often added.

Thus Niijima [Nii84] proposes the scheme

$$-\frac{\varepsilon}{h^2}(\zeta_{i+1}u_{i+1} - 2\zeta_i u_i + \zeta_{i-1}u_{i-1}) + c_i u_i = f_i, \quad u_0 = u_N = 0, \qquad (2.37)$$

with $q_i = b_i h/2\varepsilon$ and $\zeta_i = q_i \coth q_i$. It is closely related to the Il'in-Allen-Southwell scheme. See Section 2.1.6 for its motivation in a nonlinear setting. Niijima assumes additionally that

$$c(x) - |b'(x)| \geq \delta > 0 \quad \text{on } (0,1).$$

Numerical results generated by this scheme are occasionally unconvincing.

Stynes and O'Riordan [SO87] present the scheme

$$-\frac{\varepsilon}{h^2}[\theta_{i+1}u_{i+1} - (\theta_{i+1} + \theta_i)u_i + \theta_i u_{i-1}] + c_i u_i = f_i, \qquad (2.38)$$

with $\rho_i = b_i h/\varepsilon$ and

$$\theta_i = \theta(\rho_i), \quad \theta(x) = \begin{cases} \dfrac{x}{1 - e^{-x}} & \text{for } x \neq 0, \\ 1 & \text{for } x = 0. \end{cases}$$

They assume that

$$c(x) - b'(x)/2 \geq \delta > 0 \quad \text{on } (0,1).$$

The scheme can be easily motivated in a finite element framework, so we return to it in Section 2.2.

Theorem 1.13 plays a vital rôle in the convergence analysis of both schemes, which is quite complicated, and culminates in

Theorem 2.36. *Let $u_h(x)$ be the piecewise linear function that interpolates at each mesh point to the discrete solution computed by the scheme (2.37) or (2.38). Then under the respective assumptions made above, one has the L_1-norm uniform convergence result*

$$\int_0^1 |u(x) - u_h(x)|dx \leq Ch.$$

No uniform convergence result is known in the discrete maximum norm under the same hypotheses.

2.1.6 Some Nonlinear Problems

In the previous subsection we saw that precise knowledge about the asymptotic behaviour of the solution assists in the construction of schemes that are uniformly convergent in the maximum norm; otherwise only weaker stability or convergence results can be proved. This is also true of some nonlinear problems, but the nonlinear world is more complicated.

Consider first the semilinear problem

$$-\varepsilon u'' + b(x)u' = f(x, u) \quad \text{for } x \in (0, 1), \tag{2.39a}$$
$$u(0) = A,\, u(1) = B, \tag{2.39b}$$

assuming

$$b(x) > \beta > 0, \quad 0 < m \leq f_u \leq M.$$

Then estimates similar to Theorem 1.7 are valid. Using local Green's functions, which generated (2.24) in the linear case, one can construct a uniformly convergent scheme [BS90].

Remark 2.37. (The nonexistence of uniformly convergent fitted schemes) If we consider the semilinear problem

$$-\varepsilon u'' = f(x, u) \quad \text{for } x \in (0, 1), \quad u(0) = A,\, u(1) = B, \tag{2.40}$$

then under the same assumptions for f as above, the construction of a uniformly convergent fitted scheme is possible [BS89]. But relatively simple examples are presented in [FMOS98] for which it is proved that in a certain class of fitted schemes one cannot achieve uniform convergence; this is true for instance if $f(x, u) = u + u^2$. Here the use of layer-adapted meshes gives uniform convergence in the maximum norm [FMOS01] under certain conditions.   ♣

Consider now difference schemes for the quasilinear singularly perturbed boundary value problem

$$-\varepsilon u'' + b(u)u' + c(x, u) = 0 \quad \text{for } x \in (0, 1), \tag{2.41a}$$
$$u(0) = A,\, u(1) = B, \tag{2.41b}$$

under the hypothesis that

$$\frac{\partial c}{\partial s}(x, s) \geq \mu > 0, \quad \text{for } x \in (0, 1) \text{ and all } s \in R. \tag{2.41c}$$

From Section 1.3 we know that (2.41) has a unique solution. This solution u is bounded, uniformly with respect to ε, in the maximum norm; u' is uniformly bounded in the L_1 norm and (2.41) is uniformly L_1-stable – see Theorem 1.16.

Enquist and Osher [Osh81] introduced a well-known scheme for discretizing conservation laws ([LeV90] provides a good introduction to this topic) which turns out to be useful also for singularly perturbed boundary value problems. Set $e(u) := \int^u b(s)ds$. Motivated by the simple upwind idea

$$\frac{d}{dx} e(u)|_{x=x_i} \approx \begin{cases} \frac{1}{h}[e(u_i) - e(u_{i-1})] = \frac{1}{h}\int_{u_{i-1}}^{u_i} b(s)ds & \text{if } b > 0, \\ \frac{1}{h}[e(u_{i+1}) - e(u_i)] = \frac{1}{h}\int_{u_i}^{u_{i+1}} b(s)ds & \text{if } b < 0, \end{cases}$$

they set $b_+(s) = \max\{b(s), 0\}$, $b_-(s) = \min\{b(s), 0\}$ and define the *Enquist-Osher scheme*

$$-\frac{\varepsilon}{h^2}D^+D^-u_i + \frac{1}{h}\left(\int_{u_{i-1}}^{u_i} b_+(s)ds + \int_{u_i}^{u_{i+1}} b_-(s)ds\right)$$
$$+ c(x_i, u_i) = 0, \quad u_0 = A, \ u_N = B. \tag{2.42}$$

This is a special case of three-point schemes in *conservation form*:

$$-\frac{\varepsilon}{h^2}D^+D^-u_i + \frac{1}{h}\left[g(u_{i+1}, u_i) - g(u_i, u_{i-1})\right] + c(x_i, u_i) = 0, \tag{2.43}$$

where $g(v, w)$ is the *numerical flux*. For the Enquist-Osher scheme,

$$g(v, w) = \int_0^v b_-(s)ds + \int_0^w b_+(s)ds. \tag{2.44}$$

Another example is the well-known *Lax-Friedrichs scheme*, where

$$g(v, w) := \frac{1}{2}[e(v) + e(w) + \lambda(w - v)],$$

with a free parameter λ.

Consistency of the scheme requires $\partial_1 g + \partial_2 g = e'$, so the standard consistency assumption for the general form (2.43) is that

$$g(v, v) = e(v). \tag{2.45}$$

As in the linear case, we need additional tools to investigate stability. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a nonlinear mapping.

Definition 2.38. *(i) F is a Z-function if for each i the mapping*

$$v_j \mapsto \big(F(v_1, ..., v_{j-1}, v_j, v_{j+1}, ..., v_d)\big)_i$$

is nonincreasing, where $j \neq i$ and $v_1, ..., v_{j-1}, v_{j+1}, ..., v_M$ are fixed.
(ii) F is inverse-monotone if

$$Fv \leq Fw \quad implies \quad v \leq w.$$

(iii) F is an M-function if F is an inverse-monotone Z-function.

Clearly M-functions are a nonlinear generalization of M-matrices.

If F is differentiable, then F is a Z-function if and only if the Jacobian DF of F satisfies $(DFv)_{ij} \leq 0$ for $i \neq j$ and all v. From Taylor's formula, we have

$$F(w) - F(v) = \left(\int_0^1 DF\big(v + s(w-v)\big)ds \right)(w-v).$$

Write this as

$$F(w) - F(v) = \triangle F(w, v)(w-v),$$

with

$$\triangle F(w, v) := \int_0^1 DF(v + s(w-v))ds.$$

Thus *if all possible matrices $\triangle F(v, w)$ are inverse-monotone, then F is inverse-monotone.* Also, stability estimates for the nonlinear operator can often be derived by estimating $(\triangle F(v, w))^{-1}$, because

$$\|w - v\| \leq \|(\triangle F(v, w))^{-1}\| \, \|F(w) - F(v)\|. \tag{2.46}$$

Now consider again the three-point scheme (2.43). Define the nonlinear mapping F by the left-hand side of (2.43). To ensure the Z-function property, we require that the flux satisfy the *monotonicity condition*

$$\partial_1 g(v, w) \leq 0 \leq \partial_2 g(v, w). \tag{2.46}$$

For the Enquist-Osher scheme,

$$\partial_1 g(v, w) = b_-(v) \leq 0 \quad \text{and} \quad \partial_2 g(v, w) = b_+(w) \geq 0,$$

so F is a Z-function. Next, the Jacobian of F for the general case (2.43) is given by

$$(DF)_{ij} = \begin{cases} \frac{2\varepsilon}{h^2} + \partial_2 c + \frac{1}{h}[\partial_2 g(v_{i+1}, v_i) - \partial_1 g(v_i, v_{i-1})] & \text{for } j = i, \\ -\frac{\varepsilon}{h^2} - \frac{1}{h}\partial_2 g(v_i, v_{i-1}) & \text{for } j = i - 1, \\ -\frac{\varepsilon}{h^2} + \frac{1}{h}\partial_1 g(v_{i+1}, v_i) & \text{for } j = i + 1, \\ 0 & \text{in other cases.} \end{cases}$$

Thus DF satisfies the conditions

$$(DF)_{ij} \leq 0 \quad \text{for } i \neq 0, \quad (DF)_{ii} \geq \mu > 0.$$

This immediately implies that F is a Z-function. If one could prove that the row sums of DF were bounded below by a positive constant, this would imply stability in the discrete maximum norm based on the M-criterion and the majorizing element $(1, 1, ..., 1)^T$. But a close inspection reveals only that the *column sums* of DF are greater or equal to μ. This leads to stability in the discrete L_1 norm, which is defined by

$$\|v_h\|_{L_1,d} := h \sum_i |v_i|.$$

Theorem 2.39. *If the numerical flux of the scheme (2.43) is monotone in the sense of (2.46), then the scheme is stable, uniformly with respect to ε, in the discrete L_1 norm. In particular u_h satisfies the estimate*

$$\|u_h\|_{L_1,d} \leq C.$$

Remark 2.40. In [AO82] the authors use other means to prove that the variation of the discrete solution is uniformly bounded:

$$\sum_i |u_i - u_{i-1}| \leq C.$$

One can conclude from this inequality that as $h \to 0$, a subsequence of the linear interpolant to the discrete solution tends in the $L_1(0, 1)$ sense to the solution of the boundary value problem (2.41).   ♣

The convergence result of Remark 2.40 is not strong. Lorenz [Lor81, Lor84] proves more detailed convergence results in a shock layer situation where $\varepsilon = 0$ with

$$u_0(x) = \begin{cases} u_L(x) \text{ for } 0 \leq x < x_*, \\ u_R(x) \text{ for } x_* < x \leq 1; \end{cases}$$

cf. Section 1.3. Let u_i^0 be a solution of the discrete problem for $\varepsilon = 0$. Suppose that there exists u^* such that $b(s) > 0$ for $s > u^*$ and $b(s) < 0$ for $s < u^*$. Then there exists a unique index $j = j(h)$, defined by the inequality $u_j^0 \leq u^* < u_{j+1}^0$, that indicates the position of the discrete layer and satisfies

$$|x_j - x_*| \leq Ch. \tag{2.47}$$

Lorenz also proves that

$$|u_i^0 - u_L(x_i)| \leq Ch \quad \text{for} \quad i = 0, 1, ..., j-1, \tag{2.48a}$$
$$|u_i^0 - u_R(x_i)| \leq Ch \quad \text{for} \quad i = j+2, ..., N. \tag{2.48b}$$

Schemes of the form (2.43) that satisfy the monotonicity condition (2.46) have only $O(h)$ accuracy for fixed ε (see [LeV90]). To overcome this failing, we introduce the more general scheme

$$-\frac{\varepsilon}{h^2}D^+D^-u_i + \frac{1}{h}\left[g(u_{i+1}, u_i) - g(u_i, u_{i-1})\right] + \beta_i^- c_{i-1} + \beta_i^0 c_i + \beta_i^+ c_{i+1}$$
$$= 0,$$

with $c_j := c(x_j, u_j)$. Define the β_j here by

$$\beta_i^- = \beta\left(b(u_{i-1})/\sqrt{h}\right), \quad \beta_i^+ = \beta\left(-b(u_{i+1})/\sqrt{h}\right),$$
$$\beta_i^0 = 1 - \beta_{i+1}^- - -\beta_{i-1}^+,$$

where $\beta(\rho) := B(p\rho)$; the parameter p is not yet specified and

$$B(r) = \begin{cases} 0 & \text{if } r < 0, \\ r^2 & \text{if } 0 \le r \le 1/2, \\ 1/2 - (1-r)^2 & \text{if } 1/2 \le r \le 1, \\ 1/2, & \text{if } r > 1. \end{cases}$$

The choice $p = 0$ gives the original scheme (2.43). From the consistency point of view (to obtain a second-order scheme for the reduced problem), one wants to choose p as large as possible, but a computation shows that the nonlinear mapping associated with the discrete problem is no longer an M-function if p is too large. Under some restrictive assumptions, Lorenz [Lor84] gives rules for choosing p and improves the estimate (2.48).

Aiming for uniform convergence, we introduce the fitted scheme

$$-\frac{\varepsilon}{h^2}(\sigma_{i+1}u_{i+1} - 2\sigma_i u_i + \sigma_{i-1}u_{i-1}) + \frac{1}{2h}\int_{u_{i-1}}^{u_{i+1}} b(s)\, ds + c(x_i, u_i) = 0. \quad (2.49)$$

If $\sigma_i = \sigma(u_i)$, then the Jacobian of the corresponding nonlinear mapping is

$$(DF)_{ij} = \begin{cases} \frac{2\varepsilon}{h^2}(\sigma u)_i' + \partial_2 c & \text{for } j = i, \\ -\frac{\varepsilon}{h^2}(\sigma u)_{i-1}' - \frac{b(u_{i-1})}{2h} & \text{for } j = i-1, \\ -\frac{\varepsilon}{h^2}(\sigma u)_{i+1}' + \frac{b(u_{i+1})}{2h} & \text{for } j = i+1. \end{cases}$$

Thus one needs

$$(\sigma u)' \ge \frac{h}{2\varepsilon}|b(u)|.$$

Guided by this condition and by the Il'in-Allen-Southwell scheme, choose

$$(\sigma u)' = \zeta\left(\frac{b(u)h}{2\varepsilon}\right), \quad \text{where } \zeta(z) = z \coth z. \quad (2.50)$$

Theorem 2.41. *With the choice (2.50), the fitted scheme (2.49) is stable, uniformly with respect to ε, in the discrete L_1 norm.*

When written in the form (2.43), an investigation of the numerical flux shows that the scheme is not monotone in the sense of (2.46). Niijima [Nii86] proves that the scheme is first-order convergent in the L_1 norm, uniformly in ε, under the assumption $b(u) \ge b_0 > 0$. For Burgers' equation (i.e., $b(u) \equiv u$), O'Reilly [O'R86] demonstrates that the scheme cannot be uniformly convergent for any positive order in the discrete maximum norm.

2.2 Finite Element Methods on Standard Meshes

2.2.1 Basic Results for Standard Finite Element Methods

This Section presents the fundamental ideas and notation used in finite element methods for classical (non-singularly perturbed) two-point boundary value problems. Our approach is standard; see, e.g., [Cia02, GRS07].

Let V be a Hilbert space with norm $\|\cdot\|_V$ (but we shall often omit the subscript V to simplify the notation) and scalar product $(\cdot, \cdot)$. In the discretization of second-order differential equations with domain Ω, one generally chooses V as a subset of the Sobolev space $H^1(\Omega)$. Consider the following abstract variational problem:

Find $u \in V$ such that

$$a(u, v) = f(v) \quad \forall v \in V, \tag{2.51}$$

where $a(\cdot, \cdot)$ is a given bilinear form on $V \times V$ and $f(\cdot)$ a given continuous linear functional on V. We say that $a(\cdot, \cdot)$ is *continuous* on $V \times V$ if there exists a constant β, which is independent of v and w, such that $|a(v, w)| \leq \beta \|v\| \|w\|$ for all v and w in V. The bilinear form $a(\cdot, \cdot)$ is V-*elliptic* or *coercive* if

$$a(v, v) \geq \alpha \|v\|^2 \quad \forall v \in V, \tag{2.52}$$

where α is a positive constant that is independent of v.

Example 2.42. For the boundary value problem

$$-u'' + b(x)u' + c(x)u = f(x) \text{ on } (0, 1), \quad u(0) = u(1) = 0, \tag{2.53}$$

we choose $V = H_0^1(0, 1)$ so that the homogeneous Dirichlet boundary conditions are automatically satisfied and set

$$a(v, w) := \int_0^1 [v'w' + (bv' + cv)w], \qquad f(v) := \int_0^1 fv.$$

Then (2.51) is the standard variational formulation of (2.53).   ♣

The *Lax-Milgram lemma* furnishes sufficient conditions for the existence and uniqueness of solutions of the variational problem (2.51):

Theorem 2.43. *Assume that the bilinear form $a(\cdot, \cdot)$ is continuous and V-elliptic. Then for each continuous linear functional $f(\cdot)$, the problem (2.51) has a unique solution.*

In general, the space V is infinite-dimensional. We therefore approximate V by means of finite-dimensional spaces V_h and pose the variational problem in V_h. Then solving this problem in V_h is equivalent to solving a finite-dimensional system of linear equations.

Assume that the method is *conforming* – that is, that $V_h \subset V$. Then the discrete problem that corresponds to (2.51) is:

Find $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \forall v_h \in V_h. \tag{2.54}$$

This is the Ritz-Galerkin method. The Lax-Milgram lemma implies that

- the discrete problem has a unique solution
- the discrete problem is stable (viz., $\|u_h\| \le (\beta/\alpha)\|f\|_*$, where $\|\cdot\|_*$ is the norm on the dual space V^*).

Let $\{w_i : i = 1, ..., N\}$ be a basis for V_h, where $N = N(h)$ is the dimension of V_h. Then

$$u_h = \sum_{i=1}^{N} u_i w_i,$$

where the unknowns u_i satisfy the linear system

$$AU = b \tag{2.55}$$

with $A_{ij} := a(w_j, w_i)$, $U_i := u_i$, and $b_i := f(w_i)$. If $a(\cdot, \cdot)$ is symmetric and coercive, then the matrix A is symmetric and positive definite.

The discretization error can now be estimated in terms of the approximation error by means of the *Cea lemma*:

Theorem 2.44. *Assume that the hypotheses of the Lax-Milgram lemma are satisfied. Let u and u_h denote the solutions of the continuous problem (2.51) and the discrete problem (2.54), respectively. Then one has the quasi-optimal error estimate*

$$\|u - u_h\| \le \frac{\beta}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|. \tag{2.56}$$

Proof. This argument is standard in every finite element course. Nevertheless, it is given here because we shall want to modify it later.

Equations (2.51) and (2.54) imply the *Galerkin orthogonality* property

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h.$$

This property and the coercivity and continuity of $a(\cdot, \cdot)$ imply that the error $e := u - u_h$ satisfies the inequality

$$\alpha \|e\|^2 \le a(e, u - u_h) = a(e, u - v_h) \le \beta \|e\| \, \|v - v_h\| \quad \forall v_h \in V_h.$$

Hence

$$\|u - u_h\| \le \frac{\beta}{\alpha} \|u - v_h\| \quad \text{for all } v_h \in V_h. \tag{2.57}$$

The desired result follows. $\qquad\square$

Remark 2.45. If $a(\cdot, \cdot)$ is a coercive symmetric bilinear form, then we introduce the *energy product* and the related *energy norm*

$$(v, w)_E := a(v, w) \quad \text{and} \quad \|v\|_E^2 := (v, v)_E.$$

Using this norm in the proof of Theorem 2.44 gives $\alpha = 1$ and $\beta = 1$ in (2.57) so that

$$\|u - u_h\|_E = \inf_{v_h \in V_h} \|u - v_h\|_E. \tag{2.58}$$

Thus *in the symmetric case, the Ritz-Galerkin technique is optimal in the sense that, measured in the energy norm, the discretization error equals the best approximation error from the underlying discrete space.* In the general asymmetric case, the Ritz-Galerkin technique is *quasi-optimal* in the sense of (2.56).   ♣

A *finite element method* is a Ritz-Galerkin method where V_h is a spline space that is called the *finite element space*. The next step in finding a general error estimate for finite element methods is the replacement of the approximation error by an *interpolation error*. Here we assume that it is possible to define an interpolant u^I from V_h to u. Our description so far does not depend on the dimension of the discrete problem, but for the interpolation theory estimates, the precise choice of V_h and its dimension come into the game.

Consider the one-dimensional case. On a given grid

$$0 = x_0 < x_1 < ... < x_{N-1} < x_N = 1,$$

let S^k be the space of continuous splines that are polynomials of degree $k \geq 1$ on each subinterval. Then the interpolant $u^I \in S^k$ to a continuous function is determined uniquely by the correct number of interpolation conditions at suitably chosen interpolation points. Interpolation theory in Sobolev spaces [Cia02] tells us that, with $h_i = x_i - x_{i-1}$ and $h = \max h_i$,

$$\|u - u^I\|_{W^{l,q}(0,1)} \leq C h^{1/q - 1/p} h^{k+1-l} |u|_{W^{k+1,p}(0,1)}, \tag{2.59}$$

for all $u \in W^{k+1,p}(0, 1)$, where $l \leq k + 1$, $1/p + 1/q = 1$ and $1 \leq p \leq \infty$.

Example 2.46. Let us assume for the boundary value problem (2.53) that $c - b'/2 \geq \omega > 0$. Then the hypotheses of Theorems 2.43 and 2.44 are fulfilled. Thus a Ritz-Galerkin discretization, with continuous splines of degree k, results in

$$\|u - u_h\|_1 \leq C h^k |u|_{k+1} \quad \text{if } u \in H^{k+1}(0, 1).$$

In fact, it is well known that

$$|u - u_h|_0 + h|u - u_h|_1 \leq C h^{k+1} |u|_{k+1} \quad \text{if } u \in H^{k+1}(0, 1)$$

where the higher-order $L_2(0, 1)$ estimate can be derived by a duality argument [Cia02, GRS07]. If one wishes to estimate the L_∞ error of $u - u_h$, this can be

done via a *Green's function.* For fixed $\xi \in (0,1)$, define the Green's function $G(\cdot, \xi) \in H_0^1(0,1)$ by

$$a(w, G) = w(\xi) \quad \forall w \in H_0^1(0,1).$$

In the current one-dimensional case, one has $H^1(0,1)$ continuously embedded in $C[0,1]$, and consequently the linear functional $f(\cdot)$ defined by $f(w) = w(\xi)$ satisfies

$$|f(w)| = |w(\xi)| \le \|w\|_{C[0,1]} \le C \|w\|_{H^1(0,1)},$$

which shows that $f(\cdot)$ is continuous. Hence $G(\cdot, \xi) \in H_0^1(0,1)$ is well defined by the Lax-Milgram lemma. Now

$$(u - u_h)(\xi) = a(u - u_h, G) = a(u - u_h, G - v_h) \quad \forall v \in V_h,$$

so

$$|(u - u_h)(\xi)| \le \beta \|u - u_h\|_1 \inf_{v_h \in V_h} \|G(\cdot, \xi) - v_h)\|_1. \tag{2.60}$$

It is clear that the discontinuity of the derivative of $G(x, \xi)$ at $x = \xi$ may in general cause some difficulty if we try to proceed further. But if ξ is a grid point, this problem disappears and one obtains the *superconvergence* result

$$|(u - u_h)(x_i)| \le K h^{2k},$$

where K is some constant; see [DD74] for details. If ξ is not a grid point, then a direct application of (2.60) and $G \in H^1$ yields only

$$\|u - u_h\|_\infty = o(h^k).$$

A more ingenious approach [Whe73] results in the optimal estimate

$$\|u - u_h\|_\infty = \mathcal{O}(h^{k+1}),$$

provided that u is sufficiently smooth.   ♣

2.2.2 Upwind Finite Elements

We now move on to the singularly perturbed boundary value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \text{ on } (0,1), \quad u(0) = u(1) = 0. \tag{2.61a}$$

As is usual in finite element analyses, assume that

$$c(x) - b'(x)/2 \ge \omega > 0 \quad \text{for all } x \in [0,1]. \tag{2.61b}$$

We discuss discretizations on an equidistant grid with mesh size h. Set

$$a(v, w) := \varepsilon(v', w') + (bv' + cv, w),$$

where $(\cdot, \cdot)$ is the $L_2(0, 1)$ inner product. Then an analogue of (2.54),

$$a(u_h, v_h) = (f, v_h) \quad \text{for all } v_h \in V_h,$$

is the starting point when constructing discretizations.

Choose V_h to be the space of piecewise linear functions and approximate the integrals using the trapezoidal rule. This generates the central difference scheme

$$-\varepsilon D^+ D^- u_i + b_i D^0 u_i + c_i u_i = f_i,$$

so our experience with classical finite difference methods tells us that *for singularly perturbed boundary value problems, standard finite element methods are usually unsatisfactory.*

Some theoretical support for this statement will now be given. The assumption (2.61b) implies the coercivity of $a(\cdot, \cdot)$ on $H_0^1(\Omega)$ with respect to the ε-dependent norm

$$\|v\|_\varepsilon^2 := \varepsilon |v|_1^2 + \|v\|_0^2, \tag{2.62}$$

which is related to the energy norm for the symmetric case considered in Remark 2.45. Indeed, there is a positive constant $\alpha = \min\{\omega, 1\}$, which is independent of ε, such that

$$a(v, v) \geq \alpha \|v\|_\varepsilon^2 \quad \forall v \in V. \tag{2.63}$$

Furthermore, there is a positive constant β independent of ε such that

$$|a(v, w)| \leq \beta \|v\|_\varepsilon \|w\|_1 \quad \forall (v, w) \in V \times W, \tag{2.64}$$

but one does not have

$$|a(v, w)| \leq \gamma \|v\|_\varepsilon \|w\|_\varepsilon \quad \forall (v, w) \in V \times W \tag{2.65}$$

with a constant γ that is independent of ε. Using (2.63) and (2.64), the standard analysis yields

$$\|u - u_h\|_\varepsilon \leq C \inf_{v_h \in V_h} \|u - v_h\|_1. \tag{2.66}$$

But in the presence of a boundary layer, if V_h is a polynomial finite element space, then for fixed h one can show that

$$\inf_{v_h \in V_h} \|u - v_h\|_1 \to \infty \quad \text{as } \varepsilon \to 0,$$

and in fact, one does *not* have

$$\|u - u_h\|_\varepsilon \to 0 \quad \text{uniformly in } \varepsilon, \text{ as } h \to 0.$$

See also [KS97], which uses the theory of n-widths to prove that when the smoothness of the right-hand side f is specified and *any* numerical method is applied to the problem (2.61), then in practice the optimal convergence rate attainable in L_2 is inferior to that achieved in classical problems.

Remark 2.47. (Higher-degree polynomial finite element spaces) In numerical experiments it has been observed that finite element methods with continuous, piecewise polynomials of degree $k \geq 2$ behave much better than their piecewise linear counterparts; see, e.g., [BR94]. The theoretical analysis above applies to all finite element spaces and does not explain this particular phenomenon. But in the one-dimensional case one can prove for all $k \geq 2$ that there is a positive constant $\alpha > 0$, independent of the mesh size h, such that

$$\alpha |||v_h|||_\varepsilon \leq \sup_{w_h \in V_h} \frac{a(v_h, w_h)}{|||w_h|||_\varepsilon} \qquad \forall v_h \in V_h,$$

with a norm $||| \cdot |||_\varepsilon$ that is stronger than the energy norm $\| \cdot \|_\varepsilon$; one has $\|w_h\|_\varepsilon \leq |||w_h|||_\varepsilon$ for all $w_h \in V_h$. In addition, for the subspace S^{k-1} of continuous piecewise polynomials of degree $k-1$ there is a positive constant C such that

$$\|w_h\|_\varepsilon^2 + C \sum_{i=1}^{N} h_i \int_{x_{i-1}}^{x_i} (bw_h')^2 \, dx \leq |||w_h|||_\varepsilon^2 \qquad \forall w_h \in S^{k-1}.$$

This means that on the subspace S^{k-1}, as $\varepsilon \to 0$ one has control not only over the L^2 norm but also over a mesh-dependent H^1 seminorm. Hence, for the solution $u_h \in V_h$ of the discrete problem, one gets the improved stability estimate

$$|||u_h|||_\varepsilon \leq \frac{1}{\alpha} \sup_{w_h \in V_h} \frac{(f, w_h)}{|||w_h|||_\varepsilon} \leq \frac{1}{\alpha} \sup_{w_h \in V_h} \frac{(f, w_h)}{\|w_h\|_\varepsilon}$$

whereas the standard approach based on the usual coercivity of the bilinear form gives only

$$\|u_h\|_\varepsilon \leq \frac{1}{\alpha} \sup_{w_h \in V_h} \frac{(f, w_h)}{\|w_h\|_\varepsilon}.$$

For details we refer to [KT08] where the multi-dimensional case has also been investigated. ♣

Remark 2.48. (The influence of different boundary conditions) Let us consider (2.61) with $b > 0$ but with the boundary conditions

$$u(0) = 0 \quad \text{and} \quad u'(1) = 0.$$

Suppose that V_h consists of piecewise polynomials of degree k. Let u_0 be the solution of the reduced problem. Adjoining a triangle inequality to the analysis above gives

$$\|u - u_h\|_\varepsilon \leq C\big(\|u - u_0\|_1 + \inf_{v_h \in V_h} \|u_0 - v_h\|_1\big) \leq C(\varepsilon^{1/2} + h^k),$$

as u_0 is smooth (see Remark 1.5). Thus for the boundary condition $u'(1) = 0$ – which we recall induces a weaker layer – when ε is close to zero, the error behaves better than in the Dirichlet case $u(1) = 0$. ♣

In the late 1970s, researchers began to apply *Petrov-Galerkin methods* to convection-diffusion problems; see [CGMZ76, Hem77, HHZM77, HZ77]. A Petrov-Galerkin method is characterized by the use of distinct trial and test spaces, S_h and T_h respectively (with dim S_h = dim T_h), and the discretization:
Find $u_h \in S_h$ such that

$$a(u_h, v_h) = f(v_h) \quad \forall v_h \in T_h. \tag{2.67}$$

Consider first the simple differential equation

$$-\varepsilon u'' + bu' = 0$$

with constant non-zero b. The finite element spaces are piecewise linear trial and piecewise quadratic test functions. Define the splines

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h & \text{if } x \in [x_{i-1}, x_i], \\ (x_{i+1} - x)/h & \text{if } x \in [x_i, x_{i+1}], \\ 0 & \text{otherwise,} \end{cases}$$

$$\sigma_{i-\frac{1}{2}} = \begin{cases} 4(x - x_{i-1})(x_i - x)/h^2 & \text{if } x \in [x_{i-1}, x_i], \\ 0 & \text{otherwise.} \end{cases}$$

The test functions are given by

$$\psi_i(x) = \varphi_i(x) + \frac{3}{2}\kappa \left[\sigma_{i-\frac{1}{2}}(x) - \sigma_{i+\frac{1}{2}}(x) \right],$$

where κ is a user-chosen *upwind parameter*. This generates the scheme

$$-\varepsilon D^+ D^- u_i + b \left[\left(\frac{1}{2} - \kappa \right) D^+ u_i + \left(\frac{1}{2} + \kappa \right) D^- u_i \right] = 0.$$

The choice $\kappa = (\operatorname{sgn} b)/2$ produces the simple upwind finite difference scheme. In early papers on upwind schemes, the parameter κ was chosen to exclude all oscillations from the solution of the difference equation that was generated. Now for $b > 0$, one obtains an M-matrix if

$$\frac{bh}{\varepsilon} \left(\frac{1}{2} - \kappa \right) < 1$$

– a condition which is fulfilled for the simple upwind scheme. An "optimal" choice of the upwind parameter κ is got by requiring the difference scheme to yield the exact solution of the differential equation; it then generates the Il'in-Allen-Southwell scheme of Section 2.1.3.

For the more general problem (2.61), each test function ψ_i is formed by combining a linear trial function with a quadratic function using an upwind parameter $\kappa_{i-\frac{1}{2}}$, as follows:

$$\psi_i(x) = \varphi_i(x) + \kappa_{i-\frac{1}{2}}[\sigma_{i-\frac{1}{2}}(x) - \sigma_{i+\frac{1}{2}}(x)].$$

Now the midpoint quadrature rule generates the scheme

$$-\varepsilon D^+ D^- u_i + (\bar{b}_{i+\frac{1}{2}} D^+ u_i + \bar{b}_{i-\frac{1}{2}} D^- u_i) + \overline{(cu)}_{i+\frac{1}{2}} + \overline{(cu)}_{i-\frac{1}{2}}$$
$$= (\bar{f}_{i+\frac{1}{2}} + \bar{f}_{i-\frac{1}{2}}).$$

Here $\bar{q}_{i\pm1/2} := (1/2 \mp \kappa_{i\pm1/2}) q_{i\pm1/2}$, where q may be b, cu or f. This scheme can be rewritten (modulo higher-order terms) in the fitted form (2.15). Thus *some fitted upwind schemes can be generated in a Petrov-Galerkin framework, using linear trial functions and quadratic test functions.*

As with finite difference schemes, one could introduce *artificial diffusion* into the original differential equation then apply a standard approach based, e.g., on linear elements. Other combinations of polynomial test and trial functions have been proposed in the literature; for example, see the combination of quadratic trial and cubic test functions in [Hei80].

An abstract mathematical theory of Petrov-Galerkin finite element methods was developed by the early 1970s, but is nevertheless not as well known as the results mentioned in Section 2.2.1. The following generalization of Theorem 2.44 comes from [BA72, GRS07]:

Theorem 2.49. *Assume that the trial space $S_h \subset H_0^1$ and the test space $T_h \subset H_0^1$ satisfy the two conditions*

$$\inf_{\|v_h\|_\varepsilon=1} \sup_{\|w_h\|_\varepsilon=1} |a(v_h, w_h)| \geq \alpha_h > 0 \tag{2.68}$$

and

$$\sup_{v_h \in S_h} |a(v_h, w_h)| > 0 \quad \text{for each } w_h \in T_h \text{ with } w_h \neq 0. \tag{2.69}$$

Then the discrete problem (2.67) has a unique solution u_h which satisfies

$$\|u - u_h\|_\varepsilon \leq \left(1 + \frac{\beta}{\alpha_h}\right) \inf_{v_h \in S_h} \|u - v_h\|_1. \tag{2.70}$$

Remark 2.50. The quasi-optimal bound (2.70) can be found in many textbooks. A careful investigation [XZ03] shows that in fact the multiplicative factor $1 + \beta/\alpha_h$ can be replaced by β/α_h.   ♣

If one applies Theorem 2.49 to Petrov-Galerkin methods based on polynomial finite element spaces, the results are unconvincing. For instance, for linear trial and quadratic test functions one finds that

$$\alpha_h = \frac{1}{\sqrt{1 + 6\max_i \kappa_i^2}} \quad \text{and} \quad \inf_{v_h \in S_h} \|u - v_h\|_1 \leq Ch|u|_2,$$

where the κ_i are upwind parameters and $|u|_2 = \mathcal{O}(\varepsilon^{-3/2})$; see [GL78, Kun86]. Since (2.70) corresponds to the bound (2.66) for the standard Galerkin approach, the best we can hope for when applying a polynomial-based Petrov-Galerkin finite element method is that this will improve the stability properties of the discrete problem.

Morton and his coworkers [BM80, MMS92, MS85] developed a theory of "optimal" Petrov-Galerkin methods. The basic idea is quite elegant: as seen in Section 2.2.1, for symmetric problems the Ritz-Galerkin technique is optimal with respect to the energy norm, so one tries to find test functions that yield a symmetric (or nearly symmetric) discrete problem. That is, one looks for a surjective mapping $\Phi: S_h \to T_h$ such that

$$B_s(v, w) := a(v, \Phi(w))$$

is a symmetric bilinear form. For one-dimensional problems this method works well, but it is difficult to generalize it to higher-dimensional problems, so it will not be discussed further.

Instead of trial and test functions that are linear within each mesh subinterval, O'Riordan [O'R84] proposes the use of *hinged elements*; these are only piecewise linear in each mesh subinterval, thus enabling better approximation of layers. One constructs them by introducing in each subinterval an additional mesh point whose position depends on a local Reynolds number. Recently, in the context of enriching the finite element space by bubble functions, a method using two additional mesh points in each subinterval is proposed in [BHMS03]. This can be considered as an extension of [O'R84] to handle the whole range of convection-diffusion to reaction-diffusion equations.

In recent years many other finite element methods of upwind type such as

- *streamline diffusion method (SDFEM)*
- *variational multiscale method (VMS)*
- *differentiated residual method (DRM)*
- *continuous interior penalty approach (CIP)*
- *Galerkin least squares techniques (GLS)*
- *local projection stabilization (LPS)*
- *discontinuous Galerkin methods (dGFEM)*
- *combined finite volume – finite element approaches (CFVFE)*

have been developed. To give the reader some impression of how higher-order finite element methods can be designed and analysed, the first three methods of this list will be considered in the next subsections; the others are deferred to Parts II and III.

2.2.3 Stabilized Higher-Order Methods

Consider as in Section 2.2.2 the singularly perturbed boundary value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \text{ on } (0,1), \quad u(0) = u(1) = 0, \quad \text{(2.71a)}$$

under the assumption that

$$c(x) - b'(x)/2 \geq \omega > 0 \quad \text{for all } x \in [0,1]. \quad \text{(2.71b)}$$

Our aim is to create a method that is more stable than the Galerkin approach and can be used for finite elements of any order. The improved stability property will be expressed in terms of a norm stronger than the standard energy norm.

The first idea is to add weighted residuals to the usual Galerkin finite element method. The method is called the streamline-diffusion finite element method (SDFEM); the reason for its name will become clear in the multidimensional case – see the interpretation following Remark III.3.28. Multiply the differential equation (2.71) by bv', integrate over each subinterval (x_{i-1}, x_i) for $i = 1, \ldots, N$, and add this weighted sum to the standard Galerkin method; one gets the following discrete problem:

Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = f_h(v_h) \quad \text{for all } v_h \in V_h, \tag{2.72}$$

where

$$a_h(v, w) := \varepsilon(v', w') + (bv' + cv, w) + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i(-\varepsilon v'' + bv' + cv)bw' \, dx,$$

$$f_h(w) := (f, w) + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i f \, bw' \, dx.$$

Here, $(\cdot, \cdot)$ denotes the inner product in $L_2(0, 1)$, δ_i is a user chosen parameter, called the SD parameter, which is usually constant on I_i. Note that since $v \in V_h$, in general v'' in $a_h(v, w)$ is defined only piecewise. Nevertheless, for a smooth solution $u \in H^2(0, 1)$ of (2.71) we have

$$a_h(u, v_h) = f_h(v_h) \quad \text{for all } v_h \in V_h. \tag{2.73}$$

A finite element method (2.72) that satisfies (2.73) for a sufficiently smooth solution of (2.71) is said to be *consistent*. This is *not* the same as consistency of a finite difference scheme, which was discussed in Section 2.1.1. Furthermore, finite element consistency implies *Galerkin orthogonality*, viz.,

$$a_h(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h.$$

As regards coercivity of the discrete bilinear form $a_h(\cdot, \cdot)$, one has

$$a_h(v_h, v_h) = \varepsilon|v_h|_1^2 + \int_0^1 (c - b'/2) v_h^2 \, dx$$

$$+ \sum_{i=1}^{N} \|\sqrt{\delta_i} bv'\|_{0, I_i}^2 + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i \left(-\varepsilon v_h'' + cv_h\right) bv_h' \, dx$$

$$\geq \||v_h\||_{SD}^2 + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i \left(-\varepsilon v_h'' + cv_h\right) bv_h' \, dx,$$

where $I_i = (x_{i-1}, x_i)$ and $\|.\|_{0,I_i}$ denote the ith subinterval and the $L_2(I_i)$ norm. Furthermore, the streamline diffusion norm $||| \cdot |||_{SD}$ has been introduced:

$$|||v_h|||_{SD} := \left(\varepsilon|v_h|_1^2 + \omega\|v_h\|_0^2 + \sum_{i=1}^N \|\sqrt{\delta_i}bv'\|_{0,I_i}^2 \right)^{1/2}.$$

Let $h_i = x_i - x_{i-1}$ be the length of I_i. Using the inverse inequality

$$\|v_h''\|_{0,I_i} \le c_{inv}\, h_i^{-1}\|v_h'\|_{0,I_i}$$

and imposing the requirement on the SD parameter that

$$0 < \delta_i \le \frac{1}{2}\min\left\{ \frac{h_i^2}{\varepsilon c_{inv}^2}, \frac{\omega}{\|c\|_\infty^2} \right\}, \tag{2.74}$$

we estimate

$$\left| \int_{x_{i-1}}^{x_i} \delta_i \left(-\varepsilon v_h'' + cv_h \right) bv_h'\, dx \right|$$

$$\le \left(\sqrt{\frac{\varepsilon}{2}\frac{h_i}{c_{inv}}}\|v_h''\|_{0,I_i} + \sqrt{\frac{\omega}{2}}\|v_h\|_{0,I_i} \right)\|\sqrt{\delta_i}bv_h'\|_{0,I_i}$$

$$\le \frac{\varepsilon}{2}\|v_h'\|_{0,I_i} + \frac{\omega}{2}\|v_h\|_{0,I_i}^2 + \frac{1}{2}\|\sqrt{\delta_i}bv_h'\|_{0,I_i}^2.$$

In the case of piecewise linear elements one has $v_h''|_{I_i} = 0$ for $i = 1, \ldots, N$ and this inequality is still valid when (2.74) is replaced by the weaker assumption

$$0 < \delta_i \le \frac{\omega}{\|c\|_\infty^2}. \tag{2.75}$$

The above computation proves the following lemma:

Lemma 2.51. *Assume that (2.74) is satisfied. Then the SDFEM discrete bilinear form is coercive, viz.,*

$$a_h(v_h, v_h) \ge \frac{1}{2}|||v_h|||_{SD}^2 \qquad \textit{for all } v_h \in V_h.$$

For piecewise linear elements the assumption (2.74) can be replaced by (2.75).

Remark 2.52. Lemma 2.51 implies stability of the SDFEM with respect to the norm $||| \cdot |||_{SD}$. Now all $v_h \in V_h$ satisfy

$$|||v_h|||_{SD} \ge \min\{1, \omega\}\|v_h\|_\varepsilon.$$

Thus the stability of the SDFEM in the norm $||| \cdot |||_{SD}$ is stronger than the stability of the standard Galerkin method in the norm $\|\cdot\|_\varepsilon$. Furthermore, the quantity

$$\sum_{i=1}^N \|\sqrt{\delta_i}bu_h'\|_{0,I_i}^2$$

is bounded for the solution u_h of the SDFEM but in general this is not the case for the solution of the Galerkin method.   ♣

Take V_h to be the space of piecewise linear functions on an equidistant mesh ($h_i = h$ for $i = 1, \ldots, N$). Assume that b, c, f, and $\delta_i = \delta$ for $i = 1, \ldots, N$ are all constant. Then the SDFEM reduces to the scheme

$$-(\varepsilon + b^2\delta)D^+D^-u_i + bD^0u_i + cu_i = f,$$

i.e., the fitted scheme (2.15) with $\sigma(q) = 1 + b^2\delta/\varepsilon$, $q = bh/(2\varepsilon)$. Recall that for $\sigma(q) = q\coth q$ one gets the Il'in-Allen-Southwell scheme, which corresponds to choosing the SD parameter to be

$$\delta(q) = \frac{h}{2b}\left(\coth q - \frac{1}{q}\right).$$

Since

$$\coth q - \frac{1}{q} = \frac{q}{3} + \mathcal{O}(q^3) \quad \text{as } q \to 0 \quad \text{and} \quad \coth q - \frac{1}{q} = 1 + \mathcal{O}\left(\frac{1}{q}\right) \quad \text{as } q \to \infty,$$

the asymptotic limits $h \to 0$ for fixed ε, and $\varepsilon \to 0$ for fixed h, motivate the following choices of δ:

$$\delta(q) = \begin{cases} h^2/(12\varepsilon) & \text{if } 0 < q \ll 1, \\ h/(2b) & \text{if } q \gg 1. \end{cases} \tag{2.76}$$

The choice $\delta(q) = h/(2b)$ for $q \in (0, \infty)$ generates the simple upwind scheme.

We now study the convergence properties of the SDFEM in the case where $V_h \subset H_0^1(0, 1)$ is the finite element space of piecewise polynomials of degree $k \geq 1$. For the nodal interpolant $u^I \in V_h$, one has the estimates

$$|u^I - u|_l \leq Ch^{k+1-l}|u|_{k+1} \quad \text{for } l = 0, \ldots, k+1.$$

Theorem 2.53. *Let the SD parameter be specified by*

$$\delta_i = \begin{cases} C_0 h_i^2/\varepsilon & \text{if } h_i < \varepsilon, \\ C_0 h_i & \text{if } \varepsilon < h_i, \end{cases} \tag{2.77}$$

where the constant C_0 is small enough to satisfy (2.74) if $k \geq 2$ and (2.75) if $k = 1$. Then using piecewise polynomials of degree k, the solution u_h of the SDFEM satisfies the error estimate

$$|||u - u_h|||_{SD} \leq C(\varepsilon^{1/2}h^k + h^{k+1/2})|u|_{k+1}.$$

Proof. The coercivity of a_h (Lemma 2.51) and Galerkin orthogonality yield

$$\frac{1}{2}|||u^I - u_h|||_{SD}^2 \leq a_h(u^I - u_h, u^I - u_h) = a_h(u^I - u, u^I - u_h).$$

Each term in $a_h(u^I - u, u^I - u_h)$ will be estimated separately. Set $w_h = u^I - u_h$. First,

$$\left|\varepsilon((u^I - u)', w_h')\right| \le C\varepsilon^{1/2}h^k|u|_{k+1} |||w_h|||_{SD},$$
$$\left|(c(u^I - u), w_h)\right| \le Ch^{k+1}|u|_{k+1} |||w_h|||_{SD}.$$

Then, using $\varepsilon\delta_i \le C_0\, h_i^2$ and $\delta_i \le C_0\, h_i$ we obtain

$$\left|\sum_{i=1}^{N}(-\varepsilon(u^I - u)'', \delta_i bw_h')_{I_i}\right| \le C\sum_{i=1}^{N}\varepsilon^{1/2}h_i\,\|(u^I - u)''\|_{0,I_i}\|\sqrt{\delta_i}bw_h'\|_{0,I_i}$$
$$\le C\varepsilon^{1/2}h^k|u|_{k+1} |||w_h|||_{SD},$$

$$\left|\sum_{i=1}^{N}(b(u^I - u)' + c(u^I - u), \delta_i bw_h')_{I_i}\right| \le C(h^{k+1/2} + h^{k+3/2})\,|u|_{k+1} |||w_h|||_{SD}$$

It remains to estimate the convection term. The standard estimate would be

$$|(b(u^I - u)', w_h)| \le Ch^k\|w_h\|_0 \le Ch^k|u|_{k+1}|||w_h|||_{SD}$$

but thanks to the additional term $\sum_{i=1}^{N}\|\sqrt{\delta_i}bv'\|_{0,I_i}^2$ in the norm $||| \cdot |||_{SD}$, this estimate can be improved. To this end, one integrates by parts to get

$$|(b(u^I - u)', w_h)| \le |((u^I - u), bw_h')| + |((u^I - u), b'w_h)|$$

Here the second term is estimated in a standard way:

$$|((u^I - u), b'w_h)| \le Ch^{k+1}|u|_{k+1}\|w_h\|_0 \le Ch^{k+1}|u|_{k+1}|||w_h|||_{SD}.$$

The bound on the first term depends on $\varepsilon \le h_i$ or $\varepsilon > h_i$:

$$\left|\sum_{i=1}^{N}((u^I - u), bw_h')_{I_i}\right| \le C\sum_{\varepsilon \le h_i}\delta_i^{-1/2}h_i^{k+1}|u|_{k+1,I_i}\|\sqrt{\delta_i}bw_h'\|_{0,I_i}$$
$$+ C\sum_{\varepsilon > h_i}h_i^{k+1/2}|u|_{k+1,I_i}\varepsilon^{1/2}|w_h|_1$$
$$\le Ch^{k+1/2}|u|_{k+1}|||w_h|||_{SD}.$$

Collecting all these estimates completes the proof of the theorem.   □

The Cea lemma, Theorem 2.44, gives a quasi-optimal error estimate whose constant multiplier depends on the data of the problem. It says that the error is, up to a constant factor, less than or equal to the approximation error. Such an error estimate is highly desirable since it reduces the question of constructing a good solution to the corresponding task in approximation theory. In Section 2.2.2 the error $u - u_h$ has been measured in the energy norm $\|\cdot\|_\varepsilon = (\varepsilon|\cdot|_1^2 + |\cdot|_0^2)^{1/2}$, which forms part of the SD norm. But recalling (2.66), we have no uniform quasi-optimal error estimate in the norm $\|\cdot\|_\varepsilon$. Before considering the question of finding an appropriate norm in which a uniform quasi-optimal error estimate can be given, we demonstrate why the standard

H^1 norm $\| \cdot \|_1$ and the energy norm $\| \cdot \|_\varepsilon$ seem unsuited to our singularly perturbed problem.

Under the hypothesis (2.71b), the operator $\mathcal{L}_\varepsilon : H_0^1(0,1) \to H^{-1}(0,1)$ defined by

$$\langle \mathcal{L}_\varepsilon v, w \rangle = a(v, w) \quad \text{for all } v, w \in H_0^1(0,1)$$

is for each $\varepsilon > 0$ an isomorphism from $H_0^1(0,1)$ onto $H^{-1}(0,1)$. Let us consider two norms $\| \cdot \|_S$ and $\| \cdot \|_T$ on $H_0^1(0,1)$ that are equivalent for fixed ε and are such that the continuity and inf-sup conditions

$$|a(v,w)| \leq \beta \|v\|_S \|w\|_T \qquad \text{for all } v, w \in H_0^1(0,1), \qquad (2.78)$$

$$\inf_{v \in H_0^1(0,1)} \sup_{w \in H_0^1(0,1)} \frac{a(v,w)}{\|v\|_S \|w\|_T} \geq \alpha > 0, \qquad (2.79)$$

hold true. From these inequalities one can deduce immediately that

$$\|\mathcal{L}_\varepsilon^{-1}\| := \sup_{f \in H^{-1}(0,1)} \frac{\|\mathcal{L}_\varepsilon^{-1} f\|_S}{\|f\|_{*,T}} = \sup_{v \in H_0^1(0,1)} \frac{\|v\|_S}{\|\mathcal{L}_\varepsilon v\|_{*,T}} \leq \frac{1}{\alpha},$$

$$\|\mathcal{L}_\varepsilon\| := \sup_{v \in H_0^1(0,1)} \frac{\|\mathcal{L}_\varepsilon v\|_{*,T}}{\|v\|_S} = \sup_{v \in H_0^1(0,1)} \sup_{w \in H_0^1(0,1)} \frac{\langle \mathcal{L}_\varepsilon v, w \rangle}{\|v\|_S \|w\|_T} \leq \beta,$$

where $\| \cdot \|_{*,T}$ denotes the dual norm in $H^{-1}(0,1)$ defined by

$$\|f\|_{*,T} := \sup_{w \in H_0^1(0,1)} \frac{\langle f, w \rangle}{\|w\|_T}.$$

If α and β are independent of ε, then one can consider the norms $\|v\|_S$ and $\|w\|_T$ as natural for $\mathcal{L}_\varepsilon$ because for a given source term f and a perturbed source term $f + \delta f$ the relative perturbation in the solution is uniformly bounded by the relative perturbation of the source term. Indeed, if u and $u + \delta u$ denote the corresponding solutions, then

$$\frac{\|\delta u\|_S}{\|u\|_S} = \frac{\|\mathcal{L}_\varepsilon^{-1} \delta f\|_S}{\|u\|_S} \leq \frac{\beta}{\alpha} \frac{\|\delta f\|_{*,T}}{\|\mathcal{L}_\varepsilon u\|_{*,T}} = \frac{\beta}{\alpha} \frac{\|\delta f\|_{*,T}}{\|f\|_{*,T}}.$$

If however $\| \cdot \|_S = \| \cdot \|_T = \| \cdot \|_1$, then (2.78) and (2.79) hold true only with constants α and β that depend on ε; this implies that

$$\|\mathcal{L}_\varepsilon^{-1}\| \leq \frac{1}{\alpha} = \mathcal{O}\left(\frac{1}{\varepsilon}\right), \qquad \|\mathcal{L}_\varepsilon\| \leq \beta = \mathcal{O}(1).$$

On the other hand, for $\| \cdot \|_S = \| \cdot \|_T = \| \cdot \|_\varepsilon$ one obtains

$$\|\mathcal{L}_\varepsilon^{-1}\| \leq \frac{1}{\alpha} = \mathcal{O}(1), \qquad \|\mathcal{L}_\varepsilon\| \leq \beta = \mathcal{O}\left(\frac{1}{\varepsilon}\right).$$

Suppose that we have appropriate norms $\|\cdot\|_S$ and $\|\cdot\|_T$ such that (2.78) and (2.79) hold with constants α and β that are independent of ε. Then one might hope that these inequalities yield a uniform quasi-optimal convergence result with respect to $\|v\|_S$, similarly to the Cea lemma, Theorem 2.44 – but this is not true. The reason is that the inf-sup condition (2.79) is weaker than the coercivity condition (2.52): imitating the proof of Theorem 2.44 one gets

$$\|u - u_h\|_S \leq \|u - v_h\|_S + \|v_h - u_h\|_S$$

$$\leq \|u - v_h\|_S + \frac{1}{\alpha} \sup_{w \in H_0^1(0,1)} \frac{a(v_h - u_h, w)}{\|w\|_T}$$

but after using Galerkin orthogonality to replace $a(v_h - u_h, w)$ by $a(v_h - u, w)$, we are unable to take an infimum of the right-hand side over $H_0^1(0,1)$ – we can take the infimum only over the finite element space V_h where v_h lies. To surmount this obstacle, one needs an additional inf-sup condition on the discrete spaces S_h and T_h:

$$\inf_{v_h \in S_h} \sup_{w_h \in T_h} \frac{a(v_h, w_h)}{\|v_h\|_S \|w_h\|_T} \geq \alpha_1 > 0. \tag{2.80}$$

Then one can argue that

$$\|v_h - u_h\|_S \leq \frac{1}{\alpha_1} \sup_{w_h \in T_h} \frac{a(v_h - u_h, w_h)}{\|w_h\|_T} = \frac{1}{\alpha_1} \sup_{w_h \in T_h} \frac{a(v_h - u, w_h)}{\|w_h\|_T}$$

$$\leq \frac{\beta}{\alpha_1} \|v_h - u\|_S$$

and use a triangle inequality to get the uniform quasi-optimal estimate

$$\|u - u_h\|_S \leq \left(1 + \frac{\beta}{\alpha_1}\right) \inf_{v_h \in S_h} \|u - v_h\|_S.$$

An investigation of norms $\|\cdot\|_S$ and $\|\cdot\|_T$ such that (2.78)–(2.80) are satisfied has been carried out by Sangalli [San05, San08].

Following [San05], we consider the simple model problem in which $b = 1$ and $c = 0$. Thus the bilinear form $a(\cdot, \cdot)$ becomes

$$a(v, w) := \varepsilon(v', w') + (w', v) \quad \text{for all } v, w \in H_0^1(0,1).$$

Let $L_0^2(0,1)$ denote the subset of $L^2(0,1)$ comprising functions of zero mean value. Let $\Pi_0 : L^2(0,1) \to L_0^2(0,1)$ be the L^2 projection onto $L_0^2(0,1)$ such that $(\Pi_0 w - w, v) = 0$ for all $v \in L_0^2(0,1)$ and $w = \Pi_0 w + \overline{w}$ where $\overline{w}$ denotes the mean value of w. The convection term (v', w) can be estimated via

$$|(v', w)| = |((\Pi_0 v)', w)| = |-(\Pi_0 v, w')| \leq \|\Pi_0 v\|_0 \, |w|_1$$

or equivalently, integrating by parts,

$$|(v', w)| = |-(v, w')| = |-(v, (\Pi_0 w)')| = |(v', \Pi_0 w)| \le |v|_1 \, \|\Pi_0 w\|_0,$$

which results in two continuity estimates of the form (2.78):

$$|a(v, w)| \le (\varepsilon |v|_1 + \|\Pi_0 v\|_0) \, |w|_1,$$
$$|a(v, w)| \le |v|_1 \, (\varepsilon |w|_1 + \|\Pi_0 v\|_0).$$

Thus we shall consider $\varepsilon |\cdot|_1 + \|\Pi_0(\cdot)\|_0$ and $|\cdot|_1$ – or vice versa – as candidates for $\|\cdot\|_S$ and $\|\cdot\|_T$. The coercivity of $a(\cdot, \cdot)$ gives

$$\varepsilon |v|_1 \le \sup_{w \in H_0^1(0,1)} \frac{a(v, w)}{|w|_1} \qquad \text{for all } v \in H_0^1(0,1).$$

In order to show also that

$$\|\Pi_0 v\|_0 \le C \sup_{w \in H_0^1(0,1)} \frac{a(v, w)}{|w|_1} \qquad \text{for all } v \in H_0^1(0,1),$$

one uses the norm relationships $\|\Pi_0 v\|_0 \le C \, \|v'\|_{-1}$ and

$$\|v'\|_{-1} = \sup_{w \in H_0^1(0,1)} \frac{(v', w)}{|w|_1} = \sup_{w \in H_0^1(0,1)} \frac{a(v, w) - \varepsilon(v', w')}{|w|_1}$$

$$\le \sup_{w \in H_0^1(0,1)} \frac{a(v, w)}{|w|_1} + \varepsilon |v|_1 \le 2 \sup_{w \in H_0^1(0,1)} \frac{a(v, w)}{|w|_1}.$$

Hence

$$\alpha(\varepsilon |v|_1 + \|\Pi_0 v\|_0) \le \sup_{w \in H_0^1(0,1)} \frac{a(v, w)}{|w|_1}$$

where α is independent of ε. A duality argument delivers the other estimate

$$\alpha \, |v|_1 \le \sup_{w \in H_0^1(0,1)} \frac{a(v, w)}{\varepsilon |w|_1 + \|\Pi_0 w\|_0}.$$

Thus, in agreement with our earlier discussion, the norms

$$v \mapsto \varepsilon |v|_1 + \|\Pi_0 v\|_0 \quad \text{and} \quad v \mapsto |v|_1$$

are suitable for this model problem.

In [San05] Sangalli proved a discrete inf-sup condition of type (2.80) from which uniform quasi-optimality with respect to the two norms follows.

Lemma 2.54. *Consider the bilinear form $a_h(\cdot, \cdot)$ of the SDFEM (2.72) with $b = 1$ and $c = 0$. Let V_h be the space of piecewise linear functions on an equidistant mesh. Then there is a constant α_1, which is independent of ε, such that*

$$\alpha_1(\varepsilon |v_h|_1 + \|\Pi_0 v_h\|_0) \le \sup_{w_h \in V_h} \frac{a_h(v_h, w_h)}{|w_h|_1} \qquad \forall v_h \in V_h,$$

$$\alpha \, |v_h|_1 \le \sup_{w_h \in V_h} \frac{a_h(v_h, w_h)}{\varepsilon |w_h|_1 + \|\Pi_0 w_h\|_0} \qquad \forall v_h \in V_h.$$

Proof. See [San05, Lemma 3.1]. □

Lemma 2.54 is the basis for using interpolation theory to construct a family of norms in which the SDFEM yields uniform quasi-optimal estimates; see [San05]. Note that the analysis given in [San05] is restricted to the model problem ($b = 1$ and $c = 0$) in one space dimension.

Next, following [CX08, CX05], we show that a variant of the SDFEM for continuous piecewise linear finite elements on an arbitrary family of meshes yields a solution u_h that is quasi-optimal with respect to the L^∞ norm, viz.,

$$\|u - u_h\|_\infty \le C \inf_{v_h \in V_h} \|u - v_h\|_\infty.$$

To concentrate on the main ideas, consider the simple model problem

$$-\varepsilon u'' + bu' = f \quad \text{on } (0,1), \qquad u(0) = u(1) = 0,$$

where b is a positive constant and f a given function. For a positive integer N, let $\mathcal{T}_N = \{x_i : 0 = x_0 < x_1 < \cdots < x_N = 1\}$ be an arbitrary grid with $h_i = x_i - x_{i-1}$ the local mesh size and $\{\varphi_i\}$ the standard piecewise linear hat functions that satisfy $\varphi_i(x_j) = \delta_{ij}$ for $i, j = 0, 1, \ldots, N$. Let the finite element space be

$$V_h := \text{span}\{\varphi_1, \ldots, \varphi_{N-1}\} \subset H_0^1(0,1).$$

The SDFEM (2.72) can be written in the form

Find $u_h \in V_h$ such that $\quad a_h(u_h, v_h) = f_h(v_h) \quad$ for all $v_h \in V_h$

where

$$a_h(v, w) := \varepsilon(v', w') + (bv', w) + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i(-\varepsilon v'' + bv')bw' \, dx,$$

$$f_h(w) := (f, w) + \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} \delta_i f \, bw' \, dx.$$

Unlike the usual choice of a piecewise-constant SD parameter, here we take

$$\delta_i := \frac{3h_i}{b} \min\{1, q_i\} \varphi_{i-1}(x)\varphi_i(x), \qquad q_i = \frac{bh_i}{2\varepsilon}. \qquad (2.81)$$

Nevertheless the maximum of δ_i has the asymptotic behaviour (2.77) in the diffusion-dominated and convection-dominated regimes.

Let $A = (a_h(\varphi_j, \varphi_i))$, for $i, j = 1, \ldots, N-1$, be the coefficient matrix of the corresponding algebraic system. For $u_i = u_h(x_i)$ a direct calculation gives

$$-\left(\frac{\varepsilon + \overline{\delta}_i b^2}{h_i} + \frac{b}{2}\right) u_{i-1} + \left(\frac{\varepsilon + \overline{\delta}_i b^2}{h_i} + \frac{\varepsilon + \overline{\delta}_{i+1} b^2}{h_{i+1}}\right) u_i \qquad (2.82)$$

$$-\left(\frac{\varepsilon + \overline{\delta}_{i+1} b^2}{h_{i+1}} - \frac{b}{2}\right) u_{i+1} = f_h(\varphi_i)$$

where
$$\bar{\delta}_i := \frac{1}{h_i} \int_{x_{i-1}}^{x_i} \delta_i(x)\,dx = \frac{h_i}{2b}\min\{1, q_i\}.$$

Observe that

$$\frac{\varepsilon + \bar{\delta}_{i+1}b^2}{bh_{i+1}} = \frac{1 + q_{i+1}\min(1, q_{i+1})}{2q_{i+1}} = \begin{cases} \dfrac{1 + q_{i+1}^2}{2q_{i+1}} > 1 & \text{for } 0 < q_{i+1} < 1, \\[3mm] \dfrac{1 + q_{i+1}}{2q_{i+1}} > \dfrac{1}{2} & \text{for } q_{i+1} \geq 1, \end{cases}$$

so the matrix A of (2.82) is an M-matrix. The following uniform stability result is established in [CX05] by studying the properties of the discrete Green's function (compare Section 1.1.2 for the continuous analogue):

Lemma 2.55. *Define δ_i by (2.81). Then the SDFEM is uniformly $(l_\infty, w^{-1,\infty})$ stable, i.e.,*

$$\|v_h\|_{\infty,d} \leq \frac{2}{b} \max_{j=1,\ldots,N-1} \left| \sum_{k=j}^{N-1} (Av_h)_k \right| \quad \forall v_h \in V_h,$$

where the right-hand side defines the discrete analogue of the norm $W^{-1,\infty}$.

Now consider the error $e_h = u^I - u_h \in V_h$ where u^I is the nodal interpolant. The consistency property $a_h(u, v_h) = f_h(v_h)$ for all $v_h \in V_h$ implies that (provided the solution u is sufficiently smooth) the error e_h is the solution of the problem

Find $e_h \in V_h$ such that $\quad a_h(e_h, v_h) = a_h(u^I - u, v_h) \quad$ for all $v_h \in V_h$.

Using $(u^I - u)(x_i) = 0$ for $i = 0, \ldots, N$ and integration by parts, one sees that

$$(Ae_h)_k = a_h(e_h, \varphi_k) = a_h(u^I - u, \varphi_k) = r_k - r_{k+1}$$

where

$$r_k := \frac{b}{h_k}\left[-\int_{x_{k-1}}^{x_k} (u^I - u)(x)\,dx + \int_{x_{k-1}}^{x_k} \delta_k(x)\varepsilon u''(x)\,dx \right.$$
$$\left. + \int_{x_{k-1}}^{x_k} b\delta_k(x)(u^I - u)'(x)\,dx. \right]$$

Since the SD parameter δ_i vanishes at the mesh points, one can show by means of integration by parts that

$$\left| \frac{b}{h_k} \int_{x_{k-1}}^{x_k} (u^I - u)(x)\, dx \right| \le b\|u - u^I\|_\infty,$$

$$\left| \frac{b}{h_k} \int_{x_{k-1}}^{x_k} \delta_k(x)\varepsilon u''(x)\, dx \right| = \frac{b}{h_k} \left| \int_{x_{k-1}}^{x_k} \delta_k''(x)\varepsilon(u - u^I)(x)\, dx \right|$$

$$\le \frac{3b}{q_k} \min(1, q_k)\|u - u^I\|_\infty \le 3b\|u - u^I\|_\infty,$$

$$\left| \frac{b}{h_k} \int_{x_{k-1}}^{x_k} b\delta_k(x)(u^I - u)'(x)\, dx \right| = \frac{b^2}{h_k} \left| \int_{x_{k-1}}^{x_k} \delta_k'(x)(u^I - u)(x)\, dx \right|$$

$$\le 3b\|u - u^I\|_\infty.$$

Gathering all these bounds gives

$$|r_k| \le 7b\|u - u^I\|_\infty. \tag{2.83}$$

The discretization error can now be estimated using the interpolation error.

Lemma 2.56. *Let u_h be the solution of the SDFEM with δ_i given by (2.81). Then there is a positive constant C, independent of ε and the mesh, such that*

$$\|u - u_h\|_\infty \le C\|u - u^I\|_\infty.$$

Proof. By Lemma 2.55 and (2.83),

$$\|u^I - u_h\|_\infty \le \frac{2}{b} \max_{j=1,\ldots,N-1} \left| \sum_{k=j}^{N-1} (Ae_h)_k \right| = \frac{2}{b} \max_{j=1,\ldots,N-1} |r_j - r_N|$$

$$\le 28\|u - u^I\|_\infty$$

and the desired estimate follows from the triangle inequality.   □

Theorem 2.57. *Let u_h be the solution of the SDFEM with δ_i given by (2.81). Then there is a positive constant C, independent of ε and the mesh, such that*

$$\|u - u_h\|_\infty \le C \inf_{v_h \in V_h} \|u - v_h\|_\infty.$$

That is, the SDFEM is quasi-optimal in the L_∞ norm.

Proof. Let $P_h : H_0^1(0,1) \to V_h$ denote the solution operator of the SDFEM, i.e., $P_h u := u_h$. From Lemma 2.56 we infer that

$$\|u - u_h\|_\infty \le C\|u - u^I\|_\infty \le C(\|u\|_\infty + \|u^I\|_\infty) \le C\|u\|_\infty.$$

Thus the operator P_h is L_∞ stable since

$$\|P_h u\|_\infty = \|u_h\|_\infty \le \|u\|_\infty + \|u - u_h\|_\infty \le C\|u\|_\infty.$$

But $P_h^2 = P_h$, so for any $v_h \in V_h$ one has

$$\|u - u_h\|_\infty = \|(I - P_h)(u - v_h)\|_\infty \le C\|u - v_h\|_\infty.$$

The proof is then finished by taking the infimum over all $v_h \in V_h$.   □

Remark 2.58. The quasi-optimality result of Lemma 2.56 reduces the question of L_∞-norm convergence of the SDFEM to a problem in approximation. Thus if layer-adapted meshes are used, convergence can be established in the L_∞ norm uniformly with respect to ε. Of course a detailed knowledge of the analytical structure of the solution u is needed in order to create a layer-adapted mesh. ♣

Remark 2.59. A quasi-optimality result in an L_p-type norm (where $1 \le p \le \infty$ is arbitrary) for a Petrov-Galerkin finite element method is given in [SB84]. This result could also be used to get a uniform convergence result on a suitable layer-adapted mesh. Moreover, [SB84] contains an asymptotically exact error estimator; such estimators will be the main topic of Section III.3.6. ♣

2.2.4 Variational Multiscale and Differentiated Residual Methods

The variational multiscale method (VMS) [HFMQ98, Hug95, HS07] was introduced to provide a framework for a better understanding of fine-to-coarse scale effects and as a platform for the development of new numerical methods.

We derive the method for the two-point boundary value problem

$$-\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{in } (0,1), \quad u(0) = u(1) = 0, \qquad (2.84)$$

with sufficiently smooth functions b, c and f, where the parameter ε satisfies $0 < \varepsilon \ll 1$. Assume that

$$c(x) - \frac{1}{2}b'(x) \ge \omega > 0 \quad \text{for } x \in [0,1], \qquad (2.85)$$

which guarantees the unique solvability of the problem.

The weak formulation of (2.84) is given by:

Find $u \in V := H_0^1(0,1)$ such that for all $v \in V$ one has

$$a(u,v) := \varepsilon(u',v') + (bu' + cu, v) = (f,v). \qquad (2.86)$$

The basic idea of the VMS approach is to split the solution space V into resolvable and unresolvable scales. This is done by choosing a finite element space V_h that represents the resolvable scales and a projection operator $P : V \to V_h$ such that

$$V = V_h \oplus V^\diamond, \qquad \text{so } u = Pu + (I - P)u = u_h + u^\diamond.$$

Now the weak formulation (2.86) can be restated as:

Find $u_h \in V_h$ and $u^\diamond \in V^\diamond$ such that

$$a(u_h + u^\diamond, v_h) = (f, v_h) \quad \forall v_h \in V_h, \qquad (2.87a)$$
$$a(u_h + u^\diamond, v^\diamond) = (f, v^\diamond) \quad \forall v^\diamond \in V^\diamond. \qquad (2.87b)$$

To remove the unresolvable scales, define $M(u_h)$ and $F(f)$ as follows:

Find $M(u_h) \in V^\diamond$ and $F(f) \in V^\diamond$ such that

$$a(M(u_h), v^\diamond) = -a(u_h, v^\diamond) \quad \text{and} \quad a(F(f), v^\diamond) = (f, v^\diamond) \quad \forall v^\diamond \in V^\diamond. \quad (2.88)$$

Then, the solution of (2.87b) is $u^\diamond = M(u_h) + F(f)$ and its elimination from (2.87a) yields the *VMS stabilized method:*

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a(u_h + M(u_h), v_h) = (f, v_h) - a(F(f), v_h). \quad (2.89)$$

Remark 2.60. The variational multiscale approach is not restricted to the one-dimensional case; it can also be used for an arbitrary variational problem defined by a coercive, continuous bilinear form on a Hilbert space V.   ♣

Remark 2.61. (Residual-free bubble method) The VMS method can be considered as a generalization of the residual-free bubble (RFB) method, which will be discussed in Section III.3.2.2. In the RFB method, the finite element space V_h is enriched by a space $V^\diamond$ consisting of (bubble) functions that vanish at the element boundaries. Choosing the projection $P : V \to V_h$ appropriately, the functions from $V^\diamond$ in the variational multiscale method here will also vanish at the element boundaries, but they do not have this attribute in the multi-dimensional case.   ♣

Remark 2.62. The problem (2.89) is finite-dimensional but is based on the solution of the infinite-dimensional problems (2.88). Thus in practice one has to approximate the mappings M and F [ARS04, BMR98, BMR05]. In special situations (e.g., the one-dimensional case with piecewise constant coefficients) one can obtain explicit representations for these mappings.   ♣

Let $0 = x_0 < x_1 < \cdots < x_N = 1$ be a partition $\mathcal{T}_h$ of $[0, 1]$. Denote an arbitrary mesh interval (x_{i-1}, x_i) by K, its length by $h_K = x_i - x_{i-1}$, and set $h = \max_{K \in \mathcal{T}_h} h_K$. We consider two examples of the variational multiscale approach. First, let V_h be the space of piecewise linear finite elements and let $P : V \to V_h$ be the $H_0^1(0, 1)$-projection defined by

$$((Pv)', w_h') = (v', w_h') \qquad \forall w \in V_h.$$

Later it will turn out that $(Pv)(x_i) = v(x_i)$ for $i = 0, \ldots, N$, i.e., P is the piecewise linear nodal interpolant. Consequently $V^\diamond$ becomes the bubble space:

$$V^\diamond = \bigoplus_{K \in \mathcal{T}_h} H_0^1(K)$$

and the problems (2.88) can be solved locally on each mesh interval K. If one also has $c = 0$ and the functions b and f are piecewise constant, then explicit representations of the operators M and F can be found. The VMS stabilized

method (2.89) is then identical to the following SDFEM:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a(u_h, v_h) + \sum_{K \in \mathcal{T}_h} \tau_K (bu'_h, bv'_h)_K = (f, v_h) + \sum_{K \in \mathcal{T}_h} \tau_K (f, bv'_h)_K$$

with the SD parameter

$$\tau_K = \frac{h_K}{2b} \left(\coth q_K - \frac{1}{q_K} \right), \qquad q_K = \frac{bh_K}{2\varepsilon}.$$

In the second example, let V_h be the space of piecewise quadratic finite elements and let $P : V \to V_h$ be the $H_0^1(0,1)$-projection. Later we shall see that

$$(Pv)(x_i) = v(x_i), \ i = 0, \dots, N \quad \text{and} \quad \int_{x_{i-1}}^{x_i} (Pv - v)(x) \, dx = 0, \ i = 1, \dots, N.$$

Because the quadratic bubble function $x \mapsto (x_i - x)(x - x_{i-1})$ belongs to $H_0^1(x_{i-1}, x_i)$, the space of unresolvable scales is no longer the entire bubble space; instead it is the constrained bubble space

$$V^\diamond := \left\{ v^\diamond \in \bigoplus_{K \in \mathcal{T}_h} H_0^1(K) \ : \ Pv^\diamond = 0 \right\}.$$

Using the method of Lagrange multipliers, (2.87b) can be reformulated as the mixed problem

Find $(u^\diamond, \xi) \in \bigoplus_{K \in \mathcal{T}_h} H_0^1(K) \times \mathbb{R}^N$ such that

$$a(u^\diamond, v) - \sum_{K \in \mathcal{T}_h} \xi_K (1, v)_K = (f, v) - a(u_h, v) \qquad \forall v \in \bigoplus_{K \in \mathcal{T}_h} H_0^1(K),$$

$$\sum_{K \in \mathcal{T}_h} \eta_K (1, u^\diamond)_K = 0 \qquad\qquad \forall \eta = (\eta_K) \in \mathbb{R}^N.$$

Assuming that $c = 0$ and b, f are piecewise constant functions, one can find an explicit expression for the solution $(u^\diamond, \xi)$. The VMS stabilized method described by (2.89) becomes

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a(u_h, v_h) + \sum_{K \in \mathcal{T}_h} \tau_K (bu''_h, bv''_h)_K = (f, v_h) \tag{2.90}$$

with the VMS parameter

$$\tau_K = \frac{h_K^3}{72b} \left(\frac{q_K}{q_K \coth q_K - 1} - \frac{3}{q_K} \right), \qquad q_K = \frac{bh_K}{2\varepsilon}.$$

Note that the mapping Φ given by

$$\Phi(q) := \frac{q_K}{q_K \coth q_K - 1} - \frac{3}{q_K}$$

is strictly monotone on $[0, \infty)$ with $\Phi(0) = 0$ and $\lim_{q \to +\infty} \Phi(q) = 1$.

Now let us return to the general case in which b, c, and f are smooth functions that satisfy (2.85). Our aim is to design a method that uses piecewise polynomials of degree $m \in \mathbb{N}$ and is related to the VMS method in the piecewise constant coefficient case. The finite element space is given by

$$V_h := \{v_h \in V : v\big|_K \in P_m(K), \ v_h(0) = v_h(1) = 0\}$$

where $P_m(K)$ denotes the space of polynomials of degree at most m.

Assume that the solution of (2.84) belongs piecewise to $H^{k+1}(K)$. Then $(k-1)$ differentiations of the equation (2.84) gives

$$(-\varepsilon u'' + bu' + cu)^{(k-1)} = f^{(k-1)} \qquad \text{in } L^2(K) \text{ for all } K \in \mathcal{T}_h. \qquad (2.91)$$

Multiplying this equation by a user-chosen non-negative function τ_K, testing against $(bv_h')^{(k-1)}$, summing over K and adding this to the weak formulation (2.86), one sees that the solution of (2.86) satisfies

$$a_h(u, v_h) = l_h(v_h) \qquad \text{for all } v_h \in V_h \qquad (2.92)$$

where

$$a_h(u, v) := a(u, v) + \sum_{K \in \mathcal{T}_h} \left(\tau_K (-\varepsilon u'' + bu' + cu)^{(k-1)}, (bv')^{(k-1)} \right)_K,$$

$$l_h(v) := (f, v) + \sum_{K \in \mathcal{T}_h} \left(\tau_K f^{(k-1)}, (bv')^{(k-1)} \right)_K.$$

Thus the method is consistent. The associated discrete problem is:

$$\text{Find } u_h \in V_h \text{ such that } a_h(u_h, v_h) = l_h(v_h) \text{ for all } v_h \in V_h. \qquad (2.93)$$

The above derivation inspires us to call (2.93) the differentiated residual method (DRM). In the case $k = 1$ it is the same as the SDFEM, which was analysed in Section 2.2.3 both for piecewise linear elements ($m = 1$) and for higher-order elements ($m \geq 2$). In what follows only the case $k = m \in \mathbb{N}$ is considered; this coincides with the SDFEM for $m = 1$ but differs if $m \geq 2$. For constant functions b and f, $c = 0$ and $m = 2$, the method coincides with (2.90) and indeed was first derived [HS07] via a variational multiscale approach. Define the mesh-dependent DRM norm related to the discrete bilinear form $a_h(\cdot, \cdot)$ by

$$|||v|||_{DRM} := \left(\varepsilon |v|_1^2 + \omega \|v\|_0^2 + \sum_{K \in \mathcal{T}_h} \|\tau_K^{1/2} (bv')^{(k-1)}\|_{0,K}^2 \right)^{1/2}. \qquad (2.94)$$

Now we turn to the convergence properties of the method on an arbitrary mesh. The smoothness of c and inverse inequalities guarantee the existence of a general constant c_{max} such that

$$\|(cv_h)^{(k-1)}\|_{0,K} \leq C \sum_{l=0}^{k-1} \|v_h\|_{l,K} \leq c_{max}\, h_K^{-(k-1)} \|v_h\|_{0,K}. \qquad (2.95)$$

The constant c_{max} depends on the polynomial degree k, but to simplify the notation this will not be indicated. In what follows we assume that the user-chosen DRM parameter τ_K satisfies

$$0 \leq \tau_K(x) \leq \frac{\omega}{c_{max}^2}\, h_K^{2k-2} \qquad \forall x \in K,\ K \in \mathcal{T}_h. \qquad (2.96)$$

For $k = 1$ this recovers the choice (2.75) of the SDFEM for piecewise linear elements.

Lemma 2.63. *Let (2.96) be satisfied. Then the bilinear form a_h is coercive on V_h:*

$$a_h(v_h, v_h) \geq \frac{1}{2}|||v_h|||_{DRM}^2 \qquad \forall v_h \in V_h. \qquad (2.97)$$

Proof. Start from the definition of the bilinear form a_h. Integrating by parts, using (2.85) and $v_h|_K \in P_k(K)$, one gets

$$a_h(v_h, v_h) = \varepsilon |v_h|_1^2 + (bv_h' + cv_h, v_h)$$
$$+ \sum_{K \in \mathcal{T}_h} \left(\tau_K (bv_h' + cv_h)^{(k-1)}, (bv_h')^{(k-1)} \right)_K$$
$$\geq |||v_h|||_{DRM}^2 + \sum_{K \in \mathcal{T}_h} \left(\tau_K (cv_h)^{(k-1)}, (bv_h')^{(k-1)} \right)_K.$$

The second term here can be absorbed into $|||v_h|||_{DRM}^2$, as (2.96) and (2.95) give

$$\left| \sum_{K \in \mathcal{T}_h} \left(\tau_K^{1/2}(cv_h)^{(k-1)}, \tau_K^{1/2}(bv_h')^{(k-1)} \right)_K \right|$$
$$\leq \sum_{K \in \mathcal{T}_h} (\omega)^{1/2} \frac{h_K^{k-1}}{c_{max}} \|(cv_h)^{(k-1)}\|_{0,K} \|\tau_K^{1/2}(bv_h')^{(k-1)}\|_{0,K}$$
$$\leq \frac{\omega}{2} \sum_{K \in \mathcal{T}_h} \|v_h\|_{0,K}^2 + \frac{1}{2} \sum_{K \in \mathcal{T}_h} \|\tau_K^{1/2}(bv_h')^{(k-1)}\|_{0,K}^2.$$

Combining these estimates yields the conclusion of the lemma. $\qquad\square$

Next we introduce a special interpolant that will be useful later. On each $\overline{K} \in \mathcal{T}_h$ with $K = (x_{i-1}, x_i)$, define $k + 1$ nodal functionals N_l by

$$N_0(v) = v(x_{i-1}), \qquad N_k(v) = v(x_i),$$

$$N_l(v) = h_K^{-l} \int_{x_{i-1}}^{x_i} (x - x_{i-1})^{l-1} v(x) \, dx \quad \text{for } l = 1, \dots, k-1.$$

Lemma 2.64. *The set of nodal functionals $\{N_l : l = 0, \dots, k\}$ is $P_k(K)$ unisolvent.*

Proof. As $\dim P_k = k + 1$, one need only show that if any polynomial $p \in P_k(K)$ satisfies $N_l(p) = 0$ for $l = 0, \dots, k$, then p is identically zero. Let L_l denote the Legendre polynomial of degree l defined on $(-1, +1)$ and normalized by setting $L_l(1) = 1$. Transforming $K = (x_{i-1}, x_i)$ onto $(-1, +1)$, one can write the polynomial p in the form

$$p(x) = \sum_{l=0}^{k} p_l L_l \left(\frac{2x - x_{i-1} - x_i}{x_i - x_{i-1}} \right).$$

Now $N_l(p) = 0$ for $l = 1, \dots, k-1$ and the orthogonality property of the Legendre polynomials imply that $p_l = 0$ for $l = 0, 1, \dots, k-2$. Then the two remaining conditions $N_0(p) = N_k(p) = 0$ are equivalent to the algebraic system

$$0 = p_{k-1} L_{k-1}(-1) + p_k L_k(-1) = (-1)^{k-1} (p_{k-1} - p_k),$$
$$0 = p_{k-1} L_{k-1}(+1) + p_k L_k(+1) = (p_{k-1} + p_k),$$

with the unique solution $p_{k-1} = p_k = 0$.   □

Thus a local interpolant $\pi v|_K \in P_k(K)$ can be defined by $N_l(\pi v - v) = 0$ for $l = 0, \dots, k$. This can be extended to a continuous global interpolant $\pi v \in V_h$.

Next, we need the following property.

Lemma 2.65. *Let $\psi : [x_{i-1}, x_i] \to \mathbb{R}$ be a continuous function with $N_l(\psi) = 0$ for $l = 1, \dots, k-1$ where $k \geq 2$. Then there exists $F \in C^{k-1}[x_{i-1}, x_i]$ such that*

$$F^{(l)}(x_{i-1}) = F^{(l)}(x_i) = 0 \quad \text{for } l = 0, \dots, k-2,$$

$$F^{(k-1)}(x) = \psi(x), \qquad \|F\|_{0,K} \leq \left(\frac{h_K}{\sqrt{2}} \right)^{k-1} \|\psi\|_{0,K}.$$

Proof. Use induction on k. For $k = 2$ define $F \in C^1[x_{i-1}, x_i]$ by

$$F(x) := \int_{x_{i-1}}^{x} \psi(t) \, dt.$$

Clearly $F(x_{i-1}) = 0$ and $F'(x) = \psi(x)$. Furthermore, $F(x_i) = N_1(\psi) = 0$. The Cauchy-Schwarz inequality gives $|F(x)| \leq \sqrt{(x - x_{i-1})} \|\psi\|_{0,K}$, so

$$\|F\|_{0,K} \le \frac{h_K}{\sqrt{2}} \|\psi\|_{0,K}.$$

Next, assume that the lemma holds true for some $k \ge 2$. We shall deduce that it is valid for $k + 1$. By our inductive hypothesis, there exists a function $F \in C^{k-1}[x_{i-1}, x_i]$ with the properties stated in the lemma. Define $\tilde{F} : [x_{i-1}, x_i] \to \mathbb{R}$ by

$$\tilde{F}(x) := \int_{x_{i-1}}^{x} F(t)\,dt.$$

Then $\tilde{F} \in C^k[x_{i-1}, x_i]$ and $\tilde{F}^{(k)}(x) = F^{(k-1)}(x) = \psi(x)$. Moreover,

$$\tilde{F}^{(l)}(x_{i-1}) = F^{(l-1)}(x_{i-1}) = 0 \quad \text{for } l = 1, \dots, k-1,$$
$$\tilde{F}^{(l)}(x_i) = F^{(l-1)}(x_i) = 0 \quad \text{for } l = 1, \dots, k-1,$$

and

$$\|\tilde{F}\|_{0,K} \le \frac{h_K}{\sqrt{2}} \|F\|_{0,K} \le \left(\frac{h_K}{\sqrt{2}}\right)^k \|\psi\|_{0,K}.$$

It remains to show that $\tilde{F}(x_{i-1}) = 0$ and $\tilde{F}(x_i) = 0$. The first equation is true by definition of $\tilde{F}$. To verify the second equation, one integrates by parts $k-1$ times using $F^{(l)}(x_{i-1}) = F^{(l)}(x_i) = 0$ for $l = 0, \dots, k-2$, then recalls that $F^{(k-1)}(x) = \psi(x)$. That is,

$$\tilde{F}(x_i) = \int_{x_{i-1}}^{x_i} F(t)\,dt = (-1)^{k-1} \int_{x_{i-1}}^{x_i} (x - x_{i-1})^{k-1} F^{(k-1)}(t)\,dt$$
$$= (-1)^{k-1} h_K^k N_k(\psi) = 0.$$

$\square$

Some properties of our special interpolant will now be derived.

Lemma 2.66. *The special interpolant πu has the following properties:*

$$((u - \pi u)', v_h')_K = 0 \qquad\qquad \forall v_h \in V_h, \qquad\qquad (2.98a)$$
$$|u - \pi u|_{l,K} \le C\, h_K^{k+1-l} |u|_{k+1,K} \qquad \forall u \in H^{k+1}(K), \qquad (2.98b)$$
$$\|u - \pi u\|_{0,\infty,K} \le C\, h_K^{k+1} |u|_{k+1,\infty,K} \qquad \forall u \in W^{k+1,\infty}(K), \qquad (2.98c)$$

for $l = 0, \dots, k+1$ and any $K \in \mathcal{T}_h$.

Proof. We have $u(x_i) = (\pi u)(x_i)$ for $i = 0, 1, \dots, N$ and the orthogonality property $(u - \pi u, w)_K = 0$ for all $w \in P_{k-2}(K)$, so integration by parts gives

$$((u - \pi u)', v_h')_K = -(u - \pi u, v_h'')_K = 0 \quad \text{for all } v_h \in P_k(K).$$

To prove (2.98b), observe first that for $l = 0, \ldots, k + 1$ the mapping $\Phi : H^{k+1}(K) \to H^l(K)$ defined by $\Phi(u) = u - \pi u$ is linear and continuous. Moreover, $\Phi(p) = 0$ for all polynomials $p \in P_k(K)$. Using the Bramble-Hilbert Lemma and the scaling properties of the transformation between K and a reference domain, one gets (2.98b). Finally, (2.98c) is proved similarly by regarding $\Phi(u)$ as a mapping from $W^{k+1,\infty}(K)$ to $L^\infty(K)$. $\square$

Remark 2.67. The property (2.98a) is an equivalent definition of the interpolant π. The projection $R_h : H_0^1(0,1) \to V_h$ with respect to the $H_0^1(\Omega)$ inner product is also defined by $(u' - (R_h u)', v_h') = 0$ for all $v_h \in V_h$; thus $R_h = \pi$. This explains [HS07] why the $H_0^1(0,1)$ projection works well in the VMS framework, unlike the L^2 projection. $\clubsuit$

Lemma 2.68. *In the DRM, assume that for all $K \in \mathcal{T}_h$ one has*

$$0 < \tau_K = \min\left\{1, \frac{h_K}{\varepsilon}\right\} h_K^{2k-1}. \tag{2.99}$$

Let the solution u of (2.84) *belong to $H^{k+1}(0,1)$. Then*

$$|a_h(u - \pi u, v_h)| \le C \left(\varepsilon^{1/2} h^k + h^{k+1/2}\right) |u|_{k+1} |||v_h|||_{DRM} \tag{2.100}$$

for all $v_h \in V_h$ and all $h_K \le h_0$, where h_0 is some threshold value that is independent of ε.

Proof. The choice of τ_K implies that $\tau_K \le h_K^{2k-1}$, so (2.96) is true provided that $h_K \le h_0$, where h_0 is some fixed threshold value. We estimate separately each term in $a_h(u - \pi u, v_h)$. By Lemma 2.66 the first term vanishes. Integrating by parts, the convection term is split as

$$(b(u - \pi u)', v_h) = -(u - \pi u, b v_h') - (b'(u - \pi u), v_h).$$

For the first part of the convection term, Lemma 2.65 with $\psi = u - \pi u$ yields

$$-((u - \pi u), b v_h') = -\sum_{K \in \mathcal{T}_h} (F^{(k-1)}, b v_h')_K = (-1)^k \sum_{K \in \mathcal{T}_h} (F, (b v_h')^{(k-1)})_K.$$

Lemma 2.65 also gives $\|F\|_{0,K} \le C h_K^{k-1} \|u - \pi u\|_{0,K} \le C h_K^{2k} |u|_{k+1,K}$, which leads to

$$|((u - \pi u), b v_h')| \le \sum_{K \in \mathcal{T}_h} \tau_K^{-1/2} \|F\|_{0,K} \|\tau_K^{1/2} (b v_h')^{(k-1)}\|_{0,K}$$

$$\le C \left(\sum_{K \in \mathcal{T}_h} \frac{h_K^{4k}}{\tau_K} |u|_{k+1,K}^2\right)^{1/2} |||v_h|||_{DRM}.$$

The second part of the convection term is put with the reaction term:

$$\left| \left((c - b')(u - \pi u), v_h \right) \right| \leq C \left(\sum_{K \in \mathcal{T}_h} h_K^{2k+2} |u|_{k+1,K}^2 \right)^{1/2} |||v_h|||_{DRM}.$$

This brings us to the stabilizing terms. Here

$$ST = \sum_{K \in \mathcal{T}_h} \tau_K \left(-\varepsilon(u - \pi u)^{(k+1)} \right.$$

$$\left. + (b(u - \pi u)' + c(u - \pi u))^{(k-1)}, (bv_h')^{(k-1)} \right)_K$$

so

$$|ST| \leq \left[2 \sum_{K \in \mathcal{T}_h} \tau_K \left(\varepsilon^2 |u|_{k+1,K}^2 + |b(u - \pi u)' + c(u - \pi u)|_{k-1,K}^2 \right) \right]^{1/2}$$

$$\times |||v_h|||_{DRM}$$

From (2.99) one has $\varepsilon \tau_K \leq C h_K^{2k}$, and the interpolation estimate (2.98b) yields

$$|ST| \leq C \left[\sum_{K \in \mathcal{T}_h} \left(\varepsilon h_K^{2k} + \tau_K h_K^2 \right) |u|_{k+1,K}^2 \right]^{1/2} |||v_h|||_{DRM}.$$

Collecting all the bounds we obtain

$$|a_h(u - \pi u, v_h)|$$

$$\leq C \left[\sum_{K \in \mathcal{T}_h} \left(\varepsilon h_K^{2k} + \tau_K h_K^2 + \tau_K^{-1} h_K^{4k} + h_K^{2k+2} \right) |u|_{k+1,K}^2 \right]^{1/2} |||v_h|||_{DRM}.$$

Now use (2.99) to complete the proof. $\square$

Lemmas 2.63 and 2.68 together give

Theorem 2.69. *Let the coefficients of the differential equation* (2.84) *satisfy* (2.85). *Assume that the solution u lies in $H^{k+1}(0,1)$. Let τ_K be defined by* (2.99). *Then one has the error estimate*

$$|||u - u_h|||_{DRM} \leq C \left(\varepsilon^{1/2} h^k + h^{k+1/2} \right) |u|_{k+1}.$$

Proof. The triangle inequality gives

$$|||u - u_h|||_{DRM} \leq |||u - \pi u|||_{DRM} + |||\pi u - u_h|||_{DRM}.$$

To estimate the interpolation error, use (2.98b) and $\tau_K \leq C\, h_K^{2k-1}$; one gets

$$|||u - \pi u|||_{DRM} \leq C \left(\varepsilon^{1/2} h^k + h^{k+1/2} \right) |u|_{k+1}.$$

The remaining term can be bounded using the coercivity of the bilinear form (Lemma 2.63), Galerkin orthogonality (which follows from (2.92) and (2.93)), and the estimate (2.100) of Lemma 2.68:

$$\frac{1}{2}|||u_h - \pi u|||^2_{DRM} \le a_h(u_h - \pi u, u_h - \pi u) = a_h(u - \pi u, u_h - \pi u)$$

$$\le C\left(\varepsilon^{1/2}h^k + h^{k+1/2}\right)|u|_{k+1} |||u_h - \pi u|||_{DRM}.$$

This completes the argument.   □

In [Tob06] the DRM method is studied on layer-adapted meshes, which will be discussed later.

2.2.5 Uniformly Convergent Finite Element Methods

Uniformly convergent finite element methods have been derived by operator-fitted and mesh-fitted approaches. In this subsection, we discuss exponentially-fitted finite element methods which turn out to be uniformly convergent (note that the definition of uniform convergence in (2.21) can be applied to finite element methods as well as finite difference methods). Layer-adapted meshes for finite difference methods will be studied in Section 2.4, while mesh-fitted finite element methods will be deferred to the multi-dimensional case in Section III.3.5.2.

Consider again the singularly perturbed boundary value problem

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \text{ on } (0,1), \quad u(0) = u(1) = 0, \quad (2.101a)$$

under the assumptions

$$(i) \quad b(x) \ge b_0 > 0, \qquad (ii) \quad c(x) - \frac{1}{2}b'(x) \ge \omega > 0. \qquad (2.101b)$$

The constants b_0 and ω are independent of ε. The assumption (ii) is not a restriction, because if we assume only (i), then the transformation $u \mapsto e^{\sigma x}v$, for a suitably chosen σ that is bounded uniformly in ε, yields a problem in v like (2.101a) for which (i) and (ii) hold.

Discretizations on an equidistant grid with mesh size h are examined here.

For simplicity, suppose first that b is constant with $c \equiv 0$. The solution u of (2.101) has an exponential boundary layer. To get a good approximation to u it is reasonable to use exponential functions related to this layer. Thus as in Section 2.1.3 define L-splines (exponentially-fitted splines) φ_i, where $i = 1, \ldots, N-1$, by

$$-\varepsilon \varphi_i'' + b\varphi_i' = 0 \quad \text{on every open mesh subinterval,}$$
$$\varphi_i(x_j) = \delta_{ij}.$$

Let V_h be the finite element space spanned by the φ_i. Then the discrete problem is:

$$\text{Find } u_h \in V_h \text{ such that } \quad a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h, \qquad (2.102)$$

where $a(\cdot, \cdot)$ is as in Section 2.2.2. This is a *Ritz-Galerkin finite element method with exponentially-fitted splines*.

For the error analysis, let u^I be the nodal interpolant from V_h to the exact solution u of the given problem (2.101); that is,

$$u^I(x_i) = u(x_i) \quad \text{and} \quad u^I \in V_h.$$

As α in (2.63) is independent of ε, we say that the bilinear form $a(\cdot, \cdot)$ is uniformly V-elliptic. Now (2.63) yields

$$\alpha \|u - u_h\|_\varepsilon^2 \le a(u - u_h, u - u_h) = a(u - u_h, u - u^I) + a(u - u_h, u^I - u_h). \quad (2.103)$$

But (2.101a) and (2.102) imply the Galerkin orthogonality property

$$a(u - u_h, u^I - u_h) = 0,$$

so we need only estimate $a(u - u_h, u - u^I)$. For this,

$$|\varepsilon(\nabla(u - u_h), \nabla(u - u^I))| \le \varepsilon^{1/2} \|u - u_h\|_1 \, \varepsilon^{1/2} \|u - u^I\|_1$$
$$\le \gamma_1 \varepsilon \|u - u_h\|_1^2 + C(\gamma_1) \|u - u^I\|_\varepsilon^2, \qquad (2.104)$$

where γ_1 is a constant that is chosen later, the norm $\|\cdot\|_\varepsilon$ was defined in (2.62), and the generalized arithmetic-geometric mean inequality has been used in the calculation:

$$yz \le \gamma y^2 + \frac{z^2}{4\gamma}, \qquad (2.105)$$

for all $\gamma > 0$ and all $y, z \in R$.

Before moving on to the convection term in $a(u - u_h, u - u^I)$, some control of the interpolation error is required.

Lemma 2.70. *Let u^I be the interpolant to u in the space V_h of exponentially-fitted splines. Then the interpolation error $u - u^I$ satisfies the bounds*

$$\|u - u^I\|_\infty \le Ch \quad and \quad \|u - u^I\|_\varepsilon \le Ch^{1/2}. \qquad (2.106)$$

Proof. Set $Mz := -\varepsilon z'' + bz'$. On each interval (x_{i-1}, x_i), one has

$$M(u - u^I) = f, \quad (u - u^I)(x_{i-1}) = 0, \quad (u - u^I)(x_i) = 0.$$

Apply the comparison principle of Lemma 1.1 on each $[x_{i-1}, x_i]$, with $w(x) = C^*(x - x_{i-1})$ for some suitable positive constant C^*; this gives

$$|(u - u^I)(x)| \le Ch \quad \forall x \in [0, 1].$$

For the other bound on $u - u^I$, observe that

$$\alpha \|u - u^I\|_\varepsilon^2 \le a(u - u^I, u - u^I)$$

$$= \sum_i \int_{x_{i-1}}^{x_i} [-\varepsilon(u - u^I)'' + b(u - u^I)'](u - u^I)dx .$$

But $-\varepsilon(u - u^I)'' + b(u - u^I)' = f$ on each (x_{i-1}, x_i), so using the estimate on $\|u - u^I\|_\infty$ above yields the desired result. □

We now return to the convection term in $a(u - u_h, u - u^I)$. Since

$$(b(u - u^I)', u - u^I) = 0,$$

one obtains

$$|(b(u - u_h)', u - u^I)| = |(b(u^I - u_h)', u - u^I)|$$

$$\le C\|(u^I - u_h)'\|_{L_1} \|u - u^I\|_\infty.$$

As $u^I - u_h$ lies in V_h, the following paraphrase of a lemma from [OS91a, OS91b] allows us an economical replacement of the L_1 norm by the L_2 norm:

Lemma 2.71. *Let V_h be the space of L-splines. For each $v_h \in V_h$, one has*

$$\|v_h'\|_{L_1} \le Ch^{-1/2}\varepsilon^{1/2}\|v_h'\|_{L_2}.$$

Applying Lemmas 2.70 and 2.71 to the previous inequality gives

$$|(b(u - u_h)', u - u^I)| \le Ch\ h^{-1/2}\varepsilon^{1/2}\|(u^I - u_h)'\|_{L_2}$$

$$\le Ch^{1/2}\varepsilon^{1/2}(|u - u_h|_1 + |u^I - u|_1)$$

$$\le \gamma_2\varepsilon|u - u_h|_1^2 + C(\gamma_2)h + Ch, \qquad (2.107)$$

by (2.106) and (2.105). Now choose $\gamma_1 = \gamma_2 = \alpha/4$ and combine (2.103), (2.104) and (2.107) to prove:

Lemma 2.72. *In the case of constant b and $c \equiv 0$, the error of the Ritz-Galerkin L-spline finite element method satisfies the (uniform in ε) estimate*

$$\|u - u_h\|_\varepsilon \le Ch^{1/2}. \qquad (2.108)$$

Remark 2.73. (Optimality of convergence rate) In the norm $\|\cdot\|_\varepsilon$, the order of convergence (uniformly in ε) that was proved in (2.108) is in fact optimal. For consider the solution of

$$-\varepsilon z'' + z' = 1, \quad z(0) = z(1) = 0.$$

A direct computation gives

$$\varepsilon^{1/2}|z - z^I|_1 = \varepsilon^{1/2} \left[\frac{h}{2\varepsilon} \coth\left(\frac{h}{2\varepsilon}\right) - 1 \right]^{1/2},$$

where z^I is the L-spline interpolant. If $\varepsilon = h$, then $\|z - z^I\|_\varepsilon = \mathcal{O}(h^{1/2})$, which shows that, uniformly in ε, no higher order is possible. ♣

When b is not constant and $c \not\equiv 0$, we approximate b, c and f by piecewise constants as in Section 2.1.3; that is, we set

$$\bar{b} := [b(x_{i-1}) + b(x_i)]/2 \quad \text{on each mesh subinterval } (x_{i-1}, x_i),$$

with similar formulas for c and f. The space of exponential splines ($\bar{L}$-splines) is now spanned by the basis functions φ_i that satisfy

$$-\varepsilon \varphi_i'' + \bar{b} \varphi_i' + \bar{c} \varphi_i = 0 \quad \text{on every open mesh subinterval,} \qquad (2.109)$$
$$\varphi_i(x_j) = \delta_{ij}.$$

A typical $\bar{L}$-spline (for the case $\bar{c} = 0$) is drawn in the first diagram of Figure 2.4 on page 109. The associated modified bilinear form is defined by

$$a_h(v, w) := \varepsilon(v', w') + (\bar{b}v', w) + (\bar{c}v, w). \qquad (2.110)$$

Instead of using *complete exponential fitting* based on the splines (2.109), we use *partial exponential fitting* with the simpler splines φ_i defined by

$$-\varepsilon \varphi_i'' + \bar{b} \varphi_i' = 0 \quad \text{on every open subinterval,} \qquad (2.111)$$
$$\varphi_i(x_j) = \delta_{ij}.$$

Furthermore, a lumping process is used to simplify the three-point difference approximation for the term $cu - f$. Thus one replaces (2.110) by the discrete bilinear form

$$\bar{a}_h(v, w) := \varepsilon(v', w') + (\bar{b}v', w) + h \sum_i (cvw)(x_i), \qquad (2.112)$$

and, using the splines (2.111), one arrives at the following discrete problem:

$$\text{Find } u_h \in V_h \text{ such that} \quad a_h(u_h, v_h) = h \sum_i (fv_h)(x_i) \quad \forall v_h \in V_h. \quad (2.113)$$

To analyse this method, it is natural in the ε-weighted H^1 norm $\| \cdot \|_\varepsilon$ to replace the L_2 part $|v|_0$ by its discrete analogue $|v|_{0,d}$, so we set

$$\|v\|_{\varepsilon,d}^2 := \varepsilon |v|_1^2 + |v|_{0,d}^2 \quad \text{with } |v|_{0,d}^2 := h \sum_i v^2(x_i). \qquad (2.114)$$

Then the bilinear form $\bar{a}_h(\cdot, \cdot)$ turns out to be uniformly V_h-elliptic with respect to $\| \cdot \|_{\varepsilon,d}$ and the interpolation result $\|u - u^I\|_{\varepsilon,d} \leq Ch^{1/2}$ (cf. (2.106)) holds true. One can prove (see [SO91])

Theorem 2.74. *Let u_h be the solution of the lumped finite element discretization (2.113), where V_h is the space of partially exponentially-fitted L-splines. Then the error $u - u_h$ satisfies the uniform estimate*

$$\|u - u_h\|_{\varepsilon,d} \leq Ch^{1/2}. \qquad (2.115)$$

Remark 2.75. The continuous and discrete L_2 norms are equivalent, uniformly in ε, on the space of exponentially-fitted $\bar{L}$-splines. Using this equivalence, the norm $\| \cdot \|_{\varepsilon,d}$ in Theorem 2.74 can be replaced by the norm $\| \cdot \|_{\varepsilon}$.   ♣

Next we consider pointwise error estimates in finite element methods for (2.101). Let the trial space V_h be arbitrary; its basis functions φ_i are required to satisfy only

$$\operatorname{supp} \varphi_i = [x_{i-1}, x_{i+1}] \quad \text{and} \quad \varphi_i(x_j) = \delta_{ij}. \tag{2.116}$$

Petrov-Galerkin discretizations will be investigated where the test space T_h is, for the moment, arbitrary:

$$\text{Find } u_h \in V_h \text{ such that} \quad a_h(u_h, v_h) = (\bar{f}, v_h) \quad \forall v_h \in T_h, \tag{2.117}$$

where $a_h(\cdot, \cdot)$ is as in (2.110). Recall the representation for the pointwise error (in terms of a Green's function) that lead to (2.60). Hemker [Hem77] points out that *the test space of a Petrov-Galerkin method should be related to the Green's function of the problem.* We describe a test space that permits good approximation of the Green's function and give a simple derivation of uniform pointwise error estimates.

Let x_j be a given grid point. One would like to define a *discrete Green's function* $G_h(x, x_j)$ by

$$a_h(w, G_h) = w(x_j) \quad \forall w \in H_0^1(0, 1).$$

Equivalently, $G_j(\cdot) := G_h(\cdot, x_j)$ is characterized by the conditions

(i) in each open mesh subinterval, $G_j(\cdot)$ satisfies the equation

$$-\varepsilon G_j'' - \bar{b} G_j' + \bar{c} G_j = 0;$$

(ii) $G_j(\cdot)$ is continuous, with $G_j(0) = G_j(1) = 0$;
(iii) G_j satisfies the following jump condition at each inner grid point:

$$\lim_{x \to x_i - 0} (\varepsilon G_j' + \bar{b} G_j) - \lim_{x \to x_i + 0} (\varepsilon G_j' + \bar{b} G_j) = \delta_{ij}.$$

Using (i), (ii) and (iii), one can show that G_j is uniquely determined [SO86]. The weak maximum principle shows that G_j is uniformly bounded with respect to ε in the maximum norm.

Assuming that G_j belongs to our testspace T_h, one has the following representation for the pointwise error:

$$(u - u_h)(x_j) = a_h(u - u_h, G_j)$$
$$= (a_h - a)(u, G_j) + (f - \bar{f}, G_j). \tag{2.118}$$

To ensure that $G_j \in T_h$, define T_h to be the span of the basis functions ψ_k, for $k = 1, ..., N - 1$, where

$$-\varepsilon\psi_k'' - \bar{b}\psi_k' + \bar{c}\psi_k = 0 \quad \text{on every open mesh subinterval,} \qquad (2.119)$$
$$\psi_k(x_j) = \delta_{kj} \quad \text{for } j = 0, ..., N.$$

Functions belonging to T_h are called $\bar{L}^*$-splines, where L^* denotes the formal adjoint of L. See the second diagram of Figure 2.4 for a typical example.
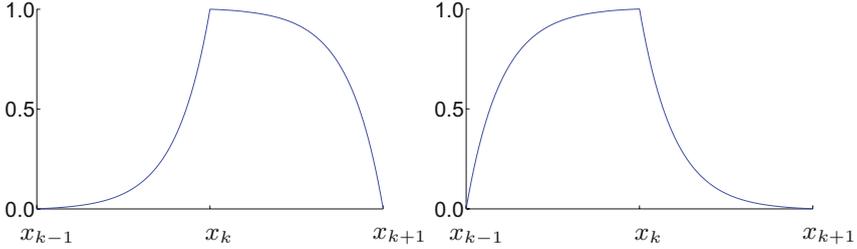
Fig. 2.4. $\bar{L}$-splines and $\bar{L}^*$-splines for $\bar{b} = 1$, $\bar{c} = 0$, and $\varepsilon/h = 0.2$

Theorem 2.76. *Let $\bar{b}, \bar{c}$ and $\bar{f}$ be $\mathcal{O}(h)$-approximations of b, c and f respectively. If arbitrary trial functions and $\bar{L}^*$-spline test functions are used in the Petrov-Galerkin method (2.117), then the error satisfies*

$$\|u - u_h\|_{\infty,d} \le Ch. \qquad (2.120)$$

Proof. From (2.118) one has

$$(u - u_h)(x_i) = (f - \bar{f}, G_i) + (u', (\bar{b} - b)G_i) + (u, (\bar{c} - c)G_i).$$

Since $\|u\|_\infty \le C$, $\|G_i\|_\infty \le C$ and $\|u'\|_{L_1} \le C$, we get immediately the desired uniform convergence result. $\qquad \square$

Which discrete problem is generated by this method? The Petrov-Galerkin method produces a tridiagonal system of equations in the unknowns u_i, where we set $u_h(x) = \sum_{i=1}^{N-1} u_i\varphi_i(x)$ and $u_i = u_h(x_i)$. This system can be written in the form

$$\alpha_{-1,k}u_{k-1} + \alpha_{0,k}u_k + \alpha_{1,k}u_{k+1} = f_k^*, \qquad (2.121)$$

where, for instance,

$$\begin{aligned}
\alpha_{-1,k} &= \int_{x_{k-1}}^{x_k} [\varepsilon\varphi_{k-1}'\psi_k' + (b_k\varphi_{k-1}' + c_k\varphi_{k-1})\psi_k] \\
&= \varepsilon\varphi_{k-1}\psi_k'|_{x_{k-1}}^{x_k} + b_k\varphi_{k-1}\psi_k|_{x_{k-1}}^{x_k} \\
&\quad + \int_{x_{k-1}}^{x_k} [-\varepsilon\psi_k'' - b_k\psi_k' + c_k\psi_k]\varphi_{k-1} \\
&= -\varepsilon\psi_k'(x_{k-1}). \qquad (2.122)
\end{aligned}$$

From (2.122) and the corresponding formulas for $\alpha_{0,k}$ and $\alpha_{1,k}$, one easily derives

Lemma 2.77. *The difference scheme generated by the Petrov-Galerkin discretization (2.117) with $\bar{L}^*$-splines as test functions is independent of the choice of the trial functions.*

Remark 2.78. A comparison of (2.122) with the formulas in Remark 2.23 shows that the scheme generated is closely related to the Il'in-Allen-Southwell scheme or to the El-Mistikawy-Werle scheme, depending on the choice of $\bar{b}$, $\bar{c}$ and $\bar{f}$. For the latter scheme, one can use (2.118) to show [SO86] that the error at the grid points is in fact $\mathcal{O}(h^2)$. ♣

Lemma 2.77 has the following analogue: the difference scheme generated by the Petrov-Galerkin discretization (2.117) with $\bar{L}$-spline trial functions is independent of the choice of the test functions. From this result and Lemma 2.77 we make the following deduction: if $\bar{L}^*$-splines are used in both the trial and the test space, the resulting difference scheme would coincide with that obtained using $\bar{L}$-splines in both the trial and the test space. Similarly, if one takes piecewise linear trial functions and $\bar{L}^*$-spline test functions, the resulting difference scheme coincides with that generated by $\bar{L}$-spline trial functions and piecewise linear test functions. (But changes of this type do affect the discretization of f.) As a consequence we get

Corollary 2.79. *Let $\bar{b}, \bar{c}$ and $\bar{f}$ be $\mathcal{O}(h)$ approximations of b, c and f respectively. If $\bar{L}$-splines are used as both trial and test functions in the finite element discretization (2.113), then the error satisfies*

$$\|u - u_h\|_{\infty,d} \le Ch.$$

Theorem 2.76 and Corollary 2.79 analyse the error in the *discrete* maximum norm. If one desires uniform $L_\infty[0, 1]$ estimates, one should choose $\bar{L}$-splines as trial functions and deduce the continuous L_∞ bound from the discrete L_∞ bound.

Corollary 2.80. *Assume that $u_h \in V_h$, where V_h is the space of $\bar{L}$-splines with basis functions specified by (2.109). If*

$$\|u - u_h\|_{\infty,d} \le Ch$$

then

$$\|u - u_h\|_\infty \le Ch.$$

Proof. We establish this result on each mesh subinterval. Set

$$\bar{L}w := -\varepsilon w'' + \bar{b}w' + \bar{c}w \quad \text{on } (x_{i-1}, x_i).$$

For the error $e := u - u_h$, one has

$$\bar{L}e = (\bar{b} - b)u' + (\bar{c} - c)u \quad \text{on } (x_{i-1}, x_i)$$

and

$$|e(x_{i-1})| \leq Ch, \qquad |e(x_i)| \leq Ch.$$

Now apply the comparison principle, using an exponential barrier function as in the proof of Theorem 1.8, to obtain the desired result. □

As the boundary layer is contained in a neighbourhood of $x = 1$, is exponential fitting needed on *all* of [0,1]? Let us make the mild assumption that ε satisfies $4\varepsilon \ln(1/\varepsilon) < b_0$, and set

$$M = \max \left\{ i \ : \ x_i \leq 1 - \frac{2\varepsilon}{b_0} \ln \frac{1}{\varepsilon} \right\}.$$

Then Lemma 1.8 implies that

$$|u'| \leq C \ \text{ and } \ |u''| \leq C \ \text{ on } (0, x_M).$$

Call $[x_M, 1]$ the *layer region*. Outside $[x_M, 1]$, let us now use piecewise linear functions in both trial and the test spaces. For $i = M, M + 1, \ldots, N - 1$, we rely on our usual partially-fitted $\bar{L}$-splines as basis functions for the trial space. Thus the trial space S_h consists of linear trials on $[0, x_M]$ and $\bar{L}$-splines on $[x_M, 1]$; in particular, φ_M is a hybrid linear/$\bar{L}$-spline.

Lemma 2.81. *Let $u^I \in S_h$ interpolate to the exact solution of the boundary value problem (2.101) at each node. Then for $x \in [x_{i-1}, x_i]$,*

$$|(u - u^I)(x)| \leq Ch^2 \qquad \text{if} \quad 1 \leq i \leq M,$$
$$|(u - u^I)(x)| \leq Ch(1 - e^{-\beta h/\varepsilon}) \quad \text{if} \quad M < i \leq N,$$

where the constant $\beta > 0$ is independent of ε.

This precise interpolation result from [SO91] gives some hope that if exponential fitting is used only in the layer region, it will work satisfactorily. Lemma 2.81 enables one to prove that a combination of $\bar{L}$-spline trials and $\bar{L}^*$-spline tests in the layer region, with piecewise linear trials and tests elsewhere, leads to uniform convergence in an energy norm [SO91].

Remark 2.82. Petrov-Galerkin finite element methods can be reformulated as *mixed finite element methods*. For consider the Petrov-Galerkin method based on (2.110) with a trial space S_h of partially-fitted $\bar{L}$-splines and a test space T_h of partially-fitted $\bar{L}^*$-splines. To explain the new approach, define $l_h(\cdot, \cdot)$ by

$$l_h(v, w) = \varepsilon(v', w') + (\bar{b}v', w)$$

and introduce the bilinear form

$$\eta(v, q) = h \sum_{i=1}^{N-1} q_i v(x_i) \quad \text{on } H_0^1 \times \mathbb{R}^{N-1}.$$

Then the function ϕ_h is in S_h if and only if some multiplier $p_h \in Q := \mathbb{R}^{N-1}$ satisfies

$$l_h(\phi_h, w) + \eta(w, p_h) = 0 \quad \forall w \in V := H_0^1(0,1). \tag{2.123}$$

This follows easily from an integration by parts:

$$l_h(\phi_h, w) = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (-\varepsilon \phi_h'' + \bar{b}\phi_h') w \, dx - \sum_{i=1}^{N-1} \varepsilon [\phi_h'(x_i^+) - \phi_h'(x_i^-)] w(x_i).$$

One can analogously characterize partially-fitted $\bar{L}^*$-splines v_h by

$$l_h(w, v_h) + \eta(w, p_h^*) = 0 \quad \forall w \in V \tag{2.124}$$

for some $p_h^* \in Q$.

Now, instead of defining $u_h \in S_h$ by

$$a_h(u_h, v_h) = (\bar{f}, v_h) \quad \forall v_h \in T_h,$$

we pose the discrete problem in the following way:

Find u_h and $w_h \in V$, and $p_h \in Q$, such that

$$a_h(u_h, v) + l_h(w_h, v) = (\bar{f}, v) \qquad \forall v \in V, \tag{2.125a}$$
$$\eta(w_h, q) = 0 \qquad \forall q \in Q, \tag{2.125b}$$
$$l_h(u_h, v) + \eta(v, p_h) = 0 \qquad \forall v \in V. \tag{2.125c}$$

As (2.125c) is identical to (2.123), it implies $u_h \in S_h$. Next, if $v_h \in T_h$, then by (2.124) one has a multiplier p_h^* with $l_h(w_h, v_h) = -\eta(w_h, p_h^*) = 0$, using (2.125b). Therefore (2.125) produces the standard Petrov-Galerkin formulation

$$a_h(u_h, v_h) = (\bar{f}, v_h) \quad \forall v_h \in T_h.$$

The main advantage of (2.125) over the standard formulation is that (2.125) holds true over V and Q and *not* over subspaces. This opens the door to a new straightforward error analysis [Fel94] that yields a result similar to Theorem 2.74. ♣

The construction of uniformly convergent finite element approximations of higher order on standard meshes is an open problem. On an equidistant mesh, one could use trial and test spaces enriched by additional polynomial or exponential functions or both (compare the HODIE technique of Section 2.1.4).

For small values of the parameter ε, de Groen and Hemker [dG81, dGH79] construct exponentially-fitted higher-order methods. They use finite-dimensional spaces spanned by polynomials of degree k and $\bar{L}$-splines or $\bar{L}^*$-splines, which we denote by E_k^h and F_k^h respectively. In the case where E_k^h is both trial and test space, they show that

$$\|u - u_h\|_\varepsilon \le C(\varepsilon + h^k), \qquad \|u - u_h\|_{\infty,d} \le C(\varepsilon + h^k),$$

while if E_k^h is the trial space and F_k^h the test space and $h + \varepsilon/h \le C$, then

$$\|u - u_h\|_\varepsilon \le C(\varepsilon + h^k), \qquad \|u - u_h\|_{\infty,d} \le C(\varepsilon^2 + h^{2k}).$$

We wrap up this section with some remarks on the semilinear problem

$$-\varepsilon u'' + b(x)u' + c(x, u) = 0 \quad \text{for } x \in (0, 1), \tag{2.126a}$$
$$u(0) = A, \ u(1) = B, \tag{2.126b}$$

under the assumption that $c_s(x, s) \ge \delta > 0$ on $[0, 1] \times \mathbb{R}$. Recall from Section 1.3 that the theory of upper and lower solutions shows that this problem has a unique solution $u(x)$, and that

$$\|u\|_\infty \le C \quad \text{and} \quad \int_0^1 |u'(x)| \, dx \le C.$$

To discretize (2.126), let us apply a Petrov-Galerkin finite element method with lumping, based on a trial space V_h that need satisfy only (2.116), with partially-fitted $\bar{L}^*$-splines as test functions. The discrete problem is:
   Find $u_h \in V_h$ such that

$$\varepsilon(u_h', \psi_i') + (\bar{b}u_h', \psi_i) + h\, c(x_i, u_h(x_i)) = 0 \quad \text{for } i = 1, ..., N - 1,$$
$$u_h(0) = A, \ u_h(1) = B,$$

where ψ_i satisfies

$$-\varepsilon\psi_i'' - \bar{b}\psi_i' = 0, \quad \psi_i(x_j) = \delta_{ij}.$$

The system of equations that this generates in the unknowns $u_h(x_i) = u_i$ is independent of the actual trial space (cf. Lemma 2.77). It takes the form

$$-\frac{\varepsilon}{h^2}\left(\theta_{i+1}u_{i+1} - (\theta_{i+1} + \theta_i)u_i + \theta_i u_{i-1}\right) + c(x_i, u_i) = 0, \tag{2.127a}$$
$$u_0 = A, \ u_N = B, \tag{2.127b}$$

with $\rho_i = b_i h/\varepsilon$ and

$$\theta_i = \theta(\rho_i), \quad \theta(x) = \begin{cases} \dfrac{x}{1 - e^{-x}} & \text{for } x \ne 0, \\ 1 & \text{for } x = 0. \end{cases}$$

We have already met this scheme (for linear problems) in (2.38). A detailed analysis [SO87] shows that:

— The piecewise linear function U_h that interpolates to the computed solution $u_0, u_1, ..., u_N$ satisfies $\|u - U_h\|_{L_1} \le Ch$.
— In the case $b(x) \ge b_0 > 0$ (i.e., no turning points), the maximum nodal error of the scheme is bounded by Ch.
— In the case of an interior turning point with a cusp layer, the maximum nodal error is bounded by Ch^λ (in the notation of Section 2.1.5).

At present the literature contains few results dealing with fitted finite element methods on standard meshes for nonlinear singular perturbation problems; on layer-adapted meshes the situation is different.

2.3 Finite Volume Methods

Finite volume methods stem from an integral balance equation over a control volume. Such an equation can often be interpreted as a conservation law for some physical unknown. If the conservation law is a natural description of the process under consideration – e.g., in certain fluid dynamics problems – then finite volume methods are often highly successful in computing approximate solutions, partly because they preserve this property on the discrete level.

The essential idea of all finite volume methods is to partition the domain into small regions called *control volumes* or *cells* or *boxes*, integrate the differential equation over each cell separately, then use the Gauss divergence theorem to convert each cell integral of the derivatives into an integral over the surface of that cell. The numerical analyst then chooses a suitable approximation of these surface integrals, thereby obtaining a difference scheme.

Consider a singularly perturbed boundary value problem in conservation form:

$$-\varepsilon u'' + (b(x)u)' + c(x)u = f(x) \text{ for } 0 < x < 1, \quad u(0) = u(1) = 0, \quad (2.128)$$

with $b(x) \geq \beta > 0$ (i.e., no turning points) and with $c(x) \geq 0$. Suppose we have an arbitrary mesh

$$0 = x_0 < x_1 < ... < x_{N-1} < x_N = 1.$$

Set $x_{i+1/2} = (x_i + x_{i-1})/2$ for $i = 0, \ldots, N-1$. Then

$$0 < x_{1/2} < x_{3/2} < ... < x_{N-1/2} < 1$$

is a *secondary grid*. It defines the cells $(0, x_{1/2}), (x_{1/2}, x_{3/2}), \ldots, (x_{N-1/2}, 1)$. Integrating the differential equation over a typical cell gives the balance equation

$$-\varepsilon u'\big|_{x_{i-1/2}}^{x_{i+1/2}} + (bu)(x_{i+1/2}) - (bu)(x_{i-1/2}) + \int_{x_{i-1/2}}^{x_{i+1/2}} cu \, dx = \int_{x_{i-1/2}}^{x_{i+1/2}} f \, dx.$$
$$(2.129)$$

Let u_i^N be the computed approximation of $u(x_i)$ for $i = 0, \ldots, N$. In the discretization step that we now discuss, the integrals in the balance equation (2.129) are approximated in a standard way and the values of u' at each secondary mesh point are expressed in terms of the u_i^N using finite differences.

Suppose for simplicity that the original (primary) mesh is equidistant, i.e., $x_i = ih$ for $i = 0, \ldots, N$ with $h = 1/N$.

Then one can choose the approximations

$$u'(x_{i+1/2}) \approx \frac{u_{i+1}^N - u_i^N}{h}, \qquad u'(x_{i-1/2}) \approx \frac{u_i^N - u_{i-1}^N}{h},$$

$$g(x_{i\pm1/2}) \approx \frac{g(x_i) + g(x_{i\pm1})}{2}, \qquad \int_{x_{i-1/2}}^{x_{i+1/2}} g(x) \, dx \approx g(x_i) \cdot h,$$

with some reasonable variant for $u'(0)$ and $u'(1)$. In the case of constant $b(\cdot)$, this leads to the central difference scheme

$$-\varepsilon \frac{u_{i+1}^N - 2u_i^N + u_{i-1}^N}{h^2} + b \frac{u_{i+1}^N - u_{i-1}^N}{2h} + c_i u_i^N = f_i,$$

where $c_i = c(x_i)$ and $f_i = f(x_i)$, which we know to be unsuitable for singularly perturbed problems.

To obtain a more stable scheme, one must give an upwind approximation of the convection terms $(bu)(x_{i\pm 1/2})$ in (2.129). Thus consider the approximation

$$(bu)(x_{i+1/2}) \approx b(x_{i+1/2})\big[\lambda_i u_{i+1}^N + (1 - \lambda_i)u_i^N\big]$$

with $0 \le \lambda_i \le 1/2$. If $\lambda_i = 1/2$, one has again the central difference scheme, while the choice $\lambda_i = 0$ yields the simple upwind scheme.

Values of λ_i between 0 and 1/2 allow us to vary the amount of upwinding. For constant $b(\cdot)$, one can in this way generate the the class of fitted schemes (2.14) of Section 2.1.2, where the fitting parameter σ and the weighting parameter λ are related by

$$\sigma_i = 1 + \frac{h}{2\varepsilon}b(1 - 2\lambda_i).$$

Hence each result for a fitted scheme yields a corresponding result for a finite volume scheme.

Remark 2.83. Finite volume methods that use secondary grids to define the integration cells are called *cell-centered* methods. One could instead use the cells defined by the original grid, and finite volume methods of this type are called *cell-vertex* methods. Cell-vertex methods are much less popular, owing to their lack of stability [BS97]; Morton [Mor96] gives a detailed account of their theory and practice. ♣

Finally, we show how to generate the Il'in-Allen-Southwell scheme of Section 2.1.3 by means of a cell-centered finite volume method. For simplicity, set $c \equiv 0$ and take b to be constant. Then (2.128) can be written in the form

$$-\varepsilon(e^{-bx/\varepsilon}u')' = e^{-bx/\varepsilon}f,$$

Integration over a cell yields

$$-\varepsilon(e^{-bx/\varepsilon}u')|_{x_{i-1/2}}^{x_{i+1/2}} = \int_{x_{i-1/2}}^{x_{i+1/2}} e^{-bx/\varepsilon} f(x)\,dx.$$

Taking

$$\int_{x_{i-1/2}}^{x_{i+1/2}} e^{-bx/\varepsilon} f(x)\,dx \approx f_i \int_{x_{i-1/2}}^{x_{i+1/2}} e^{-bx/\varepsilon}dx$$

produces the scheme

$$-\varepsilon e^{-bx_{i+1/2}/\varepsilon}\frac{u_{i+1}^{N}-u_{i}^{N}}{h}+\varepsilon e^{-bx_{i-1/2}/\varepsilon}\frac{u_{i}^{N}-u_{i-1}^{N}}{h}=$$

$$-\frac{f_{i}\varepsilon}{b}\left(e^{-bx_{i+1/2}/\varepsilon}-e^{-bx_{i-1/2}/\varepsilon}\right),$$

and this can be rewritten as the Il'in-Allen-Southwell scheme.

Remark 2.84. This derivation is routine in the field of semiconductor device modelling, where the Il'in-Allen-Southwell scheme is usually called the *Scharfetter-Gummel scheme*, which appeared already in Remark 2.22. See [Gar93] for a detailed proof of the uniform convergence of the scheme when applied to the basic equations of semiconductor physics in the one-dimensional case, or [Sel84] for a general introduction to this topic which includes finite volume discretizations. ♣

In [LMV96] various upwind discretizations of convection-diffusion problems are analysed and references to the literature are given.

The exposition of finite volume methods in this short section may give the impression that they are merely a variant of finite difference methods, but this is misleading: in multi-dimensional problems, finite volume methods are quite distinct from finite difference methods, as we shall see in later chapters.

2.4 Finite Difference Methods on Layer-adapted Grids

Solutions of singularly perturbed boundary value problems change abruptly in layers. Consequently discretization methods on equidistant meshes have difficulty in representing these solutions, and only elaborate schemes based on exponential fitting yield nodal convergence that is uniform with respect to the perturbation parameter; see Theorem 2.17. An alternative strategy to follow when computing boundary and interior layers is the use of *highly nonequidistant grids*. There are two main classes of such grids: one may *a priori choose a special mesh* based on knowledge of the behaviour of the exact solution, or one may begin with some unexceptional mesh, compute an approximate solution there, then use information from this computation to *adapt the grid a posteriori*, thereby obtaining a mesh more suited to the nature of the problem. The present section is devoted to the former approach while the latter will be presented in the following section.

Consider an arbitrary grid

$$0 = x_0 < x_1 < ... < x_{N-1} < x_N = 1.$$

Set $h_i = x_i - x_{i-1}$ for each i. Define the mesh diameter by $h := \max_i h_i$.

A family S of grids is called *quasi-equidistant* if there exists some constant K such that for each grid in S one has

$$h \leq K \min_i h_i,$$

where K is independent of the grid. A family S of grids is said to be *locally quasi-equidistant* if for each grid in S we have

$$h_i \leq K h_j \quad \text{for } |i - j| \leq 1, \tag{2.130}$$

where K is independent of the grid. Most analyses of finite difference and finite volume methods are carried out on quasi-equidistant or locally quasi-equidistant grids.

It is sometimes convenient if $|h_{i+1} - h_i| \leq K h^2$ for each i. This condition is satisfied on *locally almost equidistant grids*, which are defined by

$$h_i \leq h_j (1 + K h_j) \quad \text{for } |i - j| \leq 1.$$

Set $\hbar_i = (h_i + h_{i+1})/2$ for each i. Write u_i for $u(x_i)$. For first-order derivatives, $u'(x_i) \approx D_- u_i := (u_i - u_{i-1})/h_i$ is the backward divided difference approximation; $u'(x_i) \approx D_+ u_i := (u_{i+1} - u_i)/h_{i+1}$ is the forward difference approximation; and the central difference approximation of $u'(x_i)$ is the weighted average

$$D^0 u_i := \frac{1}{2\hbar_i}(h_i D_+ u_i + h_{i+1} D_- u_i). \tag{2.131}$$

The standard finite difference approximation of the second-order derivative is

$$u''(x_i) \approx \delta^2 u_i := \frac{1}{\hbar_i}(D_+ u_i - D_- u_i) = \frac{1}{\hbar_i}\left(\frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}\right). \tag{2.132}$$

For classical problems where the derivatives of u are bounded, (2.131) is second-order consistent on any mesh (i.e., $|u'(x_i) - D^0 u_i| = \mathcal{O}(\hbar_i^2)$), just as on equidistant meshes. The same is not true of (2.132): one has

$$u''(x_i) - \delta^2 u_i = \frac{h_i - h_{i+1}}{3} + \mathcal{O}(\hbar_i^2),$$

which is only first-order on arbitrary meshes. (In fact $u''(\bar{x}_i) - \delta^2 u_i = \mathcal{O}(\hbar_i^2)$, where $\bar{x}_i = (x_{i-1} + x_i + x_{i+1})/3$; see [Mat02] for a list of references that exploit this property.) Nevertheless, when the central difference scheme based on (2.131) and (2.132) is applied to non-singularly perturbed second-order two-point boundary value problems on arbitrary meshes, its order of convergence is still two! This enhancement of performance is called *supraconvergence* by Kreiss et al. [KMS+86], but was known much earlier [TS62]. The proof of second-order convergence is easy on locally almost equidistant grids, but becomes more difficult for arbitrary grids.

Special meshes can be constructed a priori in essentially three ways:

- A *mesh generating function* $\lambda : [0, 1] \to [0, 1]$ is a continuous and strictly increasing function with $\lambda(0) = 0$ and $\lambda(1) = 1$. It induces a mesh on $[0, 1]$ containing $N + 1$ points explicitly defined by

$$x_i = \lambda(i/N) \quad \text{for } i = 0, 1..., N.$$

If λ has additional smoothness properties, this ensures that the induced grid has special properties. For example, if $|\lambda''(x)| \leq K$ on [0,1] then the grid is locally almost equidistant.
Meshes generated in this way will be studied in Sections 2.4.1 and 2.4.2.

- A *monitor function* $M(x)$ is an arbitrary non-negative function defined on $[0, 1]$. It induces a mesh $\{x_i\}_{i=1}^N$ that is implicitly defined by the equidistribution property

$$\int_{x_{i-1}}^{x_i} M(x)\, dx = \frac{1}{N} \int_0^1 M(x)\, dx \quad \text{for } i = 1, \dots, N.$$

This approach also underpins the adaptive strategy discussed in Section 2.5.

- The mesh can be defined implicitly by a *recursive formula*. For example, to deal with an exponential boundary layer at $x = 0$, in [DL06] one finds the recipe

$$\begin{cases} x_i = i\sigma h\varepsilon & \text{for } 0 \leq i \leq 1 + (\sigma h)^{-1}, \\ x_{i+1} = (1 + \sigma h)x_i & \text{for } 1 + (\sigma h)^{-1} \leq i \leq N - 2, \\ x_N = 1, \end{cases}$$

where N is such that $x_{N-1} < 1 \leq (1 + \sigma h)x_{N-1}$; here h and σ are user-chosen positive parameters. See also [Gar88].

We now examine the construction of meshes suitable for convection-diffusion problems. Consider the usual two-point boundary value problem

$$-\varepsilon u'' + b(x)u' + c(x)u = f(x) \text{ for } 0 < x < 1, \quad u(0) = u(1) = 0, \quad (2.133)$$

with $b(\cdot) > \beta > 0$ and $c(\cdot) \geq 0$. The following two subsections present the two main classes of meshes in current use: graded meshes, which become gradually finer as one moves further into the boundary layer, and piecewise uniform meshes, where the change from coarse to fine mesh is sudden.

Our aim in these subsections will be to construct and analyse methods that are *uniformly convergent* in the discrete maximum norm; that is, the computed solution $\{u_i^N\}_{i=0}^N$ satisfies

$$\|u - u^N\|_{\infty,d} := \max_{i=0,\dots,N} |u_i - u_i^N| \leq CN^{-\alpha} \quad (2.134)$$

for some positive constants C and α that are independent of ε and of N. A power of N is a suitable measure of the error $u - u^N$ for the particular families of meshes that are discussed in Section 2.4, but a bound of this type is inappropriate for an arbitrary family of meshes; see [SW96].

2.4.1 Graded Meshes

Bakhvalov [Bak69] was the first person to use a special grid to solve a singu-
larly perturbed boundary value problem. His mesh was designed for reaction-
diffusion problems like those of Remark 1.10, but the technique is easily mod-
ified to suit convection-diffusion problems.

Assume that we have an exponential boundary layer at $x = 1$, so the
boundary layer function is $y = \exp(-\gamma(1 - x)/\varepsilon)$ for some fixed $\gamma > 0$. The
idea of [Bak69] is to use an equidistant y-grid near $y = 1$ (which corresponds
to $x = 1$), then to map this grid back to the x-axis by means of the boundary
layer function. That is, gridpoints x_i near $x = 1$ are defined by

$$\exp\left(-\frac{\gamma(1 - x_i)}{\varepsilon}\right) = \frac{i}{N}.$$

This is equivalent to

$$x_i = 1 + \frac{\varepsilon}{\gamma}\ln\left(\frac{i}{N}\right).$$

Moving away from the layer, this definition of x_i will be modified to ensure
that $x_0 = 0$.

The Bakhvalov mesh generating function for (2.133) is

$$\lambda(t) = \begin{cases} \psi(t) := 1 + A\varepsilon\ln\left(1 - \dfrac{1 - t}{q}\right) & \text{for } t \in [1 - \tau, 1], \\ \pi(t) := \psi(1 - \tau) + (t - 1 + \tau)\psi'(1 - \tau) & \text{for } t \in [0, 1 - \tau]. \end{cases} \quad (2.135)$$

If $1 - \tau$ is a mesh point, then this mesh is coarse and equidistant on $[0, 1 - \tau]$,
and graded (i.e., $h_i \geq h_{i+1}$ for all i) on $[1 - \tau, 1]$ where it changes gradually
from coarse to fine. Following standard practice for layer-adapted meshes, we
now write H for the mesh diameter. The parameters $A > 0$ and $q \in (0, 1)$ are
user-chosen; the mesh grading is affected by A (which could for example be
taken equal to β) and the fraction of mesh points used to resolve the layer
is $1 - q$, up to a term that is exponentially small in ε. (We have made the
reasonable assumption that $\varepsilon \leq q/A$; if this is not true, then an equidistant
mesh should be used on $[0, 1]$.) The transition point $1 - \tau$ must satisfy

$$\psi(1 - \tau) + (-1 + \tau)\psi'(1 - \tau) = 0 \quad (2.136)$$

so that $\lambda(0) = 0$. Geometrically, (2.136) means that the point $(0, 0)$ lies on the
tangent π to $(t, \psi(t))$ at the point $(1 - \tau, \psi(1 - \tau))$. The construction ensures
that λ is not just a continuous function but lies in $C^1[0, 1]$.

If the boundary layer were at $x = 0$, one would redefine λ by means of
$\lambda(t) \mapsto 1 - \lambda(1 - t)$.

The Bakhvalov mesh can also be generated [Lin01b] using the monitor
function

$$M_{Ba}(x) = \max\left\{1, \frac{K\gamma}{\varepsilon}e^{-\gamma(1-x)/(\sigma\varepsilon)}\right\},$$
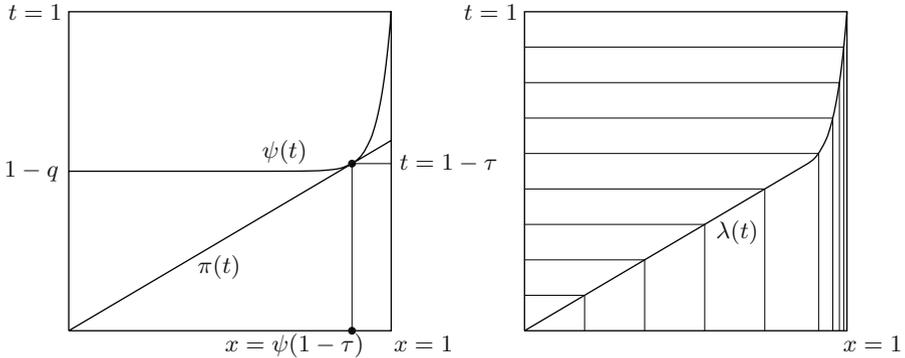
Fig. 2.5. Bakhvalov mesh generating function (left) and the mesh generated (right)

where $K > 0$ and $\sigma \geq 1$.

A drawback of Bakhvalov's original mesh is that the nonlinear equation (2.136) cannot be solved exactly for τ. The iteration

$$\tau^{(0)} = 0, \quad \tau^{(i+1)} = q - \frac{A\varepsilon(1 - \tau^{(i)})}{1 + A\varepsilon \ln(1 - \tau^{(i)}/q)} \text{ for } i = 0, 1, \ldots \quad (2.137)$$

is proved in [Bak69] to converge to τ, with $0 \leq \tau^{(i)} < \tau^{(i+1)} < \tau$ for all i. One can see that

$$q - C'\varepsilon \leq \tau \leq q - C'\varepsilon \quad (2.138)$$

for some constant C'. Here the left-hand inequality implies that $|\lambda'(t)| \leq C$ for all t and consequently $H \leq CN^{-1}$, while the right-hand inequality yields the approximation properties of the mesh. Bakhvalov meshes are not locally quasi-equidistant, uniformly in ε: for if $1 - \tau$ is the mesh point $x_m = \lambda(t_m)$ say, then

$$\frac{h_m}{h_{m+1}} = 1 - \frac{x_{m+1} - 2x_m + x_{m-1}}{x_{m+1} - x_m} \approx 1 - \frac{N^{-2}\psi''(t_m)}{N^{-1}\psi'(t_m)} = 1 + \mathcal{O}(\varepsilon^{-1}N^{-1}),$$

by (2.138).

Because the precise value of τ is not explicitly known, various authors have devised *Bakhvalov-type meshes* that approximate the original Bakhvalov mesh and are more easily computed.

For example, Kopteva [Kop99, Kop01b, Kop03] uses (2.135) but simply takes $\tau = q - A\varepsilon$, which is the value of the first iterate $\tau^{(1)}$ in (2.137) and satisfies (2.138). The $C^1[0, 1]$ property of the mesh generating function is lost, but this is not crucial: one still has a graded mesh with satisfactory approximation properties and $H \leq CN^{-1}$.

A related construction appears in [Bog84], where Boglaev considers a reaction-diffusion problem $-\varepsilon^2 u'' + c(x)u = f$ with $c(x) \geq \gamma^2 > 0$ and takes $\tau = (2/\gamma)\varepsilon|\ln \varepsilon|$ then constructs the mesh using a variant of (2.135). This

choice appears also in the A-mesh of Remarks III.2.12 and III.3.122. It ensures that the layer component of u is small on the coarse part of the mesh, but it introduces a factor $|\ln \varepsilon|$ into the final error estimate for convection-diffusion problems.

Vulanović [Vul83, Vul86, Vul89, Vul91] considers various singularly perturbed problems and uses mesh generating functions such as

$$\lambda(t) = \begin{cases} \psi(t) := 1 + A \varepsilon \dfrac{1-t}{q-(1-t)} & \text{for } t \in [1-\tau, 1], \\ \pi(t) := \psi(1-\tau) + (t-1+\tau)\psi'(1-\tau) & \text{for } t \in [0, 1-\tau]. \end{cases}$$

Then instead of (2.136) we get a quadratic equation in τ that can be solved. Convergence analyses are given in these papers.

In several papers (see [LP89] and its references) Liseikin examines the convergence of finite difference methods when using mesh generating functions $\lambda(t)$ of the given independent variable that satisfy $|\lambda'(t)| \leq C$ for all $t \in [0, 1]$. This approach generates a graded grid of Bakhvalov type. His book [Lis99] develops a general theory of grid generation. The analysis in these sources is written in terms of "layer-resolving transformations"; their relationship to mesh generating functions in a singular perturbation context is discussed in [Vul07].

A locally quasi-equidistant graded mesh is constructed in [Gar88] and a thorough analysis of the stability and convergence of various finite difference schemes on this mesh is given, but the number of meshpoints increases slowly as $\varepsilon \to 0$.

Remark 2.85. If simple upwinding applied to (2.133) is uniformly convergent in the sense of (2.134) for some constant $\alpha > 0$, and the mesh is locally quasi-equidistant (uniformly in ε), then the number N of mesh intervals must increase as $\varepsilon \to 0$. To see this, observe that the arguments of [Sty03] are still valid when slightly modified by considering a limit as $N \to \infty$ with $\varepsilon \geq h_1$ and $i = 1$; one then arrives at the conclusion of that paper that $h_1 = o(\varepsilon)$. (There are some minor extra mesh assumptions such as existence of $\lim_{N \to \infty} h_1/h_2$ and $\lim_{N \to \infty} h_2/h_1$.) But the mesh diameter is at least $1/N$, so the locally quasi-equidistant property implies that $\varepsilon K^N \geq 1/N$, where K is the constant of (2.130). Hence $N K^N \geq 1/\varepsilon$, so $N \approx \log_K(1/\varepsilon)$. ♣

We now consider the *analysis of difference schemes on Bakhvalov and Bakhvalov-type meshes*, using the example of simple upwinding. Thus our difference scheme is

$$L^N u_i^N := -\varepsilon \delta^2 u_i^N + b_i D_- u_i^N + c_i u_i^N = f_i \quad \text{for } i = 1, \dots, N-1,$$
$$u_0^N = u_N^N = 0, \tag{2.139}$$

where $z_i := z(x_i)$ for each function $z \in C[0,1]$ and the mesh is given by (2.135). Using Theorem 2.7, it is straightforward to check that the matrix L^N

associated with (2.139) is an M-matrix. To analyse the convergence of the scheme, recall the Shishkin decomposition $u = S + E$ of Lemma 1.9 and split the discrete solution $\{u_i^N\}$ in an analogous manner: define $\{S_i^N\}$ and $\{E_i^N\}$ by

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

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

$$|u_i - u_i^N| = |(S + E)_i - (S_i^N + E_i^N)| \le |S_i - S_i^N| + |E_i - E_i^N|. \qquad (2.140)$$

We shall bound each difference in (2.140) separately.

Lemma 2.86. *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

$$
\begin{aligned}
|L^N(S_i - S_i^N)| &= |L^N S_i - (LS)_i| \\
&\le 2\varepsilon \int_{x_{i-1}}^{x_{i+1}} |S'''(x)| \, dx + b_i \int_{x_{i-1}}^{x_i} |S''(x)| \, dx \\
&\le C(x_{i+1} - x_{i-1}) \\
&\le C N^{-1}
\end{aligned}
\qquad (2.141)
$$

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 (cf. Lemma 1.1) for $\{S_i - S_i^N\}$. Now

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

by (2.141), provided that C_0 is a sufficiently large constant. Clearly $w_0 = 0 = |S_0 - S_0^0|$ and $w_N = C_0 N^{-1} \ge 0 = |S_N - S_N^N|$. By a discrete comparison principle we get $|S_i - S_i^N| \le w_i \le C_0 N^{-1}$ for all i. $\square$

The difference $|E_i - E_i^N|$ must now be bounded for all i. Unsurprisingly, this is more difficult. We begin with a useful technical lemma. For $i = 0, \dots, N$, define the mesh function

$$Z_i = \prod_{j=1}^{i} \left(1 + \frac{\beta h_j}{2\varepsilon} \right)$$

(with the usual convention that if $i = 0$, then $Z_0 = 1$).

Lemma 2.87. *There exists a positive constant C such that*

$$L^N(Z_i) \geq \frac{C}{\max\{\varepsilon, h_i\}} \, Z_i \quad \text{for } i = 1, \ldots, N-1.$$

Proof. Clearly

$$D_+ Z_i = \frac{\beta}{2\varepsilon} Z_i \quad \text{and} \quad D_- Z_i = \frac{\beta}{2\varepsilon + \beta h_i} Z_i.$$

Hence

$$\varepsilon \delta^2 Z_i = -\frac{\varepsilon}{\hbar_i} \{D_+ Z_i - D_- Z_i\} = -\frac{\beta^2 h_i}{2\hbar_i(2\varepsilon + \beta h_i)} \, Z_i.$$

Thus

$$
\begin{aligned}
L^N(Z_i) &= \left[-\frac{\beta^2 h_i}{2\hbar_i(2\varepsilon + \beta h_i)} + \frac{b_i \beta}{2\varepsilon + \beta h_i} + c_i \right] Z_i \\
&\geq \frac{\beta(2b_i \hbar_i - \beta h_i)}{2\hbar_i(2\varepsilon + \beta h_i)} Z_i \\
&> \frac{\beta(b_i - \beta)}{2\varepsilon + \beta h_i} Z_i \\
&\geq \frac{C}{\max\{\varepsilon, h_i\}} Z_i,
\end{aligned}
$$

where we used $c_i \geq 0$, $h_i < 2\hbar_i$ and $b_i > \beta$. $\quad\square$

Discrete Green's functions will be used to prove that the operator L^N of (2.139) is $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$-stable, where $\|\cdot\|_{\infty,d}$ and $\|\cdot\|_{1,d}$ are the discrete analogues of the $L^\infty[0,1]$ and $L^1[0,1]$ norms defined by

$$\|v^N\|_{\infty,d} = \max_{j=1,\ldots,N-1} |v_j| \quad \text{and} \quad \|v^N\|_{1,d} = \sum_{j=1}^{N-1} \hbar_j |v_j|.$$

The *discrete Green's function* $G_{ij} = G(x_i, \xi_j)$ associated with the operator L^N of (2.139) and the mesh point $\xi_j \in \{x_0, x_1, \ldots, x_N\}$ is defined by

$$L^N G_{ij} = \delta_{ij}/\hbar_j \text{ for } i = 1, \ldots, N-1, \quad \text{with } G_{0j} = G_{Nj} = 0, \qquad (2.142)$$

where L^N is applied to the first variable in $G(\cdot, \cdot)$ and δ_{ij} is the Kronecker delta. Define $\mathcal{V}$ to be the space of grid functions v with $v_0 = v_N = 0$. For any $v \in \mathcal{V}$, one has

$$v_i = \sum_{j=1}^{N-1} \hbar_j G_{ij} [L^N v]_j \quad \text{for } i = 0, \ldots, N. \qquad (2.143)$$

We now study properties of the discrete Green's function that are analogues of those given for the continuous Green's function in Section 1.1.2.

Lemma 2.88. *For all i and j one has*

$$0 \le G_{ij} \le 2/\beta.$$

Proof. Fix $j \in \{0, \dots, N\}$. To obtain a discrete barrier function for G_{ij}, define the mesh function $w^{(j)}$ by

$$w_i^{(j)} = \begin{cases} (2/\beta) \prod_{k=i+1}^{j} (1 + \beta h_k/(2\varepsilon))^{-1}, & \text{for } i = 0, \dots, j-1, \\ 2/\beta, & \text{for } i = j, \dots, N. \end{cases} \tag{2.144}$$

The argument divides naturally into three cases.

When $0 < i < j$, one has

$$(L^N w^{(j)})_i = (2/\beta) \prod_{k=1}^{j} \left(1 + \frac{\beta h_k}{2\varepsilon}\right)^{-1} L^N(Z_i) > 0,$$

by Lemma 2.87.

When $i = j$, then $D_+ w_j^{(j)} = 0$ and

$$D_- w_j^{(j)} = \frac{2}{\beta h_j} \left(1 - \frac{1}{1 + \beta h_j/(2\varepsilon)}\right) = \frac{2}{2\varepsilon + \beta h_j},$$

so

$$(L^N w^{(j)})_j = \left(\frac{\varepsilon}{\hbar_j} + b_j\right) D_- w_j^{(j)} + c_j w_j^{(j)} \ge \frac{2(\varepsilon + b_j \hbar_j)}{\hbar_j (2\varepsilon + \beta h_j)} \ge \frac{1}{\hbar_j}.$$

Finally, when $j < i < N$, clearly $(L^N w^{(j)})_i = c_i w_i^{(j)} \ge 0$.

These calculations and (2.142) show that

$$L^N w_i^{(j)} \ge L^N G_{ij} \ge 0 \quad \text{for } i = 1, \dots, N-1,$$

and of course $w_0^{(j)} = G_{0j} = 0$, $w_N^{(j)} = G_{Nj} = 0$. As L^N is an M-matrix, from a discrete comparison principle and (2.144) it follows that

$$0 \le G_{ij} \le w_i^{(j)} \le 2/\beta,$$

as desired. □

This bound on the discrete Green's function is tantamount to the following $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ stability result.

Lemma 2.89. *Let $v \in \mathcal{V}$. Then*

$$\|v\|_{\infty,d} \le \frac{2}{\beta} \|L^N v\|_{1,d}.$$

Proof. This is immediate from (2.143) and Lemma 2.88. □

Lemma 2.89 is the discrete analogue of (1.19) for simple upwinding. It implies that

$$\|E - E^N\|_{\infty,d} \leq (2/\beta)\|L^N(E - E^N)\|_{1,d},$$

i.e., the pointwise error is bounded by the norm $\|\cdot\|_{1,d}$ of the consistency error. To finish the analysis one should bound $\|L^N(E - E^N)\|_{1,d}$ by CN^{-1} and deduce that $\|E - E^N\|_{\infty,d} \leq CN^{-1}$, but we do not give the details here as a related calculation is given in full for Shishkin meshes in Section 2.4.2; see also [AK96, AK98, AS95, KLS, LRV00, ST98]. Putting this bound together with (2.140) and Lemma 2.86, the final result obtained is

Theorem 2.90. *The simple upwind scheme (2.139) applied to (2.133) on a Bakhvalov or Bakhvalov-type mesh is first-order uniformly convergent with respect to the singular perturbation parameter:*

$$\|u - u^N\|_{\infty,d} \leq CN^{-1}$$

for some constant C.

In the convergence analysis of simple upwinding and other upwinded methods on Bakhvalov and Bakhvalov-type meshes, the $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ stability result of Lemma 2.89 is a radical departure from the numerical analysis of classical (i.e., non-singularly perturbed) two-point boundary value problems on equidistant meshes, where one typically uses $(\|\cdot\|_{\infty,d}, \|\cdot\|_{\infty,d})$ stability. This classical stability bound is evidently weaker, and is useless in the convection-diffusion context because the norm $\|\cdot\|_{\infty,d}$ of the truncation error is not bounded uniformly in ε.

When the convective term bu' is approximated by simple upwinding, the analysis is often facilitated if the standard approximation of $u''(x_i)$ given by $\delta^2 u_i$ in (2.132) is replaced by

$$u''(x_i) \approx \delta_-^2 u_i := \frac{1}{h_i}(D_+ u_i - D_- u_i).$$

After making this change in (2.139), one can simplify the proofs of Lemmas 2.87 and 2.88 by replacing the factor 2ε by ε in the functions Z_i and $w^{(j)}$ and changing $\hbar_j$ to h_j in the definition of G_{ij}. We then get $0 \leq G_{ij} \leq 1/\beta$ for all i and j.

Furthermore, the operator L^N now enjoys the following stability property, first shown in [AK98], which is even stronger than the $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ stability of Lemma 2.89.

Lemma 2.91. *Let $v \in \mathcal{V}$. Then*

$$\|v\|_{\infty,d} \leq \frac{2}{\beta}\|L^N v\|_{-1,\infty,d}, \tag{2.145}$$

where

$$\|v\|_{-1,\infty,d} := \max_{j=1,\ldots,N-1}\left|\sum_{k=1}^{j} h_k v_k\right|. \tag{2.146}$$

Proof. Sum by parts the identity (2.143), while observing that $\hbar_j$ has become h_j; this yields

$$|v_i| = \left| \sum_{j=1}^{N-1} [G_{ij} - G_{i,j+1}] \sum_{k=1}^{j} h_k [L^N v]_k \right|$$

$$\leq \left(\sum_{j=1}^{N-1} |G_{ij} - G_{i,j+1}| \right) \|L^N v\|_{-1,\infty,d}$$

for $i = 1, \ldots, N-1$. It can be shown [And01, Lin02a] that for each fixed i the function $j \mapsto G_{ij}$ is monotonically increasing for $j \leq i$ and monotonically decreasing for $j \geq i$. As we know already that $0 \leq G_{ii} \leq 1/\beta$, inequality (2.145) follows immediately. □

The norm $\|\cdot\|_{-1,\infty,d}$ is a discrete analogue of the Sobolev norm in $W^{-1,\infty}$ that was discussed in Section 1.1.2. Clearly $\|v\|_{-1,\infty,d} \leq \|v\|_{1,d}$ for all $v \in \mathcal{V}$.

The absence of absolute values inside the sum in (2.146) permits an error analysis that is simpler than our earlier analysis based on the $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ stability of Lemma 2.89. For, taking $b(x)$ to be constant for simplicity, (2.145) implies that

$$|u(x_i) - u_i^N| \leq \frac{2}{\beta} \max_{j=1,\ldots,N-1} \left| \sum_{k=1}^{j} h_k (L^N u - f)_k \right|$$

$$= \frac{2}{\beta} \max_{j=1,\ldots,N-1} \left| (-\varepsilon D_+ u_j + b u_j) - (-\varepsilon D_+ u_0 + b u_0) \right.$$

$$\left. + \sum_{k=1}^{j} h_k (c_k u_k^N - f_k) \right|.$$

But on integrating (2.133) from 0 to x_j we have

$$-\varepsilon [u'(x_j) - u'(0)] + b(u_j - u_0) + \int_0^{x_j} (cu - f)(x)\, dx = 0.$$

Thus

$$|u(x_i) - u_i^N| \leq \frac{2}{\beta} \max_{j=1,\ldots,N-1} \left| -\varepsilon [D_+ u_j - u'(x_j)] + \varepsilon [D_+ u_0 - u'(0)] \right.$$

$$\left. + \sum_{k=1}^{j} h_k (c_k u_k^N - f_k) - \int_0^{x_j} (cu - f)(x)\, dx \right|.$$

This estimate shows that we need consider only the consistency error in approximating the *first-order* derivative of u, unlike our previous analysis, which leads to the consistency error incurred in approximating u''. It follows that less regularity of u is required.

A Taylor expansion shows that

$$\left| \sum_{k=1}^{j} h_k(c_k u_k^N - f_k) - \int_0^{x_j} (cu - f)(x)\, dx \right| \leq \sum_{k=1}^{j} h_k \int_{x_{k-1}}^{x_k} |(cu - f)'(x)|\, dx$$

$$\leq \max_{k=1,\ldots,j} \int_{x_{k-1}}^{x_k} |(cu - f)'(x)|\, dx;$$

also, for each j one has

$$\varepsilon |D_+ u_j - u'(x_j)| \leq \varepsilon \int_{x_j}^{x_{j+1}} |u''(x)|\, dx \leq C \int_{x_j}^{x_{j+1}} [1 + |u'(x)|]\, dx,$$

where the second inequality is immediate from (2.133). Hence the error in the simple upwinding solution u^N satisfies

$$\|u - u^N\|_{\infty,d} \leq C \max_{k=1,\ldots,N} \int_{x_{k-1}}^{x_k} [1 + |u'(x)|]\, dx. \qquad (2.147)$$

The convergence result of Theorem 2.90 follows from this bound [AK98], and it is noteworthy that no decomposition of u from Section 1.1.3 is needed at any stage of the entire argument.

Unfortunately this approach appears to have restricted applicability: the $(\|\cdot\|_{\infty,d}, \|\cdot\|_{-1,\infty,d})$ stability property of Lemma 2.91 seems to be peculiar to problems posed in one dimension. On the other hand $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ stability does have a two-dimensional analogue, as we shall see in Section III.2.2.

Further results for Bakhvalov-type meshes will be given in Section 2.4.2, and they also appear in Section 2.5.

2.4.2 Piecewise Equidistant Meshes

The Bakhvalov mesh of Section 2.4.1 is intuitively a reasonable construction when dealing with problems whose solutions exhibit layer behaviour. Perhaps surprisingly, one can also prove uniform convergence results on an alternative class of special meshes whose construction is much simpler: *Shishkin meshes*. These are piecewise equidistant meshes that have been popular since the mid 1990s. The Shishkin mesh is discussed at length in [MOS96] and [FHM+00]; the former is concerned with the analysis of finite difference methods on this mesh for convection-diffusion and reaction-diffusion problems, while the main thrust of the latter is the presentation of detailed numerical results on Shishkin meshes for a variety of problems in one and two dimensions, together with some theoretical results.

We now describe the Shishkin mesh for the convection-diffusion problem (2.133). Set $\sigma = \min\{1/2,\ (2/\beta)\varepsilon \ln N\}$. In fact we shall assume that $\sigma = (2/\beta)\varepsilon \ln N$ as the case $\sigma = 1/2$ 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. Subdivide each of $[0, 1-\sigma]$ and $[1 - \sigma, 1]$ by an equidistant mesh with $N/2$ subintervals; see Figure 2.6.
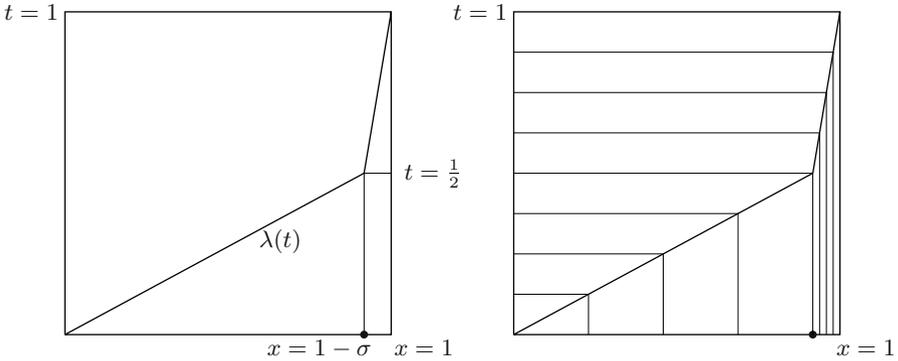
Fig. 2.6. Shishkin mesh generating function (left) and the mesh generated (right)

Remark 2.92. It is not vital that one has exactly the same number of subintervals in $[0, 1-\sigma]$ and $[1-\sigma, 1]$. All that the theory demands is that as $N \to \infty$ the number of subintervals in each of these two intervals is bounded below by CN for some constant $C > 0$.   ♣

The coarse part of this Shishkin mesh has spacing $H = 2(1-\sigma)/N$, so $N^{-1} \le H \le 2N^{-1}$. The fine part has spacing $h = 2\sigma/N = (4/\beta)\varepsilon N^{-1} \ln N$, so $h \ll \varepsilon$. Thus there is a very abrupt change in mesh size as one passes from the coarse part to the fine part. The mesh is not locally quasi-equidistant, uniformly in ε.

On the mesh, $x_i = iH$ for $i = 0, \ldots, N/2$ and $x_i = 1 - (N-i)h$ for $i = N/2+1, \ldots, N$.

Remark 2.93. (A key property of the Shishkin mesh) Nonequidistant meshes for convection-diffusion problems are sometimes described as "layer-resolving" meshes. One might infer from this terminology that wherever the derivatives of u are large, the mesh is chosen so fine that the truncation error of the difference scheme is controlled. But the Shishkin mesh does not fully resolve the layer: for

$$|u'(x)| \approx C\varepsilon^{-1} \exp(-b(1)(1-x)/\varepsilon),$$

so

$$|u'(1-\sigma)| \approx C\varepsilon^{-1} \exp(-2\ln N) = C\varepsilon^{-1}N^{-2},$$

which in general is large since typically $\varepsilon \ll N^{-1}$. That is, $|u'(x)|$ is still large on part of the first coarse-mesh interval $[x_{N/2-1}, x_{N/2}]$.

At first sight this incomplete resolution of the boundary layer seems like a flaw, but it is in fact the key property of the mesh! Shishkin's insight was that one could achieve satisfactory theoretical and numerical results without resolving all of the layer and as a consequence his mesh permits us to use a fixed number of mesh points that is independent of ε. If one set out to "repair" the Shishkin mesh by constructing 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 the derivatives of u are very large, then the number of mesh points required would have to grow like $|\ln \varepsilon|$ as $\varepsilon \to 0$. See [Wes96] and [FHM$^+$00, Section 3.6].

Although the number of mesh points is fixed independently of ε, nevertheless numerical analysis on Shishkin meshes does pay a price for the nature of the construction: typically the trickiest part of the domain to handle is the first coarse mesh interval – because the derivatives of u are large there.   ♣

Consider now the numerical solution of (2.133) by the simple upwind scheme (2.139) on the Shishkin mesh described above. As in the analysis of Section 2.4.1, split the computed solution into its smooth and layer components, viz., $u_i^N = S_i^N + E_i^N$ for all i, and we have again (2.140):

$$|u_i - u_i^N| = |(S + E)_i - (S_i^N + E_i^N)| \leq |S_i - S_i^N| + |E_i - E_i^N|.$$

It is easy to see that Lemma 2.86 is still valid, so it remains to bound $|E_i - E_i^N|$.

For this estimate, the approach is of necessity less direct than for $|S_i - S_i^N|$ because $E(x)$ has large derivatives on part of the coarse mesh (see Remark 2.93), resulting in an intractably large truncation error there. Thus 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 an error analysis like that of Lemma 2.86 will work.

From Lemma 1.9,

$$|E_i| \leq C e^{-\beta[1 - (1 - \sigma)]/\varepsilon} = C N^{-2} \leq C N^{-1} \quad \text{for } i = 0, \ldots, N/2. \qquad (2.148)$$

In the next Lemma a discrete barrier function is used to show that, like $|E_i|$, the mesh function $|E_i^N|$ is small when $i \leq N/2$.

Lemma 2.94. *There exists a constant C such that*

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

Proof. Recall the function Z_i of Section 2.4.1. Now $e^t \geq 1 + t$ for all $t \geq 0$, so

$$\frac{Z_i}{Z_N} = \prod_{j=i+1}^{N} \left(1 + \frac{\beta h_j}{2\varepsilon}\right)^{-1} \geq \prod_{j=i+1}^{N} e^{-\beta h_j/(2\varepsilon)} = e^{-\beta(1 - x_i)/(2\varepsilon)}. \qquad (2.149)$$

Set $Y_i = C_2 Z_i/Z_N$ for $i = 0, \ldots, N$. Then $L^N Y_i = (C_2/Z_N)L^N Z_i \geq 0 = |L^N E_i^N|$ for $i = 1, \ldots, N - 1$, by Lemma 2.87 and the definition of $\{E_i^N\}$. Also $Y_N = C_2 \geq |E(1)| = |E_N^N|$ if the constant C_2 is chosen sufficiently large, by the bound on $|E(x)|$ in Lemma 1.9. Finally, (2.149) and Lemma 1.9 imply that

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

provided that the constant C_2 is chosen sufficiently large. Thus we can choose C_2 so that $\{Y_i\}$ is a discrete barrier function for $\{E_i^N\}$, and it follows (cf. Lemma 1.1) that

$$|E_i^N| \leq Y_i = \frac{C_2 Z_i}{Z_N} \quad \text{for all } i. \tag{2.150}$$

For $i = 0, \ldots, N/2$,

$$\frac{Z_i}{Z_N} \leq \frac{Z_{N/2}}{Z_N} = \prod_{j=1+N/2}^{N} \left(1 + \frac{\beta h}{2\varepsilon}\right)^{-1} = \left(1 + 2N^{-1}\ln N\right)^{-N/2}.$$

But $\ln(1+t) \geq t - t^2/2$ for $t \geq 0$, so

$$-\frac{N}{2}\ln\left(1 + 2N^{-1}\ln N\right) \leq -\frac{N}{2}\left[2N^{-1}\ln N - (2N^{-1}\ln N)^2/2\right]$$
$$= -\ln N + N^{-1}\ln^2 N.$$

Taking exponentials, one gets

$$\frac{Z_i}{Z_N} \leq N^{-1}e^{(\ln^2 N)/N} \leq CN^{-1}$$

for some constant C. Combining this inequality with (2.150), the proof is complete.   □

Corollary 2.95. *There exists a constant C such that*

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

Proof. This is immediate from (2.148) and Lemma 2.94.   □

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

Lemma 2.96. *There exists a constant C such that*

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

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 Lemma 1.9, a standard consistency error analysis shows that for $i = N/2 + 1, \ldots, N - 1$,

$$|L^N(E_i - E_i^N)| = |L^N E_i - (LE)_i|$$

$$\leq 2\varepsilon \int_{x_{i-1}}^{x_{i+1}} |E'''(x)|\, dx + b_i \int_{x_{i-1}}^{x_i} |E''(x)|\, dx$$

$$\leq C \int_{x_{i-1}}^{x_{i+1}} \varepsilon^{-2} e^{-\beta(1-x)/\varepsilon}\, dx$$

$$= C\varepsilon^{-1} e^{-\beta(1-x_i)/\varepsilon} \sinh(\beta h/\varepsilon)$$

$$\leq C\varepsilon^{-1} N^{-1}(\ln N) e^{-\beta(1-x_i)/\varepsilon},$$

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

Set $\phi_i = C_3 N^{-1}(\ln N)(1 + Z_i/Z_N)$ for $i = N/2, \ldots, N$, where the constant C_3 will be chosen later. By Lemma 2.87 and (2.149),

$$L^N \phi_i \geq C_3 N^{-1}(\ln N)(L^N Z_i)/Z_N$$

$$\geq C_3 C_1 \varepsilon^{-1} N^{-1}(\ln N) Z_i/Z_N$$

$$\geq C_3 C_1 \varepsilon^{-1} N^{-1}(\ln N) e^{-\beta(1-x_i)/(2\varepsilon)}$$

for $i = N/2 + 1, \ldots, N$. 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

$$\phi_{N/2} = C_3 N^{-1}(\ln N)(1 + Z_{N/2}/Z_N) \geq C_3 N^{-1}(\ln N) \geq |E_{N/2} - E_{N/2}^N|$$

by Corollary 2.95, and $\phi_N = 2C_3 N^{-1}(\ln N) > 0 = |E_N - E_N^N|$.

Thus $\{\phi_i\}$ is a discrete barrier function for $\{E_i - E_i^N\}$, and it follows (cf. Lemma 1.1) that for $i = N/2, \ldots, N$, we have $|E_i - E_i^N| \leq \phi_i \leq 2C_3 N^{-1}\ln N$. $\square$

The final convergence result for simple upwinding on a Shishkin mesh can now be stated.

Theorem 2.97. *Let u be the solution of the convection-diffusion problem (2.133). There exists a constant C such that the solution $\{u_i^N\}$ of (2.139) satisfies*

$$\|u - u^N\|_{\infty,d} \leq CN^{-1}\ln N.$$

Proof. Combine (2.140), Lemma 2.86, Corollary 2.95 and Lemma 2.96. $\square$

Remark 2.98. (Alternative proof of Theorem 2.97) The argument above is typical of many papers dealing with Shishkin meshes, but alternatively Theorem 2.97 could have been proved in the $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ framework of Section 2.4.1, as we now outline. Again use the splitting $u_i^N = S_i^N + E_i^N$. The analysis of $|S_i - S_i^N|$ is straightforward. The proof of Lemma 2.96 demonstrates that

$$|L^N(E_i - E_i^N)| \leq C\varepsilon^{-2} \hbar_i e^{-\beta(1-x_{i+1})/\varepsilon}$$

for each i. This bound turns out to be unsuitable when $i = N/2 - 1$ and $i = N/2$, where one uses the simpler bound

$$|L^N(E_i - E_i^N)| = |L^N E_i| \leq C\hbar_i^{-1} e^{-\beta(1-x_{i+1})/\varepsilon},$$

which follows from the definition of L^N and the bounds on $|E(x)|$ and $|E'(x)|$ given in Lemma 1.9. Now $(\|\cdot\|_{\infty,d}, \|\cdot\|_{1,d})$ stability implies that

$$\|E - E^N\|_{\infty,d} \leq C\Bigg[\sum_{i<N/2-1} \varepsilon^{-2} H^2 e^{(-\sigma-H)/\varepsilon} + e^{-\sigma/\varepsilon} + e^{(-\sigma+h)/\varepsilon}$$

$$+ \sum_{i>N/2} \varepsilon^{-2} h^2 e^{-\beta(1-x_{i+1})/\varepsilon} \Bigg]$$

$$\leq CN^{-1} \ln N,$$

as $\varepsilon^{-2} H^2 e^{-H/\varepsilon} \leq C$, $e^{(-\sigma+h)/\varepsilon} \leq Ce^{-\sigma/\varepsilon} \leq CN^{-2}$ and the geometric series $\sum_{i>N/2} e^{-\beta(1-x_{i+1})/\varepsilon}$ is bounded by $CN/(\ln N)$. ♣

The condition number of the discrete linear system associated with the scheme (2.139) on a Shishkin mesh is $\mathcal{O}(\varepsilon^{-2} N^2 \ln^{-2} N)$, which is bad when ε is small, but [Roo96] an easy preconditioning by diagonal scaling (approximate equilibration) reduces this condition number to $\mathcal{O}(N^2 \ln^{-1} N)$. An extensive discussion of the iterative solution of the linear systems generated when convection-diffusion problems are discretized is given in [ESW05, Chapter 4].

Remark 2.99. (Choice of transition point) 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 2.97 reveals that σ should have the form $(k/\beta)\varepsilon\phi(N)$, 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 $\ln N$. A statement of necessary conditions on σ is given in [FHM+00, Section 3.6] (the detailed analysis appears in [Shi92b, pp.207–8]); see also [Wes96]. The earliest appearance of this transition point is in a remarkable paper by van Veldhuizen [vV78], who chooses $\sigma = C\varepsilon \ln N$ with a logarithmically graded mesh in the layer region and an equidistant mesh on the rest of the interval — this is a member of the class of Bakhvalov-Shishkin meshes that are discussed later in this section. See also [Seg82], where Segal proposes a piecewise equidistant mesh resembling Shishkin's but with $\sigma = C\varepsilon$.

The choice $k = 2$ used in our definition of σ subtly enters the proof of Lemma 2.94 during the final chain of inequalities that bound Z_i/Z_N. How to choose k in an optimal way is discussed in [ST98] and, using an argument resembling the proof above of Theorem 2.97, it is shown that for a variant of simple upwinding one has

$$\|u - u^N\|_{\infty,d} \leq C \max\{N^{-k}, kN^{-1} \ln N\} \quad \text{for } i = 0, \dots, N. \qquad (2.151)$$

The sharpness of this bound is confirmed by numerical experiments. Consequently choosing k larger than 1 diminishes the numerical accuracy of the method but does not affect the numerical rate of convergence, while choosing k smaller than 1 causes a noticeable deterioration in this rate.

Alternatively, one can check [Lin01a] that the argument leading to (2.147) remains valid on a Shishkin mesh and reproduces (2.151). ♣

The result of Theorem 2.97 can be extended to more general forms of up-winding and to other layer-adapted meshes that are designed for convection-diffusion problems. For a clear and comprehensive survey of such generalizations for problems in one and two dimensions, see [Lin03a]; we shall present some of this material in the remainder of this section.

In [Kop96] a second-order upwind 4-point scheme is examined on a Shishkin mesh and the bound $\|u - u^N\|_{\infty,d} \leq CN^{-2}\ln^2 N$ is proved. This result is extended to more general meshes in [Kop01b].

Remark 2.100. (Robin boundary conditions) Consider again the convection-diffusion problem $-\varepsilon u'' + b(x)u' + c(x)u = f(x)$ of (2.133) but with the more general Robin boundary conditions

$$\gamma_1 u(0) - \gamma_2 u'(0) = A, \quad \beta_1 u(1) + \beta_2 \varepsilon u'(1) = B,$$

where $\gamma_1, \gamma_2, \beta_1, \beta_2, A$ and B are given constants.

Simple upwinding on a Shishkin mesh is applied in [FHM+00], where the case of Neumann boundary conditions ($\gamma_1 = \beta_1 = 0$, $\gamma_2 = \beta_2 = 1$) is considered. This result is generalized in [AH03] to the case of full Robin boundary conditions (with $c(\cdot) \equiv 0$): $\gamma_1 > 0$, $\gamma_2 \geq 0$, $\beta_1 \geq 0$, $\beta_2 \geq 0$ and $\beta_1 + \beta_2 > 0$. The solution u is shown to have properties similar to those given in Lemma 1.9 and a bound like that of Theorem 2.97 is obtained.

The case $\gamma_2 = 0$ is considered in [AS96], where it is shown that a modification of Samarskiĭ's monotone scheme (see Section 2.1.2) on a Shishkin mesh yields $\|u - u^N\|_{\infty,d} \leq CN^{-2}\ln^2 N$. ♣

Remark 2.101. (Conservation form) In [AK98] a convection-diffusion problem in conservation form is examined:

$$-\varepsilon\big(p(x)u'(x)\big)' - \big(r(x)u(x)\big)' = f(x) \text{ for } 0 < x < 1, \quad u(0) = u_0, \ u(1) = u_1,$$

where $p(\cdot) \geq p_0 > 0$, $r(\cdot) \geq r_0 > 0$. The solution $u(x)$ has in general an exponential boundary layer at $x = 0$. Let the user-chosen parameter σ_i lie in $[1/2, 1]$ for $i = 1, \ldots, N - 1$. On an arbitrary mesh, consider the family of difference schemes

$$-D_\sigma\Big(\varepsilon\bar{p}_i D_- u_i^N + \sigma_i r_i u_i^N + (1 - \sigma_i)r_{i-1}u_{i-1}^N\Big) = \tilde{f}_i \text{ for } i = 1, \ldots, N - 1,$$

with $u_0^N = u_0$, $u_N^N = u_1$. Here

$$D_\sigma v_i := \frac{v_{i+1} - v_i}{h_i(1 - \sigma) + h_{i+1}\sigma_{i+1}} \quad \text{for all mesh functions } v_i$$

and $\bar{p}_i := p(x_{i-1/2})$, $r_i := r(x_i)$, $\tilde{f}_i := [f(x_i - h_i(1-\sigma_i)) + f(x_i + h_{i+1}\sigma_{i+1})]/2$. If $\sigma_i = 1/2$ for all i then this is a central difference scheme, while $\sigma_i \equiv 1$ delivers a form of simple upwinding.

Choose the weight $\sigma_i \in [1/2, 1]$ to satisfy

$$1 - \frac{\varepsilon \bar{p}_i}{h_i r_{i-1}} \leq \sigma_i \leq \frac{1}{2} + \frac{C_0 h_i}{\varepsilon} \, ,$$

where C_0 is some arbitrary but fixed constant. Here the left-hand inequality ensures that the matrix associated with the scheme is an M-matrix and the right-hand inequality guarantees formal second-order consistency on smooth grids. Then it is shown that

$$\|u - u^N\|_{\infty,d} \leq C \left\{ \max_i h_i^2 + \max_i \left[\min_i \{1, (h_i/\varepsilon)^2\} \exp(-\gamma x_{i-1}/\varepsilon) \right] \right\}, \tag{2.152}$$

where $\gamma < r(0)/p(0)$ is an arbitrary constant.

This revealing bound gives precise information on how small h_i has to be relative to $\exp(-\gamma x_{i-1}/\varepsilon)$ inside the layer (i.e., where $\exp(-\gamma x_{i-1}/\varepsilon)$ has not yet decayed) if one is to get second-order convergence. That is, (2.152) quantifies the crucial property that a good layer-adapted mesh should have. It follows [AK98] from (2.152) that $\|u - u^N\|_{\infty,d} \leq CN^{-2}$ on a Bakhvalov mesh and $\|u - u^N\|_{\infty,d} \leq CN^{-2} \ln^2 N$ on a Shishkin mesh. A related analysis in [Kop01b] derives similar results for a 4-point discretization of this problem.

The analogue of (2.152) for first-order convergence is

$$\|u - u^N\|_{\infty,d} \leq C \left\{ \max_i h_i + \max_i \left[\min_i \{1, h_i/\varepsilon\} \exp(-\gamma x_{i-1}/\varepsilon) \right] \right\},$$

which is proved in [AK98, Theorem 3].   ♣

Remark 2.102. Error estimates in various norms for numerical methods on Shishkin meshes usually include a multiplicative factor $\ln^\gamma N$ for some $\gamma > 0$. This factor is unimportant relative to the main convergence factor N^{-k}, where $k > 0$. On Bakhvalov meshes the $\ln N$ factor disappears, so these meshes yield a higher rate of convergence, but they are more complicated to construct.

Table 2.1 shows the rates of convergence $N^{-\alpha}$ observed when the actual rate is $N^{-2} \ln^2 N$.

Table 2.1. Observed convergence rates $N^{-\alpha}$ with actual rate of $N^{-2} \ln^2 N$

N	32	64	128	256	512	1024	2048	4096
α	1.3561	1.4739	1.5552	1.6147	1.6601	1.6960	1.7250	1.7489

In [Xen03] the optimality of the $\ln^\gamma N$ factor in error estimates on Shishkin meshes is discussed in the context of reaction-diffusion problems. See also [Shi08] where n-widths are used to address this question.   ♣

Several authors have studied variants of the original Shishkin mesh that use a transition point $\sigma = (k/\beta)\varepsilon \ln N$ for some positive constant k, are equidistant (or at least quasi-equidistant) on $[0, 1-\sigma]$ and are graded in some way on

$[1 - \sigma, 1]$. We shall refer to these as *Shishkin-type meshes*. Thus Shishkin-type meshes have a mesh generating function

$$
\lambda(t) = \begin{cases} 2t(1 - \sigma) & \text{for } t \in [0, 1/2], \\ 1 - \dfrac{k\varepsilon}{\beta}\tilde{\lambda}(1 - t) & \text{for } t \in [1/2, 1], \end{cases}
$$

where $\tilde{\lambda} : [0, 1/2] \to [0, \ln N]$ is strictly increasing. (Once again we have placed half the mesh points in each of the intervals $[0, 1 - \sigma]$ and $[1 - \sigma, 1]$, but this can be varied as in Remark 2.92.)

Now, following [RL99], define the *mesh-characterizing function*

$$
\psi = \exp(-\tilde{\lambda}) : [0, 1/2] \to [1, 1/N].
$$

This function is strictly decreasing. In the case of a standard Shishkin mesh one has $\psi(t) = \exp(-2t \ln N)$. Then an analysis like that of Theorem 2.97 yields the following result:

Theorem 2.103. *Consider the solution $\{u_i^N\}$ of (2.133) using the simple upwind scheme on a Shishkin-type mesh. Assume that $k \geq 2$, that $\tilde{\lambda}$ is piecewise differentiable, and that*

$$
\max_{t \in [0, 1/2]} \tilde{\lambda}'(t) \leq C_0 N \quad \text{and} \quad \int_0^{1/2} \left[\tilde{\lambda}'(t)\right]^2 dt \leq C_0 N
$$

for some constant C_0. Then there exists a constant C such that

$$
|u_i - u_i^N| \leq \begin{cases} C(H + N^{-1}) & \text{for } i = 0, \ldots, N/2 - 1, \\ C(H + N^{-1} \max_{t \in [0, 1/2]} |\psi'(t)|) & \text{for } i = N/2, \ldots, N. \end{cases}
$$

This result is more general than Theorem 2.97, since for the standard Shishkin mesh one has $H \leq 2N^{-1}$ and $\max_{t \in [0, 1/2]} |\psi'(t)| = 2 \ln N$. For example, it encompasses Bakhvalov-Shishkin meshes [Lin99, Lin00a], where the mesh on $[1 - \sigma, 1]$ is graded as for a standard Bakhvalov mesh; with these meshes $\psi(t) = 1 - 2t(1 - N^{-1})$ and $\max_{t \in [0, 1/2]} |\psi'(t)| = 2(1 - N^{-1}) \leq 2$, so Theorem 2.103 yields $\|u - u^N\|_{\infty,d} \leq C(H + N^{-1}) \leq C(\varepsilon + N^{-1})$, since in general on these meshes one has $H = \mathcal{O}(\varepsilon)$.

Simple upwinding yields at best first-order convergence, irrespective of the mesh used. Several papers have constructed other difference schemes on Shishkin-type meshes that achieve a higher order of convergence on some or all of the domain. Before describing these, we caution the reader that the construction of stable difference schemes attaining higher-order convergence is much easier for ordinary differential equations than for partial differential equations.

For central differencing on a Shishkin mesh, it is shown in [AK96] that the computed solution $\{u_i^N\}$ satisfies

$$
\|u - u^N\|_{\infty,d} \leq C N^{-2} \ln^2 N. \tag{2.153}
$$

The proof, which uses discrete Green's functions, is difficult as the scheme does not satisfy a discrete maximum principle. A generalization of this result is applied to Bakhvalov meshes in [Kop99].

In Figure 2.7 we show the computed solution for a typical problem using central differencing on a Shishkin mesh, with the computed nodal values joined by straight line segments. The left-hand diagram shows the solution on the coarse mesh; in the right-hand diagram the x-axis is rescaled to show the computed solution on the fine mesh.
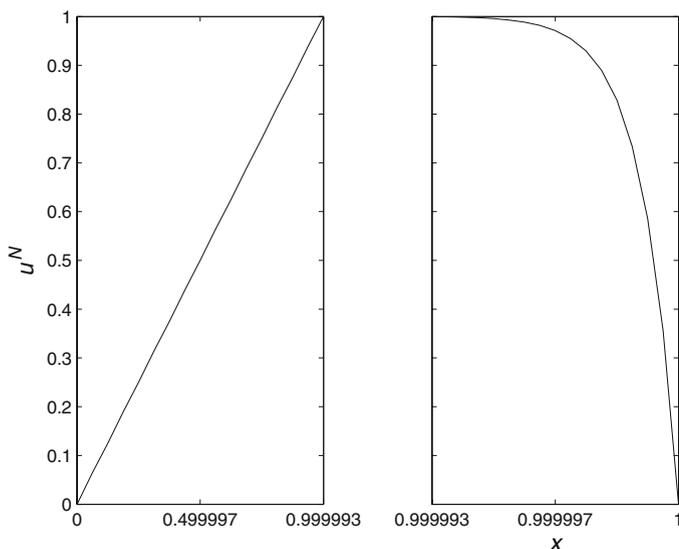
Fig. 2.7. Solution computed by central differencing on a Shishkin mesh

Although the graph of Figure 2.7 seems satisfactory, nevertheless this computed solution has small oscillations on the coarse mesh. This unwelcome fact is implied by Figure 2.8, which displays the error in the computed solution of Figure 2.7, with once again the fine mesh rescaled in order to show the data more clearly.

Despite the oscillations in the computed solution, if one selects alternate points in this solution (i.e., those corresponding to $\{x_{2i}\}$, or those corresponding to $\{x_{2i+1}\}$) then one obtains a non-oscillatory solution. This idea dates back to [AC85] but was first analysed in [Len00a], where the bound (2.153) is proved. In a later paper [Len00b] the technique is applied to a piecewise-quadratic Galerkin finite element method and it is shown that the computed solution satisfies $\|u - u^N\|_{\infty,d} \leq C N^{-4} \ln^4 N$.
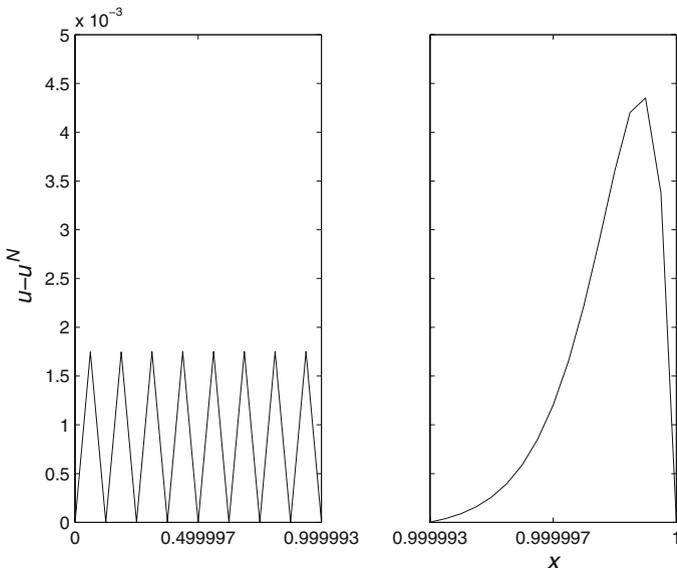
Fig. 2.8. Error in solution computed by central differencing on a Shishkin mesh

Another unorthodox way of extracting useful information from the oscillatory central difference solution is described and analysed in [SYZ07]. Using central differencing, solve (2.133) on an equidistant mesh $\{x_i\}_0^N$, then add an extra grid point in the mesh interval $(x_{N-1}, 1)$ and solve the problem again; for each of these two (oscillatory) computed solutions join the nodal values by straight lines (i.e., construct the piecewise linear finite element solution); finally, find the points $(\hat{x}_i, \hat{y}_i)$ where these two piecewise linear solutions intersect. Then one has $x_{i-1} \leq \hat{x}_i \leq x_i$ for all i and $\max_{0 \leq i \leq N-1} |u(\hat{x}_i) - \hat{y}_i| \leq CN^{-2}$.

Numerical experience with central differencing on Shishkin meshes for two-dimensional problems [LS01b] reveals that despite the theoretical accuracy of the method it is computationally expensive to solve the discrete linear system, so we shall not pursue this approach.

The central difference approximation of u' can be used where the mesh is fine without destroying the M-matrix property, while on the coarse mesh a midpoint upwind approximation also has the correct sign pattern and is formally second-order consistent. These observations are the basis for the methods of [SR97, ST98], where (under the simplifying assumption $c(\cdot) \equiv 0$) it is proved that

$$|u_i - u_i^N| \leq \begin{cases} CN^{-1}(\varepsilon + N^{-1}) & \text{for } i = 0, \ldots, N/2 - 1, \\ CN^{-2} \ln^2 N & \text{for } i = N/2, \ldots, N. \end{cases}$$

In Remark 2.101 we described an inverse-monotone scheme that, like central differencing, yields (2.153) on a Shishkin mesh and $\|u - u^N\|_{\infty,d} \leq CN^{-2}$ on a Bakhvalov mesh. A related inverse-monotone scheme in [Lin01a] achieves the same orders of convergence on these meshes; see also [Lin01b] for Shishkin-type meshes. The bound (2.153) is derived in [AS95] for a modified version of Samarskiĭ's monotone scheme on a Shishkin mesh.

The HODIE technique of Section 2.1.4 is used in [CG04, CGL99] to generate two schemes on Shishkin meshes for which one obtains $\|u - u^N\|_{\infty,d} \leq C(N^{-1} \ln N)^k$ for $k = 2, 3$ respectively. It is not clear if this approach can be extended to elliptic convection-diffusion problems in two dimensions.

High-order pointwise convergence results can be deduced from the hp finite element methods of Melenk and Schwab [MS99a, MS99b], but the construction of such methods in a finite difference framework remains an open problem.

Remark 2.104. On Shishkin-type meshes with transition point $1 - \sigma$, where $\sigma = (k/\beta)\varepsilon \ln N$, the analyses in papers such as [ST98] disclose a relationship between the user-chosen constant k and the order of convergence: if the method is expected to produce convergence of order $(N^{-1} \ln N)^m$, then one should choose $k \geq m$. ♣

Instead of seeking a difference scheme that yields a solution achieving a high order of accuracy, one can take an easily-implemented low-order scheme such as simple upwinding and apply some postprocessing technique to the computed solution to improve its accuracy. Here we discuss two such techniques: *Richardson extrapolation* and *defect correction*.

Richardson extrapolation is applied in [NS03] to the solution obtained from simple winding on a Shishkin mesh. Two solutions are computed initially: $\{v_i^N\}$ on a standard Shishkin mesh $\mathcal{S}$ and $\{\tilde{v}_i^{2N}\}$ on a mesh obtained by bisecting each subinterval in $\mathcal{S}$. Thus $\tilde{v}_{2i}^{2N}$ lies at the same mesh point as v_i^N for $i = 0, \ldots, N$. Apply the extrapolation formula

$$u_i^N := 2\tilde{v}_{2i}^{2N} - v_i^N \text{ for } i = 0, \ldots, N.$$

It is shown in [NS03] that the extrapolated solution $\{u_i^N\}$ satisfies the error bound (2.153). An analogue of this result is valid for problems in two dimensions [Kop03], but the analysis is more delicate in this setting. Iterated extrapolation – to improve the order of accuracy still further – is described but not analysed in [NS03].

Defect correction has been applied to many non-singularly perturbed problems [BR84]. Its philosophy is to generate a stable higher-order scheme by combining a stable low-order scheme such as simple upwinding with a higher-order but possibly unstable scheme such as central differencing on the same mesh. Thus, on a Shishkin mesh let $\hat{u}^N$ be the simple upwind solution, i.e., $L_{up}^N \hat{u}^N = f^N$, where L_{up} is the discrete operator corresponding to simple upwinding. Compute the "defect" $\sigma^N := f^N - L_c^N \hat{u}^N$ where L_c^N is the central difference operator. Then compute the "defect correction" δ^N by solving $L_{up}^N \delta^N = \sigma^N$. Finally, set $u^N := \hat{u}^N + \delta^N$.

This method avoids computational difficulties by solving only discrete systems that involve the stable upwind operator L_{up}^N, yet it aims to attain the higher-order convergence associated with the operator L_c^N. For convection-diffusion problems, [FLR01] gives an analysis showing that the solution u^N does indeed satisfy (2.153). Auspicious numerical results for defect correction applied to a convection-diffusion problem in two dimensions appear in [LS01a].

Although defect correction is superficially different from Richardson extrapolation, the two are nevertheless related, and Linß [Lin04] furnishes a unified error analysis that is applicable to both methods.

Remark 2.105. (Derivative approximation) Sometimes when solving (2.133), one wishes to compute the derivative u'. We know that $|u'(1)| = \mathcal{O}(\varepsilon^{-1})$, so at the boundary $x = 1$ it is more reasonable to estimate the computed approximation of $\varepsilon u'(x)$ than the approximation of $u'(x)$.

Consider the problem $-\varepsilon(p(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x)$, with $p(\cdot) > 0$, $b(\cdot) > 0$, $c(\cdot) \geq 0$ and Dirichlet boundary conditions. Its solution u has properties similar to the solution of (2.133). Let u^N be the approximate solution obtained using a modified Samarskiĭ scheme on a Shishkin mesh. Then [AS96] a weighted divided difference Γu_N computed from u^N satisfies $|-\varepsilon p(1)u'(1) - \Gamma u_N| \leq CN^{-2}\ln^2 N$.

When simple upwinding is applied on a Shishkin mesh to solve (2.133) for either Dirichlet or Neumann boundary conditions at $x = 1$, the computed solution u^N is proved in [FHM$^+$00, Theorem 3.17] to satisfy

$$\max_i |\varepsilon(D_- u_i^N - u'(x_i))| \leq CN^{-1}\ln N. \tag{2.154}$$

A bound like (2.154) is reasonable near $x = 1$, but away from this boundary the ε weighting is too strong. For $i = 0, \ldots, N/2 - 1$ the bound can be sharpened [KS01b] to $|D_- u_i^N - u'(x_i)| \leq CN^{-1}\ln N$. The analysis in this paper is also applicable to other layer-adapted meshes.

For the central difference scheme on Shishkin-type meshes, one has [RL01a]

$$\max_i |\varepsilon(D_- u_i^N - u'(x_{i-1/2}))| \leq C(H + N^{-1}\max|\psi'|)^2,$$

where H is the mesh diameter and ψ the mesh-characterizing function.   ♣

Remark 2.106. (Reaction-diffusion) Consider the reaction-diffusion problem of Remark 1.10:

$$-\varepsilon^2 u'' + c(x)u = f(x) \text{ for } 0 < x < 1, \quad u(0) = u(1) = 0, \tag{2.155}$$

with $c(\cdot) \geq \gamma > 0$. The solution $u(x)$ typically has layers at $x = 0$ and $x = 1$. To construct a Shishkin mesh for the this problem, let N be a positive integer that is divisible by 3. Set $\tau = (2/\gamma)\varepsilon \ln N$. Then the Shishkin meshpoints divide each of the intervals $[0, \tau]$, $[\tau, 1-\tau]$ and $[1-\tau, 1]$ into $N/3$ equal subintervals. (As in Remark 2.92, it's not necessary to distribute the meshpoints

in exactly this way: it suffices to partition each of the above three intervals by a number of equal subintervals that is bounded below by a fixed fraction of N.) Now apply the difference scheme

$$L^N u_i^N := -\varepsilon \delta^2 u_i^N + c_i u_i^N = f_i \text{ for } i = 1, \ldots, N-1, \quad u_0^N = u_N^N = 0.$$

The condition number of the associated discrete linear system is shown in [Roo96] to be $\mathcal{O}(N^2 \ln^{-2} N)$.

Savin [Sav95] proves that there exists a constant C such that

$$\|u - u^N\|_{\infty,d} \leq C N^{-2} \ln^2 N. \tag{2.156}$$

The proof of (2.156) is along the lines of Remark 2.98; cf. [LM03] where a system of reaction-diffusion equations is analysed. A nonlinear generalization of (2.155)–(2.156) is discussed in [KS04].

A problem whose diffusive and convective terms are each multiplied by a small parameter is examined in [RU03]. Here one is in a convection-diffusion or reaction-diffusion regime, depending on the relative sizes of these parameters. Discrete Green's functions are used to prove the bound (2.156) for the solution of a difference scheme on a Shishkin mesh.

Non-upwinded finite element methods (which give rise to schemes of central-difference type) on Shishkin meshes for higher-order differential equation of both convection-diffusion and reaction-diffusion type are considered in [SS95a, SS95b], where uniform convergence results are proved in the ε-weighted energy norm of (2.62). This analysis is extended in [LX05, LX06] and applied to other meshes.   ♣

Remark 2.107. (Turning point problems) In [LV01] Linß and Vulanović investigate the boundary turning point problem

$$-\varepsilon u''(x) - x b(x) + x c(x, u(x)) = 0 \text{ for } 0 < x < 1, \quad u(0) = u(1) = 0,$$

where $b(x) > 0$, $c(0, u(0)) = 0$ and $c_u(x, u) \geq 0$. The solution u^N of an upwind scheme on a Shishkin mesh is shown to satisfy $\|u - u^N\|_{\infty,d} \leq C N^{-1} \ln^2 N$. (A generalization of this problem is considered in [Lin03b].) Liseikin [Lis90] transforms the problem to one whose second-order derivatives are bounded, so simple upwinding on a uniform mesh can be used to generate a solution u^N for which $\|u - u^N\|_{\infty,d} \leq C N^{-1}$.

In [SS94] a finite element method on a generalized Shishkin mesh is applied to a problem with interior turning points and convergence results, uniform in ε, are proved in weighted energy norms.   ♣

Remark 2.108. (Nonlinear problems) A quasilinear analogue of (2.133) is

$$-\varepsilon u'' + b(x, u)' + c(x, u) = f(x) \text{ for } 0 < x < 1, \quad u(0) = u(1) = 0, \tag{2.157}$$

with $b_u(\cdot, \cdot) \geq \beta > 0$ and $c_u(\cdot, \cdot) \geq 0$. This problem has a unique solution u for which the a priori bounds of Lemma 1.8 hold true – see [Vul89]. Many

of the earlier convergence results for (2.133) are also valid for (2.157); one can derive them by linearizing (2.157) about the solution u (cf. the proof of Lemma 2.111 below) then applying our previous analysis. We shall not give details here, but merely refer the reader to the papers [FOMS01, KL01, Kop01a, KS01b, Lin01c, Lin01a, LRV00, Shi92a, Vul01].

For singularly perturbed nonlinear problems that do not have unique solutions, it can be tricky to compute reliably accurate numerical solutions. In the case of the nonlinear reaction-diffusion problem considered in [KS07b] it is found that if one takes a Shishkin mesh suited to the asymptotic structure of the interior layer that appears in a true solution, and centres this mesh at *any* point in $(0, 1)$, then the numerical solution obtained will have an interior layer at that point!   ♣

Remark 2.109. (Systems) Systems of convection-diffusion problems are solved numerically in [Lin07b, OSS] using forms of upwinding on Shishkin meshes and the bound $\|u - u^N\|_{\infty,d} \leq CN^{-1} \ln N$ is proved. An adaptive approach for these problems is followed in [Lin07a].

The numerical solution of systems of reaction-diffusion problems on layer-adapted meshes is handled using finite difference techniques in [LM]. A related finite element method is analysed in [Lin08a].

2.5 Adaptive Strategies Based on Finite Differences

In the previous section we discussed special meshes that were chosen a priori for the solution of convection-diffusion and reaction-diffusion problems. There is a growing interest in the construction of meshes suitable for these problems by means of an alternative methodology: *adaptive mesh generation*, where information extracted from the computed solution on the current mesh is used to form a new mesh that is more appropriate for the problem, and this process is repeated iteratively until some stopping criterion is met.

Adaptive mesh generation using finite element methods will be discussed in the context of multi-dimensional problems in later chapters, since the theory is no different for one-dimensional problems and most numerical experimentation has been for partial differential equations.

Thus in the current section we confine ourselves to adaptive techniques based on finite difference methods. Our aim is to develop an adaptive mesh algorithm that, starting from some ordinary coarse mesh, will eventually generate a problem-fitted mesh on which the computed solution is guaranteed to be an accurate approximation of the solution to a convection-diffusion two-point boundary value problem.

This aspiration is not new. An early attempt to generate satisfactory meshes based on computational data is found in [KNB86], where systems of convection-diffusion problems comprising both turning point problems (see Section 1.2) and non-turning point problems are considered, and numerical

results are presented. More recently, [LG97] contains an adaptive algorithm that is based on reformulating the two-point boundary value problem as an integral equation; plausible numerical results are obtained for a variety of difficult singular perturbation problems, but no rigorous analysis is given of the dependence of the algorithm on the singular perturbation parameter.

To devise a rigorous analysis of any adaptive algorithm, the first ingredient one needs is an *a posteriori bound on the error in the computed solution*; that is, an error bound that is expressed entirely in terms of local data of the current computed solution. Such estimates give implicit guidance for improvement of the mesh. Bounds such as (2.147), although they depend on the current mesh, are not full a posteriori bounds since they involve the (unknown) true solution u.

As we are dealing with finite difference methods, it is natural to seek an a posteriori bound for $\|u - u^N\|_{\infty,d}$, where $\{u_i^N\}_{i=1}^N$ is the computed solution. The only published bound of this type is by Kopteva [Kop01a], and her exposition is followed here.

Consider a quasilinear two-point boundary value problem in conservation form:

$$Tu(x) := -\varepsilon u''(x) + b(x, u(x))' = f(x) \text{ for } x \in (0,1), \quad u(0) = u(1) = 0,$$
(2.158)

where a prime denotes differentiation with respect to x. It is assumed that $b \in C^1([0,1] \times \mathbb{R})$, $f \in C([0,1])$ and

$$0 < \beta \le b_u(x,u) \le \bar{\beta} \quad \text{for all } x \in [0,1] \text{ and all } u \in \mathbb{R},$$
(2.159)

for some constants β and $\bar{\beta}$. Then (2.158) has a unique solution $u \in C^2[0,1]$ that has an exponential boundary layer at $x = 1$; see [Vul89].

Set

$$Av(x) = -\varepsilon v'(x) + b(x, v(x))$$
(2.160)

for all $v \in C^2[0,1]$. Clearly (2.158) can be written as $(Au)' = f$.

Consider an arbitrary mesh $\bar{\omega} := \{x_0, x_1, \ldots, x_N\}$, where the discretization parameter N is a positive integer and $0 = x_0 < x_1 < \cdots < x_N = 1$. We use the notation $h_i, \hbar_i, H, D_-v_i$, etc. from Section 2.4. No assumption can be made that the mesh satisfies $H \le CN^{-1}$ for some constant C; such a property must be shown to hold true for the meshes generated by any adaptive algorithm.

Define the discrete operator A^N for any mesh function $\{v_i\}$ by

$$A^N v_i = -\varepsilon D_+ v_i + b(x_i, v_i).$$

To solve (2.158) we shall examine the conservation-form upwind scheme of [Kop01a, LRV00, Vul01]:

$$T^N u_i^N := D_-(A^N u_i^N) = f_i \text{ for } i = 1, \ldots, N-1, \quad u_0^N = 0, \ u_N^N = 0. \quad (2.161)$$

Here $\{u_i^N\}$ is the solution computed on the mesh $\{x_i\}$.

Recall from Section 1.1.2 the norm

$$\big\|v(x)\big\|_* = \min_{V:V'=v} \big\|V(x)\big\|_\infty \tag{2.162}$$

on the Sobolev space $W^{-1,\infty}$.

First consider the linear case of (2.158):

$$Lu \equiv -\varepsilon u'' + (p(x)u)' = f(x), \text{ for } x \in (0,1), \quad u(0) = u(1) = 0, \tag{2.163}$$

where $p \in C[0,1]$, and in accordance with (2.159),

$$0 < \beta \le p(\cdot) \le \bar{\beta}. \tag{2.164}$$

In the subsequent analysis the function f in (2.163) will often have the form $f(x) = F'(x)$, where $F(x)$ is a bounded piecewise continuous function, so f may have isolated singularities similar to the Dirac delta distribution and problem (2.163) is to be understood in the sense of distributions. In the next lemma, $u \in C^{0,1}[0,1] \subset H^1(0,1) \subset C[0,1]$, where $C^{0,1}[0,1]$ and $H^1(0,1)$ are the standard Hölder space and Sobolev space.

Lemma 2.110. *Suppose that $f(x) = F'(x)$, where $F(x)$ is a bounded piecewise continuous function. Then (2.163) has a unique solution $u \in C[0,1]$ and*

$$\|u\|_\infty \le (2/\beta)\|Lu\|_*.$$

Proof. By integration one can see that (2.163) has the unique continuous solution $u(x) = W(x) - W(0)V(x)/V(0)$, where

$$W(x) = \int_x^1 \frac{1}{\varepsilon} F(s) \exp\left\{-\frac{1}{\varepsilon}\int_x^s p(t)\,dt\right\}ds,$$

$$V(x) = \int_x^1 \frac{1}{\varepsilon} \exp\left\{-\frac{1}{\varepsilon}\int_x^s p(t)\,dt\right\}ds.$$

Now $|W(x)| \le \|F\|_\infty V(x)$, and (2.164) easily yields $0 \le V(x) \le \beta^{-1}$. Hence $|u(x)| \le 2\|F\|_\infty V(x) \le (2/\beta)\|F\|_\infty$, and the desired result follows. $\square$

This result can be regarded as an extension of the bound on $\|u\|_\infty$ given by (1.20) to a larger class of f in (2.163). We now generalize it to the quasilinear operator T using a standard linearization technique.

Lemma 2.111. *Let $v, w \in H^1(0,1)$ with $v(0) = w(0)$, $v(1) = w(1)$ and*

$$Tv(x) - Tw(x) = F'(x),$$

where $F(x)$ is a bounded piecewise continuous function. Then

$$\|v - w\|_\infty \le (2/\beta)\|Tv - Tw\|_*.$$

Proof. Now $Tv(x) - Tw(x) = L[v(x) - w(x)]$, where the operator L is linear and defined by (2.163) with

$$p(x) := \int_0^1 b_u\big(x, w(x) + s[v(x) - w(x)]\big)\, ds.$$

Then (2.159) implies that the condition (2.164) is satisfied, and the argument is completed by invoking Lemma 2.110. $\square$

Define $u^N(x)$ to be the piecewise linear interpolant through the knots (x_i, u_i^N) given by the computed solution. The desired *a posteriori error estimate* can now be stated.

Theorem 2.112. *Suppose that* $f(x) \in C^1[0,1]$. *Then there exists a constant* C *such that*

$$\|u^N - u\|_\infty \leq (2/\beta)\Big[\bar{\beta} \max_i |u_i^N - u_{i-1}^N| + CH\Big].$$

Proof. By Lemma 2.111, it suffices to prove that

$$\big\|Tu^N(x) - Tu(x)\big\|_* \leq \bar{\beta} \max_i |u_i^N - u_{i-1}^N| + CH. \qquad (2.165)$$

It follows from (2.158) and (2.160) that

$$Tu^N(x) - Tu(x) = Tu^N(x) - f(x) = \big(Au^N(x) - F(x) - a\big)', \qquad (2.166)$$

where $F(x) := \int_0^x f(x)\, dx$ and a is any constant. Set $F_i^N = \sum_{j=1}^i h_j f_j$ for $i = 1, \ldots, N$. By (2.161),

$$A^N u_i^N - F_i^N = A^N u_0^N \quad \text{for } i = 1, 2, \ldots, N-1.$$

Taking $a = A^N u_0^N$ in (2.166) yields $Tu^N(x) - Tu(x) = \eta'(x)$ for $x \in (x_{i-1}, x_i)$ and $i = 1, 2, \ldots, N$, where $\eta(x) = \bar{\eta}(x) + \tilde{\eta}(x)$ with $\bar{\eta}(x) = Au^N(x) - A^N u_i^N$ and $\tilde{\eta}(x) = F_i^N - F(x)$. It is easy to check that $|\tilde{\eta}(x)| \leq CH$; thus (2.162) gives

$$\|Tu^N(x) - Tu(x)\|_* \leq \|\eta(x)\|_\infty \leq \|\bar{\eta}(x)\|_\infty + CH. \qquad (2.167)$$

Now for $x \in (x_{i-1}, x_i)$, we have

$$\sup_{x \in (x_{i-1}, x_i)} |\bar{\eta}(x)| = \sup_{x \in (x_{i-1}, x_i)} \big|b(x_i, u_i^N) - b(x, u^N(x))\big|$$

$$= \sup_{x \in (x_{i-1}, x_i)} \left| \int_x^{x_i} \frac{d}{dx} b(x, u^N(x)) dx \right|$$

$$\leq \int_{x_{i-1}}^{x_i} \left| \frac{d}{dx} b(x, u^N(x)) \right| dx$$

$$\leq h_i \sup_{x \in (x_{i-1}, x_i)} \big|b_x(x, u^N(x))\big| + \bar{\beta}|u_i^N - u_{i-1}^N|,$$

since $\left[u^N(x)\right]' = (u_i^N - u_{i-1}^N)/h_i$. But the linearization of T^N around 0 yields an M-matrix, so a modification of [Lin01b, Lemma 1] shows that

$$\|u^N\|_\infty \le \frac{2}{\beta}(\|b(x,0)\|_\infty + \|f\|_\infty). \tag{2.168}$$

Consequently the above bound on $\bar\eta$ and $b \in C^1\big([0,1] \times \mathbb{R}\big)$ imply that

$$\|\bar\eta\|_\infty \le CH + \bar\beta \max_i |u_i^N - u_{i-1}^N|.$$

Recalling (2.167), we have verified (2.165) and the proof is complete.   □

Remark 2.113. Theorem 2.112 is a first-order estimate because at best one can deduce $\mathcal{O}(H)$ convergence from it. It is shown in [Kop01a] to hold true also for certain difference schemes other than (2.161), and for some of these schemes a second-order a posteriori estimate is derived.

In [Kop05] a second-order a posteriori estimate is proved for a reaction-diffusion problem.   ♣

The bound of Theorem 2.112 is, up to a constant factor, equivalent to

$$\|u^N - u\|_\infty \le C \max_i h_i \sqrt{1 + (D^- u_i^N)^2}\,.$$

The right-hand side here measures the length of the longest line segment from (x_{i-1}, u_{i-1}^N) to (x_i, u_i^N) in the piecewise linear function $u^N(x)$. To obtain an accurate approximate solution, we should strive to make $\max_i h_i \sqrt{1 + (D^- u_i^N)^2}$ as small as possible. Observations such as this have led many authors (e.g., [BM00, HRR94, Kop01a, KS01c, Lin01b, Mac99, RL99, QS99, QST00]) to focus on the twin ideas of monitor functions, which were defined on page 118, and *equidistribution.*

A mesh $\{x_i\}_{i=1}^N$ is said to equidistribute a monitor function $M(\cdot)$ if

$$\int_{x_{i-1}}^{x_i} M(x)\,dx = \frac{1}{N}\int_0^1 M(x)\,dx \quad \text{for } i = 1,\ldots,N.$$

Most authors choose M in their algorithms to be some discrete analogue of the standard arc-length function

$$M_{arc}(x) = \sqrt{1 + (u'(x))^2}. \tag{2.169}$$

While many published papers use monitor functions, only [KS01c] gives a full and rigorous analysis of an adaptive mesh algorithm based on them. The results of this paper are described in the remainder of this section.

Our adaptive mesh algorithm differs from one proposed by de Boor [dB74] only in its choice of stopping criterion. The algorithm solves (2.158) by means of the difference scheme (2.161), using equidistribution with the monitor function $M^N(x) := \sqrt{1 + ((u^N)'(x))^2}$, which is a discrete analogue of (2.169).

Thus, setting $M_i := \sqrt{1 + (D^- u_i^N)^2}$ for $i = 1, \ldots, N$, we are concerned with the following *equidistribution problem*: find $\{(x_i, u_i^N)\}$, with the $\{u_i^N\}$ computed from the $\{x_i\}$ by means of (2.161), such that

$$h_i M_i = \frac{1}{N} \sum_{j=1}^{N} h_j M_j \text{ for } i = 1, 2, \ldots, N. \tag{2.170}$$

Note here that both $\{x_i\}$ and $\{u_i^N\}$ are a priori unknown. Consequently, even if (2.158) is linear, the equidistribution problem is nonlinear because it requires the simultaneous solution of (2.161) and (2.170). This nonlinearity is a serious obstacle to the analysis of adaptive algorithms based on equidistribution.

To give a sense of the issues involved in analysing an adaptive algorithm, four fundamental questions will be addressed:

1. Does the equidistribution problem have a solution?
2. If $\{(x_i, u_i^N)\}$ is a solution of the equidistribution problem, will $u^N(x)$ be an accurate approximation of $u(x)$ on $[0, 1]$?
3. Is there an easily-implemented algorithm that can be proved to yield an accurate solution to the equidistribution problem when it terminates?
4. Can one prove that such an algorithm yields an accurate approximate solution after a predetermined number of iterations?

In order to answer these questions, we introduce our algorithm. The number N of mesh intervals is fixed throughout.

Adaptive Mesh Algorithm:

1. Initialize mesh: The initial mesh $\{0, 1/N, 2/N, \ldots, 1\}$ is equidistant.
2. For $k = 0, 1, \ldots$, given the mesh $\{x_i^{(k)}\}$, compute the discrete solution $\{u_i^{(k)}\}$ satisfying

$$T^{(k)} u^{(k)} = f^{(k)} \quad \text{on } \{x_i^{(k)}\}, \qquad \text{with} \quad u_0^{(k)} = u_N^{(k)} = 0,$$

where $f^{(k)} = \{f_i\}$. Let $h_i^{(k)} = x_i^{(k)} - x_{i-1}^{(k)}$ for each i. Let

$$\ell_i^{(k)} = h_i^{(k)} \sqrt{1 + (D^- u_i^{(k)})^2} = \sqrt{(u_i^{(k)} - u_{i-1}^{(k)})^2 + (h_i^{(k)})^2}$$

be the arc-length between the points $(x_{i-1}^{(k)}, u_{i-1}^{(k)})$ and $(x_i^{(k)}, u_i^{(k)})$ in the piecewise linear computed solution $u^{(k)}(x)$. Then the total arc-length of the solution curve $u^{(k)}(x)$ is

$$L^{(k)} := \sum_{i=1}^{N} \ell_i^{(k)} = \sum_{i=1}^{N} h_i^{(k)} \sqrt{1 + (D^- u_i^{(k)})^2}$$

$$= \sum_{i=1}^{N} \sqrt{(u_i^{(k)} - u_{i-1}^{(k)})^2 + (h_i^{(k)})^2}.$$

3. Test mesh: Let C_0 be a user-chosen constant with $C_0 > 1$ (see Remark 2.114). If

$$\frac{\max_i \ell_i^{(k)}}{L^{(k)}} \leq \frac{C_0}{N}, \qquad (2.171)$$

then go to Step 5. Otherwise, continue to Step 4.

4. Generate a new mesh by equidistributing the arc-length of the current computed solution: Choose points $0 = x_0^{(k+1)} < x_1^{(k+1)} < \cdots < x_N^{(k+1)} = 1$ such that for each i the distance from $(x_{i-1}^{(k+1)}, u^{(k)}(x_{i-1}^{(k+1)}))$ to $(x_i^{(k+1)}, u^{(k)}(x_i^{(k+1)}))\}$, *measured along the polygonal solution curve* $u^{(k)}(x)$, equals $L^{(k)}/N$. (This clearly determines the $x_i^{(k+1)}$ uniquely.) Our new mesh is then defined to be $\{x_i^{(k+1)}\}$. Return to Step 2.

5. Set $\{x_0^*, x_1^*, \ldots, x_N^*\} = \{x_i^{(k)}\}$ and $u^* = u^{(k)}$ then stop.

Remark 2.114. In (2.171) we can choose any constant C_0 that satisfies $C_0 > 1$. The larger C_0 is, the fewer iterations needed by the algorithm, but the constant factor in the final error bound of Theorem 2.118 increases with C_0. If we set $C_0 = 1$, then the algorithm is attempting to compute a fixed point of Theorem 2.116, so when $C_0 \approx 1$, we expect that the computed solution lies near such a fixed point. ♣

First, we prove a preliminary result.

Lemma 2.115. *Let $\{u_i^N\}$ be the solution of (2.161) on an arbitrary mesh $\{x_i\}$. There exists a constant C_1 such that $|D^- u_i^N| \leq C_1(N + \varepsilon^{-1})$ for $i = 1, 2, \ldots, N$.*

Proof. From (2.168) we have $|u_i^N| \leq C$ for all i. As the mesh has N subintervals, $h_m \geq N^{-1}$ for some m. These inequalities imply that $|D^- u_m^N| \leq CN$.

Let $i \in \{1, 2, \ldots, N\}$ be arbitrary. Assume that $i \leq m$, as the other case is similar. Now

$$A^N u_m^N - A^N u_i^N = \sum_{j=i}^{m-1} \left(A^N u_{j+1}^N - A^N u_j^N \right) = -\sum_{j=i}^{m-1} h_{j+1} f_j,$$

by (2.161). Hence $|A^N u_i^N| \leq \varepsilon |D^- u_m^N| + |b(x_m, u_m^N)| + \|f\|_\infty \leq C(\varepsilon N + 1)$. Consequently $\varepsilon |D^- u_i^N| \leq C(\varepsilon N + 1) + |b(x_i, u_i^N)|$, and the result follows. □

The following existence result answers Question 1 of page 146.

Theorem 2.116. *The equidistribution problem has a solution, i.e., there exists a mesh that equidistributes the arc-length monitor function along the piecewise linear interpolant to the solution of (2.161).*

Proof. One can regard Steps 2 and 4 of the Adaptive Mesh Algorithm as a mapping $\Phi : (h_1, h_2, \ldots, h_N) \mapsto (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_N)$, where the h_i and $\tilde{h}_i$ are the mesh-widths before and after regridding.

We claim that $\Phi : S_Q \to S_Q$, where

$$S_Q = \left\{ (h_1, h_2, \ldots, h_N) \in \mathbb{R}^N : \ h_i \geq Q \text{ for } i = 1, \ldots, N, \ \sum_{i=1}^{N} h_i = 1 \right\}$$

and $Q = Q(\varepsilon, N)$ satisfies $0 < Q < 1/N$. Note that these bounds on Q imply that S_Q is nonempty.

To prove this claim, let $\{u_i^N\}$ be the solution to (2.161) computed on the mesh with mesh-widths $h_1, h_2, \ldots, h_N$. Set $\ell_i = h_i \sqrt{1 + |D^- u_i^N|^2}$ for each i. By Lemma 2.115 the slope of each segment of $u^N(x)$ is at most $C_1(N + \varepsilon^{-1})$, so for each i the total arc-length of $u^N(x)$ on an interval of length $\tilde{h}_i$ is at most $\sqrt{\tilde{h}_i^2 + C_1 \tilde{h}_i^2 (N + \varepsilon^{-1})^2}$. Hence when Step 4 of the regridding algorithm is applied to partition the piecewise linear function $u^N(x)$ into N pieces, we have

$$\sqrt{\tilde{h}_i^2 + C_1 \tilde{h}_i^2 (N + \varepsilon^{-1})^2} \geq \frac{1}{N} \sum_i \ell_i \geq \frac{1}{N} \sum_i h_i = \frac{1}{N},$$

which implies that

$$\tilde{h}_i \geq Q := \frac{1}{N \sqrt{1 + C_1 (N + \varepsilon^{-1})^2}} \quad \text{for all } i.$$

Hence $0 < Q < 1/N$ and we see that Φ maps S_Q into itself.

The set S_Q is a nonempty convex compact subset of $\mathbb{R}^N$, and Φ is clearly continuous. By the Brouwer fixed-point theorem [Sma74], the function Φ has a fixed point in S. That is, there is a mesh on which the computed solution satisfies $\ell_i = \ell_j$ for all i and j. $\square$

Let $\{u_i^N\}$ be the solution of (2.161) on an arbitrary mesh $\{x_i\}$. Let $\mathcal{L}$ be the total arc-length along the solution curve $u^N(x)$. Then it can be shown [KS01c] that

$$1 \leq \mathcal{L} \leq C_2, \tag{2.172}$$

where $C_2 = 1 + 2\|f - T^N 0\|_\infty / \beta$.

We claim that on the meshes generated by the algorithm, for all i we have

$$h_i^{(k)} \leq C_2 N^{-1}. \tag{2.173}$$

Inequality (2.173) clearly holds true when $k = 0$ since $C_2 \geq 1$, so assume that $k > 0$. By Step 4 of the algorithm, the distance from $(x_{i-1}^{(k)}, u^{(k-1)}(x_{i-1}^{(k)}))$ to $(x_i^{(k)}, u^{(k-1)}(x_i^{(k)}))$, measured along the solution curve $u^{(k-1)}(x)$, equals $L^{(k-1)}/N$. Hence $h_i^{(k)} = x_i^{(k)} - x_{i-1}^{(k)} \leq L^{(k-1)}/N \leq C_2/N$, by (2.172).

Question 2 can now be answered by the next theorem: any solution to the equidistribution problem is an accurate approximation of the true solution u.

Theorem 2.117. *Let $\{u_i^N\}$ be the solution to (2.161) computed on a mesh $\{x_i\}$ that satisfies (2.170). Then*

$$\|u - u^N\|_\infty \le CN^{-1} \quad \text{for some constant } C, \tag{2.174}$$

and $\ell_i = L/N$ for $i = 1, \ldots, N$, where $\ell_i = \sqrt{(x_i - x_{i-1})^2 + (u_i^N - u_{i-1}^N)^2}$ is the local arc-length and the total arc-length of the solution curve is $L = \sum_{i=1}^N \ell_i$.

Proof. This result essentially appears in [Kop01a, Section 6]. First, $\ell_i = L/N$ for all i since (2.170) is true. To prove (2.174), observe that by (2.172) we have $\ell_i = L/N \le C_2/N$ for all i. Now Theorem 2.112 and (2.173) yield (2.174). □

The next result deals with Question 3 by showing that the final solution generated by our algorithm is an accurate approximation of u; for its proof, Theorem 2.112 is needed again.

Theorem 2.118. *Suppose that the Adaptive Mesh Algorithm reaches its stopping criterion and halts. Let the final mesh generated be $\{x_i^*\}$. Let $\{u_i^*\}$ be the discrete solution computed on this mesh, and let $u^*(x)$ be the piecewise linear interpolant of $\{(x_i^*, u_i^*)\}$. Then there exists a constant C such that*

$$\|u - u^*\|_\infty \le CN^{-1}.$$

Proof. Let $\ell_i^* = \sqrt{(u_i^* - u_{i-1}^*)^2 + (x_i^* - x_{i-1}^*)^2}$ denote the arc-length between successive knots in the polygonal computed solution on the final mesh. Inequalities (2.172) and (2.171) imply that $\ell_i^* \le C_0 C_2/N$ for all i. The result now follows from Theorem 2.112 and (2.173). □

Our final result gives a precise answer to Question 4 in the case where (2.158) is linear. The proof is much more difficult than the arguments presented thus far; for it we refer the reader to [KS01c].

Theorem 2.119. *Assume that the two-point boundary value problem (2.158) is linear. Let N be sufficiently large, independently of ε. Assume that $\varepsilon \le N^{-1}$. Then there exists a positive integer K, with $K \le C_3 |\ln \varepsilon|/(\ln N)$, such that $\|e^{(K)}\|_\infty \le C_4 N^{-1}$, where $e^{(k)}$ is the error in the kth solution computed by the algorithm. Here C_3 and C_4 are constants that are independent of N, ε and the meshes generated.*

Numerical results presented in [KS01c] show that in practice the number of iterations taken by the algorithm (with $C_0 = 2$) is indeed $\mathcal{O}(|\ln \varepsilon|/(\ln N))$.

Remark 2.120. Experimental evidence shows that the final mesh computed by the algorithm is strikingly close to a Bakhvalov mesh inside the boundary layer; see [KS01c, Fig. 2]. In contrast, most adaptive algorithms will not generate a mesh resembling a Shishkin mesh. ♣

Parabolic Initial-Boundary Value Problems in One Space Dimension

In Part II we leave two-point boundary value problems and move on to time-dependent (i.e., unsteady) problems. Only problems posed in one space dimension are examined here; this enables us to make full use of the experience gained in Part I. These time-dependent problems are more difficult than two-point boundary value problems but less demanding than elliptic (steady) problems in two space dimensions.

The second-order differential equations that will be discussed in Part II encompass a wealth of applications, as Chapter 1 will reveal. Our presentation concentrates on the motivation and analysis of numerical methods; for detailed numerical results we recommend [HV03, VK93]. The classical theory of parabolic partial differential equations is presented in [Fri64, LSU67].

Unsteady problems in more than one space dimension will be discussed in Chapter III.4.

1

Introduction

In Part II we shall work with parabolic partial differential equations on the rectangle $(0,1) \times (0,T]$ in the space-time domain, where T is some fixed positive time. It is not essential to have a rectangle; one can transform many other domains to rectangular form.

Thus, consider the initial-boundary value problem

$$u_t(x,t) - \varepsilon u_{xx}(x,t) + b(x,t)u_x(x,t) + d(x,t)u(x,t) = f(x,t) \quad \text{(1.1a)}$$

where $(x,t) \in Q := (0,1) \times (0,T]$, and

$$u(x,0) = s(x) \quad \text{on } S_x := \{(x,0) : 0 \le x \le 1\}, \quad \text{(1.1b)}$$
$$u(0,t) = q_0(t) \quad \text{on } S_0 := \{(0,t) : 0 < t \le T\}, \quad \text{(1.1c)}$$
$$u(1,t) = q_1(t) \quad \text{on } S_1 := \{(1,t) : 0 < t \le T\}. \quad \text{(1.1d)}$$

As in Part I, the notation L is used for the differential operator. That is,

$$Lu := u_t - \varepsilon u_{xx} + bu_x + du$$

throughout Part II.

Here, as in Part I, ε is a parameter satisfying $0 < \varepsilon \ll 1$. The functions b and c are assumed to be smooth on $\bar{Q} := [0,1] \times [0,T]$, the closure of Q; the data s of the initial condition (1.1b) and the data q_0 and q_1 of the boundary conditions (1.1b) and (1.1c) are usually assumed to be smooth on the closures of their respective domains. For the present we take f to be in $L_2(Q)$, but more smoothness of this function will sometimes be needed. Full hypotheses will be stated for each theoretical result.

Without loss of generality, one may assume that $d(x,t) \ge \gamma > 0$ on $\bar{Q}$ for some constant γ, since this property can be ensured by the change of variable $u(x,t) = v(x,t)e^{Ct}$ with some suitable constant C.

When b is not identically zero, L is a *convection-diffusion* operator: $-\varepsilon u_{xx}$ models diffusion while the combined first-order terms $u_t + bu_x$ represent convection. To see this clearly in the case where b is constant and $c \equiv f \equiv 0$, set $x' = x - bt$ and $w(x',t) = u(x' + bt, t)$. Then $w_t = bu_x + u_t$ and $w_{x'x'} = u_{xx}$.

Consequently w satisfies the heat equation $w_t = \varepsilon w_{x'x'}$. Restating these facts in terms of the original variables, we see that u is convected along lines of the form $x = bt + C$ at unit speed relative to a coordinate system moving with speed b in the positive x direction, and is also subject to diffusion of magnitude ε.

Returning to the solution of the general problem (1.1), the convection (or flow, as it is often called) travels in the direction of propagation of the first-order differential operator $w \mapsto w_t + bw_x$. The direction of propagation is often called the direction of flow. Continuing in this vein, a point $A \in \bar{Q}$ is said to be upstream of a point $B \in \bar{Q}$ if the curve of flow through A (that is, the characteristic curve of $w_t + bw_x$ that passes through A) also passes through B, with the direction of propagation along this curve pointing from A to B. Equivalently, one can say that B is downstream of A.

We shall concentrate on the case where $b > 0$ on $\bar{Q}$, much as we did in Part I. Once again, the mapping $x \mapsto 1 - x$ transforms the case $b < 0$ to the case $b > 0$.

The case when $b \equiv 0$ will also be mentioned occasionally. In this situation, the u_t term may again be interpreted as a convective term, but now the zero-order term du plays a significant role. Consequently this case is commonly referred to as being of *reaction-diffusion* type. See in particular Remark 2.11 and Section 3.4.3.

When all the data b, d, f, s, q_0, q_1 are smooth and $b > 0$, the solution u of (1.1) will be smooth on most of Q. Near the boundary $x = 1$ of Q, the solution will in general exhibit a boundary layer. For each fixed value of $t > 0$, the dependence of this layer on x is exactly the same as in the boundary layer of Section I.1.1. Thus once again we are dealing with a singularly perturbed problem. If we replace the Dirichlet condition (1.1d) by a Neumann condition, then as in Remark I.1.5 the solution has a less severe boundary layer at $x = 1$.

The solution u may also have one or more interior layers. Such layers have no exact counterpart in the solutions of ordinary differential equations. They can be caused by discontinuities in s or q_0, by insufficient compatibility of the initial-boundary data at the corner (0,0) of Q, or by singularities in f. The layer generated lies downstream of the discontinuity or singularity at which it is triggered. For example, if $b \equiv 2$, $d \equiv 0$, $f \equiv 0$, $s(x) \equiv 0$ and $q_0(t) \equiv 1$, then the discontinuity in the initial-boundary conditions at (0,0) propagates along the line $x = 2t$ as a layer in u; roughly speaking,

$$u(x,t) \approx \begin{cases} 1 \text{ if } x < 2t, \\ 0 \text{ if } x > 2t, \end{cases}$$

and the continuous function u changes rapidly as (x,t) crosses the line $x = 2t$. We discuss this phenomenon more rigorously immediately after Theorem 2.6.

Applications involving (1.1) arise for example in the linearized Navier-Stokes equations of fluid dynamics [Hir88, KL04], simulation of oil extraction from underground reservoirs [Ewi83], convective heat transport problems with large Péclet numbers [Jak59], electromagnetic field problems in moving media

[HBS87], miscible and multiphase flows [EW01], semiconductor device modelling [MRS90], and meteorology [Sal98]. Biological and chemical applications appear in [HV03]. A problem closely related to (1.1), where one of the boundary conditions is at $x = \infty$, appears in the study of unsteady hydromagnetic flow over a continuous moving flat surface for large suction Reynolds number [VR90].

If one tries to solve (1.1) using standard numerical methods for partial differential equations, then very inaccurate solutions are obtained unless the mesh discretization used is extremely fine (see [HBS87] for an example). That is, the situation is just as for the singularly perturbed ordinary differential equations of Part I: in order to get inexpensive but accurate numerical results, it will be necessary to devise methods that can cope with boundary and interior layers.

2

Analytical Behaviour of Solutions

2.1 Existence, Uniqueness, Comparison Principle

For a general discussion of the properties enjoyed by solutions of parabolic differential equations, the standard reference books are Friedman [Fri64] and Ladyženskaja et al. [LSU67]. The broad analysis presented there is classical in nature, dealing with solutions lying in Hölder spaces. Here we shall state only those fundamental results from [Fri64] that are necessary to provide a basis for our later work.

First, some notation and definitions. Let $\alpha \in (0,1)$. A function $w : Q \to R$ is said to be Hölder continuous on Q with exponent α if

$$|w|_\alpha^Q := \sup_{(x,t),(x',t')\in Q} \frac{|w(x,t) - w(x',t')|}{\left[\operatorname{dist}((x,t),(x',t'))\right]^\alpha} < \infty,$$

where we set $\operatorname{dist}((x,t),(x',t')) = \left((x-x')^2 + |t-t'|\right)^{1/2}$. For such a function w, let

$$\|w\|_\alpha^Q := |w|_\alpha^Q + \sup_{(x,t)\in Q} |w(x,t)|.$$

For all sufficiently smooth functions w, set

$$\|w\|_{2+\alpha}^Q := \|w\|_\alpha^Q + \|\partial w/\partial x\|_\alpha^Q + \|\partial^2 w/\partial x^2\|_\alpha^Q + \|\partial w/\partial t\|_\alpha^Q.$$

Now define the space

$$C^{2+\alpha}(Q) := \{w \in C(Q) : \|w\|_{2+\alpha}^Q < \infty\}.$$

From [Fr64, p. 65], one gets

Theorem 2.1. *Let* $\alpha \in (0,1)$. *Let* $s \in C^2[0,1], q_0 \in C^1[0,T]$ *and* $q_1 \in C^1[0,T]$ *with* $q_0(0) = s(0)$ *and* $q_1(0) = s(1)$. *Set*

$$\Psi(x,t) = s(x) + (1-x)[q_0(t) - q_0(0)] + x[q_1(t) - q_1(0)],$$

so that Ψ interpolates to the initial-boundary conditions. Assume that $\Psi \in C^{2+\alpha}(Q)$. Assume also that

$$q_0'(0) - \varepsilon s''(0) + b(0,0)s'(0) + d(0,0)s(0) = f(0,0), \qquad (2.1a)$$
$$q_1'(0) - \varepsilon s''(1) + b(1,0)s'(1) + d(1,0)s(1) = f(1,0). \qquad (2.1b)$$

Let b, d and f be Hölder continuous on Q with exponent α. Then (1.1) has exactly one solution in $C^{2+\alpha}(Q)$.

In this theorem, the equations $q_0(0) = s(0)$, $q_1(0) = s(1)$ and (2.1) are called *compatibility conditions*. When the data b, d, f, s, q_0 and q_1 of (1.1) are sufficiently differentiable, the solution u will be pointwise differentiable to any desired degree on the region Q. But if differentiability on the closed region $\bar{Q}$ is required, then regularity of the data alone does not suffice: one also needs a sufficient amount of compatibility of that data at the corners $(0,0)$ and $(1,0)$. Theorem 2.1 conforms to this statement, as it shows that $\|u\|_{2+\alpha}^Q$ is bounded; when regularity of u is lost at a point on the boundary ∂Q, one expects the relevant norm of u to blow up as one approaches that point from inside Q. If the above compatibility conditions were not satisfied while the data b, d and f remained Hölder continuous on Q, then a unique classical solution u of (1.1) would still exist but its regularity would not be guaranteed on all of ∂Q.

For a discussion of the relationship between compatibility conditions and the regularity of solutions to general parabolic differential equations see [Fri64, Sections 3.3 and 3.5]. These ideas are applied to (1.1) in [SO89]. We shall also mention some effects of compatibility conditions in Section 2.2.

Continuing with our examination of classical solutions to (1.1), one has the following *comparison principle*, which is equivalent [PW67] to the usual maximum principle for the operator L. (A more general *weak comparison principle* can be found in [GFL$^+$83] – cf. [GT83] for the elliptic case.)

Theorem 2.2. *Let b and c be bounded functions. Let $v, w \in C(\bar{Q})$. Assume that v and w are twice differentiable in x and once in t on Q. Suppose that*

$$|Lv(x,t)| \leq Lw(x,t) \quad \text{for all } (x,t) \in Q,$$
$$|v(x,t)| \leq w(x,t) \quad \text{on } S_x \cup S_0 \cup S_1.$$

Then

$$|v(x,t)| \leq w(x,t) \quad \text{for all } (x,t) \in \bar{Q}.$$

Proof. See Friedman [Fri64]. □

This theorem, and its discrete analogue that will appear in Lemma 3.12, are very useful in the analysis of asymptotic expansions and numerical methods for (1.1). As in Part I, the function v above will be the error in the asymptotic expansion or numerical solution and w will be chosen carefully to act as a barrier function for v.

2.2 Asymptotic Expansions and Bounds on Derivatives

Assume throughout Section 2.2 that in (1.1) one has $b > \beta > 0$ on $\bar{Q}$ for some constant β, so this is a convection-diffusion problem. Assume also that this problem has a unique solution u. Chapter 1 sketched the behaviour of u. We now elucidate this behaviour by showing how to construct an asymptotic expansion of u and by examining bounds on the derivatives of u.

For the asymptotic expansion, we shall follow the approach of Bobisud [Bob67]. (See [GFL+83] for a more general approach by means of a weak maximum principle.) Not all of the details are given here, as our main aims are to impart a sense of the methods used and an understanding of the nature of the solution of (1.1).

Lemma 2.3. *Assume in (1.1) that* b, d *and* f *are bounded on* Q. *Assume also that* $s \in C^2[0,1]$, $q_0 \in C^1[0,T]$ *and* $q_1 \in C^1[0,T]$, *with* $s(0) = q_0(0)$ *and* $s(1) = q_1(0)$. *Then there exists a constant* C, *which is independent of* x, t *and* ε, *such that the bounds*

$$|u(x,t) - s(x)| \le Ct, \tag{2.2}$$
$$|u(x,t) - q_0(t)| \le Cx \tag{2.3}$$

hold true for all $(x,t) \in \bar{Q}$.

Proof. Set $v(x,t) = u(x,t) - s(x)$. Then

$$Lv(x,t) = f(x,t) + \varepsilon s''(x) - b(x,t)s'(x) - d(x,t)s(x),$$
$$v(x,0) = 0 \text{ for } 0 < x < 1,$$
$$v(0,t) = q_0(t) - s(0) \text{ and } v(1,t) = q_1(t) - s(1) \text{ for } 0 \le t \le 1.$$

In particular our hypotheses imply that for $0 \le t \le 1$ one has

$$|v(0,t)| = |q_0(t) - q_0(0)| \le Mt \quad \text{and} \quad |v(1,t)| \le M't$$

for some constants M and M'.

On the other hand, setting $w(x,t) = Ct$ for any constant C, clearly

$$Lw(x,t) = C + Ctd(x,t),$$
$$w(x,0) = 0 \text{ for } 0 < x < 1,$$
$$w(0,t) = w(1,t) = Ct \text{ for } 0 \le t \le 1.$$

Using the hypotheses on the data of the problem, we can easily verify that, provided C is chosen sufficiently large, Theorem 2.2 applies to v and w. This proves (2.2). The proof of (2.3) is similar. □

Remark 2.4. Lemma 2.3 shows that the solution u of (1.1) does not have a layer near the sides $x = 0$ and $t = 0$ of $\bar{Q}$. For if $u \in C^1(\bar{Q})$, then (2.2) and (2.3) imply that $|u_t(x,0)| \le C$ and $|u_x(0,t)| \le C$ respectively.

In general any attempt at a similar argument will fail to prove that

$$|u(x,t) - q_1(t)| \leq C(1-t) \quad \forall (x,t) \in Q. \tag{2.4}$$

The inequality (2.4), if true, would imply that u had no boundary layer at the side $x = 1$ of $\bar{Q}$.

Note how the compatibility conditions $s(0) = q_0(0)$ and $s(1) = q_1(0)$ are central to the proof of Lemma 2.3. This hints that, without such assumptions, the solution u may not be so well behaved. This is indeed the case: for example, $s(0) \neq q_0(0)$ causes an interior layer in the solution, as we described in Chapter 1.   ♣

We now construct an asymptotic expansion for the solution u of (1.1). The basic approach is a natural generalization of the technique used in Section I.1.1. Nevertheless, the nonsmoothness of the boundary ∂Q of Q at the point (0,0) causes particular difficulties that require special treatment.

Definition 2.5. *The reduced problem associated with (1.1) when $b > 0$ on $\bar{Q}$ is defined by*

$$(u_0)_t + b(u_0)_x + du_0 = f \quad \text{on } Q, \tag{2.5a}$$
$$u_0(x,0) = s(x) \quad \text{on } S_x, \tag{2.5b}$$
$$u_0(0,t) = q_0(t) \quad \text{on } S_0. \tag{2.5c}$$

Since $b > 0$ on $\bar{Q}$, this first-order problem has a unique solution, which we denote by $u_0(x,t)$ and call the *reduced solution*. Analogously to Section I.1.1, the reduced problem is obtained from (1.1) by formally setting $\varepsilon = 0$ in the differential equation and discarding the boundary condition from the side of Q where u has a boundary layer.

Theorem 2.6. *Let $b, d, f \in C^2(\bar{Q})$, $s \in C^4[0,1]$ and $q_0, q_1 \in C^3[0,T]$. Assume that $s(0) = q_0(0)$ and $s(1) = q_1(0)$. Then the solution u of (1.1) has the asymptotic expansion*

$$u(x,t) = u_0(x,t) + v(x,t) + w(x,t), \tag{2.6}$$

where u_0 is the solution of the reduced problem (2.5), $v(x,t)$ is a boundary layer function (defined in (2.12) below) that decays exponentially as one moves away from $x = 1$, and $|w(x,t)| \leq C\sqrt{\varepsilon}$.

Proof. A simplified version of the argument in [Bob67] will be presented. First, we show that it is sufficient to consider the case of homogeneous initial-boundary conditions. Set

$$p(x,t) = s(x) + (1-x)[q_0(t) - q_0(0)] + x[q_1(t) - q_1(0)],$$

so $p = u$ on $S_x \cup S_0 \cup S_1$. Let $v = u - p$. Then

$$Lv = f - p_t + \varepsilon p_{xx} - bp_x - dp \tag{2.7}$$

with $v \equiv 0$ on $S_x \cup S_0 \cup S_1$.

This is almost what we want; it is not quite perfect because the right-hand side of (2.7) depends on ε. To remedy this defect, set $v = v_1 + \varepsilon v_2$, where v_2 is the solution of the problem

$$Lv_2 = p_{xx} \quad \text{on } Q,$$
$$v_2 = 0 \quad \text{on } S_x \cup S_0 \cup S_1.$$

It is easy to verify that $|v_2| \leq \|p_{xx}\|_{L_\infty(Q)}/\gamma$ on $\bar{Q}$, using Theorem 2.2. Hence $|\varepsilon v_2| \leq C\varepsilon$ on $\bar{Q}$, so εv_2 can be absorbed into w in (2.6). This leaves v_1, which satisfies $Lv_1 = f - p_t - bp_x - dp$ with homogeneous initial-boundary conditions.

During the rest of the proof, suppose that $s \equiv q_0 \equiv q_1 \equiv 0$ in (1.1). Then the reduced solution u_0 will not in general be C^1 across the characteristic curve of (2.5a) that passes through (0,0). This lack of smoothness hinders our later arguments, so we shall show that u_0 can be approximated to order $\sqrt{\varepsilon}$ by a C^2 function $\tilde{u}_0$ that is the solution of a problem closely related to (2.5).

Integrating (2.5) along its characteristic curves, one can check that u_0 lies in $C^2(\bar{Q})$ if and only if f, f_x and f_t satisfy a certain compatibility condition at (0,0); the details of this computation are in [Bob67]. For the present homogeneous initial-boundary data, the compatibility condition holds true if $f(0,0) = f_x(0,0) = f_t(0,0) = 0$. This observation motivates the construction of $\tilde{u}_0$ below, where we introduce a cut-off function that is tantamount to f being identically zero in a neighbourhood of (0,0).

Let $z : [0, \infty) \to [0, 1]$ be C^∞, with $z(y) = 1$ for $0 \leq y \leq 1$ and $z(y) = 0$ for $y \geq 2$. Then $u_0 = \tilde{u}_0 + \hat{u}_0$, where these new functions are defined by

$$\big((\tilde{u}_0)_t + b(\tilde{u}_0)_x + d\tilde{u}_0\big)(x,t) = \big[1 - z(t/\sqrt{\varepsilon})\big]f(x,t) \quad \text{on } Q, \tag{2.8}$$
$$\tilde{u}_0 = 0 \quad \text{on } S_x \cup S_0,$$
$$\big((\hat{u}_0)_t + b(\hat{u}_0)_x + d\hat{u}_0\big)(x,t) = z(t/\sqrt{\varepsilon})f(x,t) \quad \text{on } Q, \tag{2.9}$$
$$\hat{u}_0 = 0 \quad \text{on } S_x \cup S_0.$$

Now $\tilde{u}_0 \in C^2(\bar{Q})$, since $[1 - z(t/\sqrt{\varepsilon})]f(x,t)$ satisfies the compatibility conditions mentioned earlier. Also, integrating (2.9) along its characteristic curves easily yields $|\hat{u}_0| \leq C\sqrt{\varepsilon}$ on $\bar{Q}$. The term $\hat{u}_0$ will later be absorbed into w in (2.6).

Next, the boundary layer term $v(x,t)$ of (2.6) is constructed. Lemma 2.3 implies that a boundary layer can occur only near the side $x = 1$ of $\bar{Q}$. Thus define a local variable by $\xi := (x-1)/\varepsilon$ and, setting $v^*(\xi,t) := v(x,t)$, rewrite the homogeneous differential equation $Lv = 0$ in terms of ξ:

$$v_t^* - \varepsilon^{-1}v_{\xi\xi}^* + \varepsilon^{-1}bv_\xi^* + dv^* = 0. \tag{2.10}$$

We wish to choose v^* to satisfy (2.10) up to $\mathcal{O}(\varepsilon^{-1})$, with $v^*(0,t) = -\tilde{u}_0(1,t)$ for $0 \leq t \leq T$, and $\lim_{\varepsilon \to 0} v^*(\xi,t) = 0$ for each fixed $x < 1$ and $t \in [0, T]$.

Fix $t \in [0, T]$. For $0 \leq x \leq 1$, one has $\varepsilon^{-1}b(x, t) = \varepsilon^{-1}b(1, t) + \mathcal{O}(\xi)$. Substituting this into (2.10) and equating the coefficients of ε^{-1} to zero yields

$$-v^*_{\xi\xi} + b(1, t)v^*_\xi = 0. \tag{2.11}$$

Now define v^* by requiring it to satisfy (2.11), $v^*(0, t) = -\tilde{u}_0(1, t)$ and $\lim_{\varepsilon \to 0} v^*(\xi, t) = 0$ for each fixed $x < 1$. This forces the choice

$$v^*(\xi, t) = -\tilde{u}_0(1, t)e^{\xi b(1, t)}$$

for $\xi < 0$ and $0 \leq t \leq T$. That is,

$$v(x, t) = -\tilde{u}_0(1, t)e^{-b(1, t)(1-x)/\varepsilon} \tag{2.12}$$

for $(x, t) \in \bar{Q}$.

To complete the proof, we need to show that $|u - (u_0 + v)|$ is $\mathcal{O}(\sqrt{\varepsilon})$. Since $u_0 = \tilde{u}_0 + \hat{u}_0$ and $\hat{u}_0$ is $\mathcal{O}(\sqrt{\varepsilon})$, it is sufficient to bound $|u - (\tilde{u}_0 + v)|$ by $C\sqrt{\varepsilon}$. This will be done by means of the comparison principle of Theorem 2.2.

Set $\eta = u - (\tilde{u}_0 + v)$. Since $\tilde{u}_0 \in C^2(\bar{Q})$ one can apply the operator L to η. Furthermore, writing down an explicit formula for $\tilde{u}_0$ (see [Bob67]) shows that on $\bar{Q}$ one has

$$|\tilde{u}_0| = \mathcal{O}(1), \quad |(\tilde{u}_0)_t| = \mathcal{O}(1), \quad |(\tilde{u}_0)_{xx}| = \mathcal{O}(\varepsilon^{-1/2}). \tag{2.13}$$

Now by (1.1a), (2.8) and (2.13),

$$\begin{aligned}
|L\eta(x, t)| &= |f(x, t) + \varepsilon(\tilde{u}_0)_{xx}(x, t) - (1 - z(t/\sqrt{\varepsilon}))f(x, t) \\
&\quad - (v_t - \varepsilon v_{xx} + bv_x + dv)(x, t)| \\
&\leq z(t/\sqrt{\varepsilon})|f(x, t)| + C\sqrt{\varepsilon} \\
&\quad + |(b(1, t) - b(x, t))v_x(x, t) - (v_t + dv)(x, t)| \\
&\leq z(t/\sqrt{\varepsilon})|f(x, t)| + C\sqrt{\varepsilon} + Ce^{-b(1, t)(1-x)/2\varepsilon}
\end{aligned} \tag{2.14, 2.15}$$

for $(x, t) \in Q$, where we used (2.12) and the inequality

$$\varepsilon^{-1}(1 - x)e^{-b(1, t)(1-x)/2\varepsilon} \leq C \quad \text{for } x \leq 1.$$

Hence

$$|L\eta(x, t)| \leq C[\sqrt{\varepsilon} + e^{-\beta(1-x)/2\varepsilon} + z(t/\sqrt{\varepsilon})] \quad \text{for } (x, t) \in Q. \tag{2.16}$$

Furthermore, $\eta = 0$ on $S_x \cup S_1$ and $|\eta|$ is exponentially small on S_0.

Consider the function $\theta(x, t) = M\sqrt{\varepsilon}e^{t/\sqrt{\varepsilon}}$, where the constant M will be chosen later. We show that θ is a barrier function for η. Clearly any value of M satisfying $M \geq 1$ will yield $|\eta| \leq \theta$ on $S_x \cup S_0 \cup S_1$. Also,

$$(L\theta)(x, t) = M(1 + d\sqrt{\varepsilon})e^{t/\sqrt{\varepsilon}} \quad \text{for } (x, t) \in Q.$$

From (2.16) it is clear that the constant M can be chosen (independently of ε) so that θ is a barrier function for η.

Of course θ is quite large on most of Q and consequently useless there. Nevertheless, it does show that

$$|\eta(x,t)| \leq Me^2\sqrt{\varepsilon} \leq C\sqrt{\varepsilon} \quad \text{for } 0 \leq x \leq 1 \quad \text{and} \quad 0 \leq t \leq 2\sqrt{\varepsilon}. \qquad (2.17)$$

The significance of this bound is that it estimates $|\eta|$ in a satisfactory way on that part of Q where there is a contribution to (2.16) from $z(t/\sqrt{\varepsilon})$.

Observe next that the comparison principle of Theorem 2.2 holds true also on the (x,t)-domain $Q' := (0,1) \times (2\sqrt{\varepsilon}, T]$, when one makes appropriate superficial changes in the statement of this Lemma. This will be used to bound $|\eta|$ on Q'.

From (2.17) and our previous comments regarding η on S_0 and S_1, the initial-boundary data for η on Q' satisfies

$$|\eta| \leq C\sqrt{\varepsilon}. \qquad (2.18)$$

For $(x,t) \in Q'$, (2.16) becomes

$$|L\eta(x,t)| \leq C(\sqrt{\varepsilon} + e^{-\beta(1-x)/2\varepsilon}). \qquad (2.19)$$

Let $\phi(x,t) = M'(\sqrt{\varepsilon} + \varepsilon e^{-\beta(1-x)/2\varepsilon})$ be our barrier function, where the constant M' will be chosen in a moment. Any choice of $M' \geq C$, where C is as in (2.18), gives $|\eta| \leq \phi$ for the initial-boundary data on Q'. Now

$$L\phi(x,t) = M'\{\sqrt{\varepsilon}\,d(x,t) + [b(x,t)\beta/2 - \beta^2/4 + d(x,t)]e^{-\beta(1-x)/2\varepsilon}\}.$$

Hence, using $d \geq \gamma > 0$ and $b > \beta$ on $\bar{Q}$ and (2.19), one can choose an M' that is bounded independently of ε to yield $L\phi(x,t) \geq |L\eta(x,t)|$ on Q'.

The comparison principle now gives $|\eta| \leq \phi \leq C\sqrt{\varepsilon}$ on Q'. Combining this with (2.17) gives finally $|\eta| \leq C\sqrt{\varepsilon}$ on Q.

Recalling our earlier remarks in the proof, we have shown that

$$|u - (u_0 + v)| \leq C\sqrt{\varepsilon} \quad \text{on } Q.$$

Set $w = u - (u_0 + v)$ to complete the argument. □

Theorem 2.6 gives us a good understanding of the structure of the solution u. In particular it says that the boundary layer along $x = 1$ is, for each fixed value of t, of the same form as we encountered in Part I; compare (2.12) and the term $v_0((1-x)/\varepsilon)$ of (I.1.6).

The theorem also indicates the effect of any discontinuity in the initial-boundary data on $S_x \cup S_0$. For then the reduced solution u_0 will clearly be discontinuous along the characteristic curve of (2.5a) that passes through the point of discontinuity in $S_x \cup S_0$. But $u \in C^1(Q)$ and, away from $x = 1$, we have $|u - u_0| \leq C\sqrt{\varepsilon}$ by (2.6). Hence u must have an interior layer that lies along this characteristic curve.

The characteristic curves of (2.5a) appear frequently in our exploration of numerical methods for (1.1). We shall in future refer to these curves as the *subcharacteristics* of (1.1a).

Remark 2.7. Suppose that for homogeneous initial-boundary data the compatibility conditions $0 = f(0,0) = f_x(0,0) = f_t(0,0)$ are satisfied, in addition to the hypotheses assumed in Theorem 2.6. Then the bound on w in Theorem 2.6 can be strengthened to $|w| \leq C\varepsilon$; see [Bob67].   ♣

In several papers [Shi96a, Shi01, GS04] Shih constructs asymptotic expansions for the solutions of problems related to (1.1), and in [Shi07a] discusses an asymptotic expansion for a pure initial-value problem with a discontinuity in the initial data. Hirsch [Hir90, Section 22.4] also considers this initial-value problem, assuming b is constant and $d \equiv f \equiv 0$. His initial conditions are $u \equiv u_1$ for $x > 0$ and $u \equiv u_2$ for $x < 0$, where the u_i are constants, with the boundary conditions $u \equiv u_1$ as $x \to \infty$ and $u \equiv u_2$ as $x \to -\infty$. Then the exact solution is

$$u(x,t) = \frac{u_2 + u_1}{2} - \frac{u_2 - u_1}{2} \operatorname{erfc}\left(\frac{x - bt}{2\sqrt{\varepsilon t}}\right),$$

where $\operatorname{erfc}(\cdot)$ is the usual complementary error function. From this explicit formula and standard properties of $\operatorname{erfc}(\cdot)$, one can see that u has an interior layer as described above.

Remark 2.8. Suppose that in (1.1) we have homogeneous initial-boundary data. Assume that b, d and f are sufficiently smooth and that the following compatibility conditions are satisfied: $f(1,0) = 0$ and

$$\left|\frac{\partial^{k+m} f(0,0)}{\partial x^k \partial t^m}\right| = 0 \quad \text{for } k + 2m \leq 3.$$

Then one can obtain pointwise estimates for low-order derivatives of u, viz.,

$$\left|\frac{\partial^{k+m} u(x,t)}{\partial x^k \partial t^m}\right| \leq C(1 + \varepsilon^{-k} e^{-\beta(1-x)/\varepsilon}), \tag{2.20}$$

for $(x,t) \in Q$, $k = 0,1$ and $k + m \leq 2$. Details of this work are in [SO89]. Inequality (2.20) implies that no interior layer is present. By making further compatibility assumptions and differentiating (1.1a) with respect to t one or more times, the bounds of (2.20) are easily extended to larger values of k and m.   ♣

Remark 2.9. In Section I.1.1.3 an S-decomposition of the solution of a two-point boundary value problem was constructed. An analogue for (1.1) appears in [Shi92b, p.221] and [Shi]: assume that the solution u has no interior layers (equivalently, assume that a sufficient number of compatibility conditions are

satisfied at the corners $(0,0)$ and $(1,0)$). Then $u(x,t) = U(x,t) + V(x,t)$ for $(x,t) \in Q$, with $LU = f$, $LV = 0$,

$$\left| \frac{\partial^{k+m} U(x,t)}{\partial x^k \partial t^m} \right| \leq C \quad \text{and} \quad \left| \frac{\partial^{k+m} V(x,t)}{\partial x^k \partial t^m} \right| \leq C\varepsilon^{-k} e^{-\beta(1-x)/\varepsilon}, \qquad (2.21)$$

for $0 \leq k, m \leq 3$ and some constant C.

Starting from (2.20), one can deduce the existence of a decomposition $u = U + V$ satisfying (2.21) by means of the same argument that was presented for two-point boundary value problems in Section I.1.1.3, but the properties $LU = f$ and $LV = 0$ are then not guaranteed. ♣

Example 2.10. Suppose that the problem (1.1) were posed on $[0,1] \times [0,\infty)$ instead of $[0,1] \times [0,T]$. Suppose also that the data b, d and f of the problem are constants and that the $q_i(t)$ are continuous with $\lim_{t\to\infty} q_i(t) = \bar{q}_i$ for $i = 1, 2$, where the $\bar{q}_i$ are constants. What happens to the solution $u(x,t)$ as $t \to \infty$?

Let $\bar{u}(x)$ denote the solution of the two-point boundary value problem

$$-\varepsilon \bar{u}'' + b\bar{u}' + d\bar{u} = f \quad \text{for } 0 < x < 1,$$
$$\bar{u}(0) = \bar{q}_0 \quad \text{and} \quad \bar{u}(1) = \bar{q}_1.$$

Problems of this type are quite familiar from Part I.

Then, by a repeated use of comparison principles, it is not difficult to show that $\max_{0 \leq x \leq 1} |u(x,t) - \bar{u}(x)|$ is arbitrarily small for all sufficiently large t. That is, $\bar{u}$ is the *steady-state* solution of (1.1). ♣

In [VR90] the authors give an asymptotic expansion for a problem like (1.1) except that the boundary condition at $x = 0$ is replaced by a decay condition as $t \to \infty$ and one has initial data for $u(x,0)$ for $-\infty < x < 1$.

Remark 2.11. In the case of the *reaction-diffusion problem* obtained by taking $b \equiv 0$ in (1.1), the asymptotic nature of the solution u is quite different from the convection-diffusion case: now u will have boundary layers along both $x = 0$ and $x = 1$. Each of these layers contributes a leading term to the asymptotic expansion of u that is the solution w of the parabolic partial differential equation $Lw = 0$ on Q, subject to certain boundary conditions. Thus they are called *parabolic boundary layers*. They have a more complicated asymptotic structure than the exponential boundary layers we have met up to now, and have no analogue in the solutions of ordinary differential equations. We discuss this type of layer in more detail in Part III.

In the reaction-diffusion case the reduced problem is defined by formally setting $\varepsilon = 0$ in (1.1a) and using only the boundary data (1.1b). Parabolic boundary layers arise where the flow (i.e., the direction of propagation of the reduced problem) is parallel to the boundaries $x = 0$ and $x = 1$ of Q, which are also subcharacteristics of (1.1a). For this reason, parabolic boundary layers are also known as *characteristic boundary layers*.

An S-decomposition of the solution u of a reaction-diffusion problem is given in [HSS00, Appendix A.1]. After making precise assumptions about the regularity of the data and the compatibility conditions that are satisfied at the corners $(0,0)$ and $(1,0)$ of Q, it is shown that

$$u(x,t) = U(x,t) + W_1(x,t) + W_2(x,t) \quad \text{for } (x,t) \in Q,$$

with $LU = f$, $LW_1 = 0$, $LW_2 = 0$ and

$$\left| \frac{\partial^{k+m} U(x,t)}{\partial x^k \partial t^m} \right| \leq C, \tag{2.22a}$$

$$\left| \frac{\partial^{k+m} W_1(x,t)}{\partial x^k \partial t^m} \right| \leq C\varepsilon^{-k/2} e^{-\beta x/\sqrt{\varepsilon}}, \tag{2.22b}$$

$$\left| \frac{\partial^{k+m} W_2(x,t)}{\partial x^k \partial t^m} \right| \leq C\varepsilon^{-k/2} e^{-\beta(1-x)/\sqrt{\varepsilon}}, \tag{2.22c}$$

for some constant C and values of k and m that depend on the compatibility assumptions. Note how each extra x-derivative introduces a factor $\varepsilon^{-1/2}$ into (2.22b) and (2.22c), unlike the corresponding factor ε^{-1} that appears in (2.21) for convection-diffusion problems.   ♣

3

Finite Difference Methods

3.1 First-Order Problems

Assume that $b(x,t) > \beta > 0$ on $\bar{Q}$. Then the solution u of (1.1) has in general a boundary layer along the side $x = 1$ of $\bar{Q}$. Away from this layer and from any interior layers – i.e., on almost all of Q – one expects that u is very close to the solution u_0 of the reduced problem

$$L^0 u_0 := (u_0)_t + b(u_0)_x + d u_0 = f \quad \text{on } Q, \tag{3.1a}$$

$$u_0(x,0) = s(x) \quad \text{on } S_x, \tag{3.1b}$$

$$u_0(0,t) = q_0(t) \quad \text{on } S_0. \tag{3.1c}$$

Consequently any method that computes an accurate numerical approximation of u must be closely related to a method that yields an accurate numerical solution of the first-order problem (3.1). We therefore begin our investigation by considering finite difference methods that are applicable to (3.1).

Several concepts that are frequently used in the analysis of finite difference methods for first-order problems are presented and explained below. For a more complete discussion see [Str04].

Let M and N be positive integers. When working on $\bar{Q}$, we use a rectangular grid $Q_{h,\tau}$ whose nodes are (x_i, t_j) for $i = 0, \ldots, M$ and $j = 0, \ldots, N$. Here $0 = x_0 < x_1 < \cdots < x_M = 1$ and $0 = t_0 < t_1 < \cdots < t_N = T$. Such grids are also known as tensor-product grids. Given any function v that is defined on the grid, set $v_i^j = v(x_i, t_j)$.

For simplicity throughout Section 3.1 only equidistant grids are considered, viz., $x_i - x_{i-1} = h$ (say) for $i = 1, \ldots, M$ and $t_j - t_{j-1} = \tau$ (say) for $j = 1, \ldots, N$.

3.1.1 Consistency

A basic theoretical concept is *consistency*. Essentially, a scheme for (3.1) is consistent if it provides a good approximation to (3.1) when the mesh is

sufficiently fine. This does *not* imply that the solution of the scheme must be a good approximation of the solution of (3.1).

To state this idea more precisely, let the scheme be $L^0_{h,\tau} u_{h,\tau} = \tilde{f}$, where $u_{h,\tau}$ and $\tilde{f}$ are column vectors and $L^0_{h,\tau}$ is a matrix. The matrix $L^0_{h,\tau}$ approximates L^0 by means of difference quotients. The vector $\tilde{f}$ approximates the values of f at the gridpoints (x_i, t_j). The grid function $u_{h,\tau}$ interpolates to $s(x)$ on S_x and to q_0 on S_0; this initial-boundary information may also be incorporated into $L^0_{h,\tau} u_{h,\tau} = \tilde{f}$.

When discussing the solution $u(x,t)$ we set $u^j_i := u_{h,\tau}(x_i, t_j)$ for all i and j. Discrete operators such as $L^0_{h,\tau}$ are often applied to functions v defined on all of $\bar{Q}$ by first restricting v to $\{v^j_i\}$. To describe this precisely one should introduce a restriction operator R_h as in Section I.2.1.1, but we do not bother with this here and simply write $L^0_{h,\tau} v$.

If $w \in C(\bar{Q})$ or w is defined on the grid $Q_{h,\tau}$, then define the discrete maximum norm of w by $\|w\|_{\infty,d} = \max\{|w(x_i, t_j)| : (x_i, t_j) \in Q_{h,\tau}\}$.

Definition 3.1. *Similarly to Part I, the scheme $L^0_{h,\tau} u_{h,\tau} = \tilde{f}$ is said to be consistent with (3.1) with respect to $\|\cdot\|_{\infty,d}$ if one has*

$$\|L^0 u_0 - L^0_{h,\tau} u_0\|_{\infty,d} + \|f - \tilde{f}\|_{\infty,d} \to 0 \quad \text{as } h, \tau \to 0.$$

In general, consistency is easy to check by a Taylor expansion of the functions u_0 and f about each mesh point (x_i, t_j). If a scheme is consistent, it does not automatically follow that it yields an accurate numerical approximation to the solution of (3.1). The next example illustrates this important point and introduces a scheme that we shall encounter frequently later.

Example 3.2. Take $b \equiv 1$ and $d \equiv f \equiv 0$ in (3.1a), so the differential equation becomes $(u_0)_t + (u_0)_x = 0$. Let the initial-boundary data be $u_0(x,0) = 0$ on S_x and $u_0(0,t) = t^3$ on S_0. The solution to this problem is

$$u_0(x,t) = \begin{cases} (t-x)^3 & \text{if } t \geq x, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

for all $(x,t) \in \bar{Q}$.

Approximate this differential equation by the scheme

$$(L^0_{h,\tau} u_{h,\tau})^j_i := \frac{u^j_i - u^j_{i-1}}{h} + \frac{u^{j+1}_i - u^j_i}{\tau} = 0$$

for $i = 1, \ldots, M$ and $j = 0, \ldots, N-1$. The discrete solution $u_{h,\tau}$ interpolates to the initial-boundary data of the original problem, i.e., we set $u^0_i = 0$ for all i and $u^j_0 = t^3_j$ for all j.

This scheme is commonly called the *simple upwind scheme*. It is straightforward to verify that it is consistent. One can rewrite it as

$$u_i^{j+1} = (1 - \tau/h)u_i^j + (\tau/h)u_{i-1}^j, \tag{3.3}$$

which is a more convenient form for computation.

Since $u_i^0 = 0$ for all i (from the initial data), taking $j = 0$ in (3.3) yields $u_i^1 = 0$ for $i = 1, \dots, M$. Repeating this argument for $j = 1, \dots, N - 1$, we conclude that the computed solution $u_{h,\tau}$ satisfies $u_i^j = 0$ for $i \geq j$. That is, $u_{h,\tau} = 0$ at those nodes (x_i, t_j) that lie in $\{(x, t) \in \bar{Q} : \tau x/h \geq t\}$. If, say, $\tau = 2h$, then the computed solution will be zero at all nodes of $Q_{h,\tau}$ that lie below the line $t = 2x$. This is clearly a very poor approximation to u_0 on much of $\bar{Q}$. ♣

3.1.2 Stability

We have just seen that consistency alone fails to ensure that the solution of a difference scheme is accurate. A more subtle additional attribute called *stability* is needed to guarantee accuracy. It essentially says that the discrete operator $L_{h,\tau}^0$ has an inverse that is bounded, uniformly in h and τ, in some norm, but the material below is not presented from this viewpoint.

One can define various forms of stability; see [GKO95, RM94] for a discussion of the issues involved. Here we confine our attention to the most standard form.

Suppose that the difference scheme, including boundary conditions, can be written in matrix-vector form as

$$u^j = Au^{j-1} + w^{j-1} \quad \text{for } j = 1, \dots, N, \tag{3.4}$$

where $u^j := (u_0^j, \dots, u_M^j)^T$, A is an $(M + 1) \times (M + 1)$ matrix, and w^{j-1} is a vector that depends only on f, the initial-boundary data, and the mesh. Consequently

$$u^j = A^j u^0 + \sum_{k=1}^{j} A^{j-k} w^{k-1}. \tag{3.5}$$

In the case where $q_0 \equiv f \equiv 0$, the solution $u_0(\cdot, t)$ of (3.1) will be bounded in $L_p(0, 1)$ for each $t \in [0, T]$ and $1 \leq p \leq \infty$.

In general, a difference scheme is stable if its solution has a certain boundedness property that is already enjoyed by the solution of the differential equation that it models; thus we are led by (3.5) to the next definition.

Definition 3.3. *The scheme (3.4) is stable with respect to the discrete L_p norm $\|\cdot\|_{p,d}$ over a family $\mathcal{F}$ of meshes if*

$$\|A^j\|_{p,d} \leq K \text{ for all } j, \tag{3.6}$$

where K is some constant that is fixed for all meshes in $\mathcal{F}$ and $\|A^j\|_{p,d}$ is the matrix norm induced on A^j by the discrete L_p norm of $v \in \mathbb{R}^{M+1}$, which is defined by

$$\|v\|_{p,d} = \begin{cases} \left(h \sum_{i=0}^{K} |v_i|^p\right)^{1/p} & \text{when } 1 \le p < \infty, \\ \max_i |v_i| & \text{when } p = \infty. \end{cases}$$

This definition is known as the *matrix criterion for stability analysis*. The family $\mathcal{F}$ of meshes may be defined, for instance, as those meshes satisfying $\tau \le h \le h_0$ where the constant h_0 depends on the data b and d. We emphasize that the constant K in (3.6) must be independent of the mesh; an examination of the eigenvalues of A may show that for each fixed mesh $\|A^j\|_{2,d}$ is bounded for all j, but this does not prove (3.6). For further discussion of this point, which is sometimes incorrectly presented in the literature, see [Str04, Section 11.5].

Setting $q_0 \equiv f \equiv 0$, as we did earlier, means that $w^{j-1} = 0$ in (3.4) except for a finite number of j. Then stability implies that $\|u^j\|_{p,d}$ is bounded for all j, just as the solution of the differential equation was bounded.

In principle, the matrix criterion provides necessary and sufficient conditions for the stability, with respect to any vector norm, of any scheme of the form (3.4). In practice, however, it is often quite difficult to obtain a satisfactory analysis of (3.6). For an example see [Str04, Section 11.5]. This drawback prompts us to use the simpler L_2-stability test that we now describe.

For first-order problems like (3.1) and parabolic problems such as (1.1), one can decompose the stability analysis into two separate parts. One part examines the difference scheme when it is applied to a pure initial-value problem, with initial data on the entire x-axis. The other analyses the interaction between the scheme and the boundary condition(s) imposed in the scheme. This separation is not necessary for a theoretical investigation – recall the matrix criterion above – but it is frequently convenient.

We shall describe the analysis of the initial-value problem but refer to [GKO95, Str04] for theory and examples that deal with the boundary conditions, because a full description of stability for initial-boundary value problems is technical and lengthy.

Suppose therefore that (3.1) is altered to a pure initial-value problem. That is, (3.1a) becomes

$$w_t + bw_x + dw = f \quad \text{on } \hat{Q} := (-\infty, \infty) \times [0, T] \tag{3.7}$$

and (3.1b) and (3.1c) are replaced by $w(x, 0) = \hat{s}(x)$ on $(-\infty, \infty)$, where the function $\hat{s}$ is given. The functions b, d and f are taken to be defined on $\hat{Q}$. Assume that $\hat{s}, b, d$ and f are continuous and bounded on $\hat{Q}$, with $b(x, t) > \beta > 0$ and $d(x, t) \ge \gamma > 0$. Also assume that $\hat{s}(x) \to 0$ as $|x| \to \infty$, with $\hat{s} \in L_2(-\infty, \infty)$.

The above assumptions are designed to remove the effects of boundary conditions, but otherwise alter the original problem (3.1) as little as possible.

The following stability analysis assumes that $f \equiv 0$. This is not a restriction. It turns out [Str04, Section 9.3] that the stability of a scheme depends only on the difference approximation to L^0 and on the boundary conditions present.

Place an equidistant tensor-product mesh (x_i, t_j) on $\hat{Q}$, where $x_0 = 0$, $x_i = ih$ for $-\infty < i < \infty$, and $t_0 = 0$, $t_j = j\tau$ for $j = 0, \ldots N$. The difference scheme "matrix" (it is now infinite-dimensional) is still written as $L^0_{h,\tau}$.

Assume that $f \equiv 0$. Then the scheme is: $L^0_{h,\tau} u^j_i = 0$ for $-\infty < i < \infty$ and $j = 0, \ldots N - 1$. This scheme is L_2 *stable* if and only if there exists a fixed non-negative integer J and a constant K such that for each $j \in \{0, \ldots, N\}$,

$$\|u_{h,\tau}(\cdot, t_j)\|^2_{2,d} := h \sum_{i=-\infty}^{\infty} |u^j_i|^2 \leq Kh \sum_{k=0}^{J} \sum_{i=-\infty}^{\infty} |u^k_i|^2 = K \sum_{k=0}^{J} \|u_{h,\tau}(\cdot, t_k)\|^2_{2,d}$$

$$(3.8)$$

for all h and τ sufficiently small (with perhaps some restriction on the relative sizes of h and τ such as $\tau \leq 2h$).

The left-hand side of (3.8) is the square of the usual discrete $L_2(-\infty, \infty)$ norm, while the right-hand side is a sum of $J+1$ such squares. The inequality is reasonable: it says that the discrete L_2 norm of the solution at any time level is bounded by a constant times a discrete L_2 norm near $t = 0$.

The value of J in (3.8) depends on how the scheme makes use of the initial data, as we now describe.

A *one-step scheme* is a difference scheme where the computation of u^{j+1}_i for each i and j does not depend on $u^n_{(\cdot)}$ for any $n < j$. The simple upwind scheme of Example 3.2 is a one-step scheme. A scheme such as

$$\frac{u^{j+1}_i - u^{j-1}_i}{2\tau} + \frac{u^j_{i+1} - u^j_{i-1}}{2h} = 0$$

(which is consistent with the equation $w_t + w_x = 0$) is not a one-step scheme. Schemes that are not one-step are called *multi-step schemes*.

For L_2-stable one-step schemes, one always has $J = 0$. For multi-step schemes, unlike one-step schemes, it is obvious that we need initial data on more than one time level in order to commence iterating. This extra initial data may come from the original problem (in which case one must take $J > 0$ in (3.8) in order to include all externally supplied initial data in the right-hand side), or it may be generated by using a one-step method that needs only the initial data from $t = 0$ (and one then takes $J = 0$).

Example 3.4. Consider the L_2 stability of the simple upwind scheme

$$(L^0_{h,\tau} u)^j_i := \frac{u^{j+1}_i - u^j_i}{\tau} + b\frac{u^j_i - u^j_{i-1}}{h} = 0,$$

for $-\infty < i < \infty$ and $j = 0, \ldots, N - 1$. This scheme is consistent with the differential equation $w_t + bw_x = 0$, where b is constant.

For each $j \geq 1$, the inequality $2|u^j_i u^j_{i-1}| \leq |u^j_i|^2 + |u^j_{i-1}|^2$ yields

$$\sum_{i=-\infty}^{\infty} |u_i^{j+1}|^2 = \sum_{i=-\infty}^{\infty} |(1 - b\tau/h)u_i^j + (b\tau/h)u_{i-1}^j|^2$$

$$\leq \sum_{i=-\infty}^{\infty} \{[(1 - b\tau/h)^2 + (b\tau/h)|1 - b\tau/h|]|u_i^j|^2$$

$$+ [(b\tau/h)^2 + (b\tau/h)|1 - b\tau/h|]|u_{i-1}^j|^2\}$$

$$= \sum_{i=-\infty}^{\infty} [|1 - b\tau/h| + (b\tau/h)]^2 |u_i^j|^2.$$

If $|1 - b\tau/h| + (b\tau/h) \leq 1$, then

$$\sum_{i=-\infty}^{\infty} |u_i^{j+1}|^2 \leq \sum_{i=-\infty}^{\infty} |u_i^j|^2 \leq \sum_{i=-\infty}^{\infty} |u_i^{j-1}|^2 \leq \cdots \leq \sum_{i=-\infty}^{\infty} |u_i^0|^2,$$

i.e., the scheme is L_2 stable. The sufficient condition $|1 - b\tau/h| + (b\tau/h) \leq 1$ is equivalent to the inequality $b\tau \leq h$.

When this sufficient condition is violated, it does not follow from our calculation that the scheme is L_2 unstable. Nevertheless Example 3.2 shows that when $b\tau > h$ the computed solution is unsatisfactory; this poor behaviour will be examined from another viewpoint after Theorem 3.7. ♣

To prove L_2 stability sometimes entails algebraic manipulations that are more ingenious than those of Example 3.4; see [GKO95, RM94].

If the coefficients in the scheme are variable (as will usually be the case when either b or d is not constant), then to prove L_2 stability one begins by "freezing" each coefficient. This means that one replaces each variable coefficient by its value at some arbitrary point in the domain of definition of the differential equation. Thus the variable coefficient scheme is replaced by one with constant coefficients, but these constant coefficients are not known precisely; one can say only that they lie in the range of values of the original coefficients. If one can prove that this frozen coefficient scheme is L_2 stable, then under certain conditions it follows that the original variable coefficient scheme is also L_2 stable. For a discussion of this topic see [Str04].

3.1.3 Convergence in L_2

Now that consistency and stability have been defined, we can address the issue of deciding which schemes yield "good" approximations of the solution of (3.7). (Strictly speaking consistency was defined only for (3.1) and for the norm $\|\cdot\|_{\infty,d}$, but it is easy to see how to adapt this definition to fit (3.7) and $\|\cdot\|_{2,d}$.) We again work in a framework of L_2 norms.

Definition 3.5. *(Convergence in the discrete L_2 norm) Let w be the solution of (3.7) with $w(x,0) = \hat{s}(x)$ on $(-\infty, \infty)$. Let $u_{h,\tau}$ be the solution of the*

scheme $L_{h,\tau}^0 u_{h,\tau} = \tilde{f}$ with some initial conditions, where all meshes considered come from some family $\mathcal{F}$. We say that $u_{h,\tau}$ converges to w if

$$\max_{t_j \in [0,T]} h \sum_{i=-\infty}^{\infty} |w(x_i, t_j) - u_i^j|^2 \to 0 \quad \text{as } h, \tau \to 0,$$

where in this limit we consider only values of τ such that t/τ is an integer. ♣

All of these concepts come together in the following celebrated result, which is proved in, e.g., [GKO95, Str04].

Theorem 3.6 (Lax-Richtmyer theorem). *An L_2-consistent finite difference scheme for (3.7) is L_2-convergent for a family of meshes $\mathcal{F}$ if and only if it is L_2 stable for $\mathcal{F}$.*

This theorem tells us that we should concentrate on schemes that are both consistent and stable. In general schemes that seem intuitively to be reasonable approximations of (3.7) are consistent. It is less obvious which schemes are L_2 stable; one must carefully verify the condition of the definition. We now give a simple necessary condition for L_2 stability that enables us to exclude many plausible but inaccurate schemes from consideration.

A difference scheme for (3.7), with $f = 0$, is said to be *explicit* if for each i and j it can be written in the form

$$u_i^j = \sum \alpha_{(\cdot)}^n u_{(\cdot)}^n$$

where the $\alpha_{(\cdot)}^n$ depend only on b, d and the grid, the sum has a fixed finite number of terms and each n satisfies $n < j$. That is, each u_i^j can easily be calculated from the previously computed solution at earlier time levels without having to solve a linear system of equations.

Theorem 3.7. *Consider*

$$w_t + bw_x + dw = 0 \quad \text{on } \hat{Q}, \tag{3.9}$$

with initial data on the x-axis. Assume that b is a positive constant. Approximate (3.9) on an equidistant tensor-product grid by the explicit one-step scheme

$$u_i^j = \alpha_{i-1}^{j-1} u_{i-1}^{j-1} + \alpha_i^{j-1} u_i^{j-1} + \alpha_{i+1}^{j-1} u_{i+1}^{j-1},$$

where the coefficients α are constants depending on h, τ, b and d. Assume that this scheme is consistent with (3.9). Then a necessary condition for L_2 stability is the Courant-Friedrichs-Lewy (CFL) condition

$$\frac{b\tau}{h} \le 1.$$

Proof. The form of the scheme implies that u_i^j is computed using only those initial values u_k^0 for which $i - j \leq k \leq i + j$. That is, the only data used from the x-axis lies in the interval $[x_i - jh, x_i + jh]$.

Suppose that the scheme violates the CFL condition. Then $[x_i - jh, x_i + jh]$ is contained in the interior of the interval $[x_i - bjk, x_i + bjk]$.

Now the characteristic curve of (3.9) that passes through (x_i, t_j) intersects the x-axis at the point $x = x_i - bjk$. Thus the value of w_i^j depends on the initial data at $(x_i - bjk, 0)$, but we have just seen that data from this point is not used to compute u_i^j. As (x_i, t_j) was an arbitrary mesh point, we infer that $u_{h,\tau}$ cannot in general converge to w. The Lax-Richtmyer theorem now implies that the scheme is not L_2 stable.   □

The quantity $b\tau/h$ is called the *Courant number*. Theorem 3.7 can easily be modified so that its argument and conclusion apply to any explicit scheme.

The CFL condition is extremely useful because of its simplicity and widespread applicability. For instance, it complements the analysis of Example 3.4, where the simple upwind scheme was shown to be stable when $b\tau \leq h$. The CFL condition implies that the scheme is unstable when $b\tau > h$, which agrees with the conclusion of Example 3.2.

Explicit schemes have the desirable property that solutions can be computed cheaply as one moves from each time level to the next. Nevertheless, for any explicit scheme, the CFL condition places a restriction on the maximum permissible time step. Thus if the mesh in the x-direction is fine (a situation that might arise at a boundary layer), then large time steps will not be permitted and consequently an excess of computational effort may be needed to reach $t = T$.

All the schemes we have seen so far are explicit, but *implicit* (i.e., non-explicit) schemes are quite common. More work per time step is needed when one uses an implicit scheme. This may be offset by the fact that implicit schemes generally have much less restrictive conditions (or perhaps none at all) on the maximum permissible value of τ; see, e.g., Example 3.8.

Theorem 3.7 cannot be applied to implicit schemes. We now describe an alternative method for L_2-stability analysis, devised by von Neumann and based on Fourier transforms, which can be used with any one-step or multi-step scheme.

Given a constant coefficient difference scheme for (3.7), where the mesh is equidistant, replace u_m^j by $\xi^j e^{\imath m\theta}$ for each m and j, and set $f \equiv 0$. Here $\imath = \sqrt{-1}$, $\xi \in \mathbb{C}$ and $\theta \in \mathbb{R}$. Solve this equation for the *amplification factor* ξ. Then the scheme is L_2 stable if and only if the *von Neumann condition*

$$|\xi(\theta, h, \tau)| \leq 1 + K\tau$$

holds true for all θ and all sufficiently small h and τ, where K is some fixed positive constant. If $d \equiv 0$ in (3.7), then [RM94] the above von Neumann condition should be replaced by

$$|\xi(\theta, h, \tau)| \leq 1.$$

Example 3.8. Here, for the first time, we meet one of the most commonly used schemes in the literature. It is often called the *(Keller) box scheme*, since it is analysed in [Kel71] and is derived by integrating the differential equation over each rectangular box formed by the grid, but the earliest description of the scheme appears to be that of Wendroff [Wen60] and indeed it also travels under the name of *Wendroff's implicit scheme.*

The box scheme for (3.7), when b is constant and $d \equiv 0$, is

$$\frac{b}{2h}[(u_{m+1}^j + u_{m+1}^{j+1}) - (u_m^j + u_m^{j+1})] + \frac{1}{2\tau}[(u_m^{j+1} + u_{m+1}^{j+1}) - (u_m^j + u_{m+1}^j)]$$
$$= f_m^j, \tag{3.10}$$

for $-\infty < m < \infty$ and $j = 0, \ldots, M - 1$.

For a pure initial-value problem such as (3.7) this scheme is implicit, but for the initial-boundary value problem (3.1) one can sequentially compute $u_1^{j+1}, u_2^{j+1}, \ldots, u_M^{j+1}$ at successive time levels t_{j+1} without solving linear systems of equations.

It is easy to check that (3.10) is consistent with (3.7). Following the von Neumann procedure with $f \equiv 0$, a calculation yields

$$\xi(\theta, h, \tau) = \frac{1 + \nu + (1 - \nu)e^{i\theta}}{1 - \nu + (1 + \nu)e^{i\theta}},$$

where we have set $\nu = b\tau/h$. It then follows that $|\xi(\theta, h, \tau)| = 1$ for all values of θ, h and τ. That is, the box scheme satisfies the von Neumann condition on every equidistant mesh. ♣

Example 3.9. If one applies the von Neumann analysis to the simple upwind scheme of Example 3.4, then $\xi = 1 - \nu(1 - e^{-i\theta})$ with $\nu = b\tau/h$, so $|\xi|^2 = 1 + 2\nu(\nu - 1)(1 - \cos\theta)$. Consequently the von Neumann condition is satisfied if and only if $\nu \leq 1$. This conclusion resembles our earlier stability results for this scheme. ♣

Further examples of the von Neumann stability analysis can be found in [Hir88, Str04]. Like the CFL condition, the von Neumann condition is *a necessary but not a sufficient* L_2-stability condition for the initial-boundary value problem (3.1). In practice, however, this pair of necessary conditions – when taken together – often turn out to be also sufficient for L_2 stability.

3.2 Convection-Diffusion Problems

Consider once again the parabolic convection-diffusion problem (1.1):

$$u_t(x,t) - \varepsilon u_{xx}(x,t) + b(x,t)u_x(x,t) + d(x,t)u(x,t) = f(x,t)$$

where $(x,t) \in Q := (0,1) \times (0,T]$, and

$$u(x,0) = s(x) \quad \text{on } S_x := \{(x,0) : 0 \le x \le 1\},$$
$$u(0,t) = q_0(t) \quad \text{on } S_0 := \{(0,t) : 0 < t \le T\},$$
$$u(1,t) = q_1(t) \quad \text{on } S_1 := \{(1,t) : 0 < t \le T\}.$$

Assume that $d(x,t) \ge \gamma > 0$ and $b(x,t) > \beta > 0$.

The ideas and techniques that we encountered in Section 3.1 will apply (for the most part) to finite difference schemes for (1.1) after making some minor changes in notation. Again place the equidistant rectangular grid

$$Q_{h,\tau} := \{(x_i, t_j) : i = 0, \dots, M \text{ and } j = 0, \dots, N\}$$

on $\bar{Q}$, with $h = x_i - x_{i-1}$ for all i and $\tau = t_j - t_{j-1}$ for all j. The difference scheme is written as $L_{h,\tau} u_{h,\tau} = \tilde{f}$ on $Q_{h,\tau}$, where $u_{h,\tau}$ interpolates to the initial-boundary data. As before, u_i^j stands for $u_{h,\tau}(x_i, t_j)$.

3.2.1 Consistency and Stability

Assume that all our meshes come from some family $\mathcal{F}$. Consistency of the scheme $L_{h,\tau} u_{h,\tau} = \tilde{f}$ with respect to the discrete maximum norm is defined analogously to Definition 3.1.

The scheme $L_{h,\tau} u_{h,\tau} = 0$ is L_2 *stable* on $\mathcal{F}$ with respect to the initial-boundary data if there exists a constant K such that for each $j \in \{0, \dots, N\}$,

$$\|u_{h,\tau}(\cdot, t_j)\|_{2,d}^2 := h \sum_{i=0}^{M} |u_i^j|^2$$

$$\le K \left[h \sum_{i=0}^{M} |u_i^0|^2 + \tau \sum_{k=1}^{j} (|u_0^k|^2 + |u_M^k|^2) \right]$$

$$= K \left[\|s\|_{2,d}^2 + \|q_0\|_{2,d;[0,t_j]}^2 + \|q_1\|_{2,d;[0,t_j]}^2 \right] \qquad (3.11)$$

for all h and τ sufficiently small, where $\|q_k\|_{2,d;[0,t_j]}$ denotes the discrete L_2 norm of $q_k(t)|_{t \in [0,t_j]}$.

The von Neumann condition of Section 3.1.3 can be applied to any difference scheme for (1.1). This gives a necessary condition for L_2 stability. One can also use the matrix criterion to get necessary and sufficient conditions for L_2 and L_∞ stability.

Example 3.10. Hirsch [Hir88, Section 10.4] compares these two approaches in giving a detailed L_2-stability analysis of the scheme

$$\frac{u_i^{j+1} - u_i^j}{\tau} - \varepsilon \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + b \frac{u_{i+1}^j - u_{i-1}^j}{2h} = 0, \qquad (3.12)$$

which uses central differences in space to approximate (1.1a). Here b is constant and $d \equiv f \equiv 0$. In particular he shows from a von Neumann analysis that

$$\tau \leq \frac{2\varepsilon}{b^2} \quad \text{and} \quad \frac{\tau}{h^2} \leq \frac{1}{2\varepsilon}$$

are necessary for L_2 stability. Thus the scheme has little practical value because the condition $\tau \leq 2\varepsilon/b^2$ is very restrictive. It should be noted that for the *cell Reynolds number* bh/ε, the L_2-stability analysis requires only the upper bound $2h/(b\tau)$. Thus even for large cell Reynolds numbers one can achieve L_2 stability for nonupwinded schemes for time-dependent problems. Nevertheless a simple barrier function argument based on Theorem 2.2 shows that

$$\max_{(x,t)\in\bar{Q}} |u(x,t)| \leq \max\{|u(x,t)| : (x,t) \in S_x \cup S_0 \cup S_1\},$$

and requiring that the discrete solution obey the same bound (this is the L_∞-stability condition (3.15) with $K = 1$) forces the cell Reynolds number to be bounded by 1 – see [Str04, Section 6.4].   ♣

Example 3.11. Suppose that we attempt to stabilize the central difference in space scheme by taking a backward difference in time. That is, approximate (1.1a), with $b \equiv r \equiv 1$ and $d \equiv f \equiv 0$ for convenience, by

$$\frac{u_i^j - u_i^{j-1}}{\tau} - \varepsilon \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + \frac{u_{i+1}^j - u_{i-1}^j}{2h} = 0, \qquad (3.13)$$

for $i = 1, \ldots, M-1$ and $j = 1, \ldots, N$. A von Neumann analysis of this scheme shows that it is L_2 stable for all choices of h and τ. Nevertheless it is in general an unsatisfactory scheme for (1.1), as we now show heuristically.

Suppose that the data of the problem are such that the solution u has no interior layers but a boundary layer at $x = 1$. Assume that ε is much smaller than h and τ, as is usually the case in practice. Then all interior nodes lie outside the boundary layer and the ε/h^2 term in (3.13) can be ignored. Assume also that $h \leq \tau$.

Take $i = M - 1$ in (3.13) for each j, giving

$$u_{M-2}^j = u_M^j + \frac{2h(u_{M-1}^j - u_{M-1}^{j-1})}{\tau}.$$

But if the computed solution is accurate outside the layer, then $u_{M-1}^j - u_{M-1}^{j-1}$ is small. It follows that $u_{M-2}^j \approx u_M^j$, which is not true of the analytical solution $u(x,t)$ because of the boundary layer.   ♣

Example 3.11 shows that one must be careful when interpreting the results obtained from an L_2-stability analysis. Discussions of this issue and of the effect of boundary conditions are given in [HGG84, Hir88, Mor80, Str04].

For difference schemes for general parabolic differential equations, the CFL condition for L_2 stability is inapplicable. Nevertheless – recall the first paragraph of Section 3.1 – for a convection-dominated problem such as (1.1), if setting $\varepsilon = 0$ in our difference scheme yields an explicit scheme for the reduced problem, then in practice the CFL condition for this "reduced" scheme is a necessary condition for L_2 stability of the original scheme.

A further way of analysing stability is the following. The difference operator $L_{h,\tau}$ satisfies a *discrete maximum principle* (cf. the discrete comparison principle of Section I.2.1.1) if

$$L_{h,\tau}u_{h,\tau} \le 0 \text{ and } u_{h,\tau}|_{S_x \cup S_0 \cup S_1} \le 0 \text{ together imply that } u_{h,\tau} \le 0 \text{ on } Q_{h,\tau}.$$

By Theorem 2.2 the differential operator satisfies a comparison principle, which is equivalent to a maximum principle. We now describe a discrete analogue of this result. In this lemma, all vector and matrix inequalities hold componentwise.

Lemma 3.12. *Suppose that the difference scheme (excluding initial-boundary conditions) can be written in the form*

$$(L_{h,\tau}u_{h,\tau})^{j+1} := A\hat{u}^{j+1} - Bu^j = w^j \quad \text{for } j = 0, \ldots, N-1, \qquad (3.14)$$

where $u^j = (u_0^j, \ldots, u_M^j)^T$, $\hat{u}^{j+1} = (u_1^{j+1}, \ldots, u_{M-1}^{j+1})^T$, w^j *is a vector that depends only on f and the mesh, and A and B are matrices. Suppose also that A is an M-matrix and $B \ge 0$.*

Let y and z be functions defined on the mesh. Set $y^j = (y_0^j, \ldots, y_M^j)^T$ and $z^j = (z_0^j, \ldots, z_M^j)^T$ for each j. Assume that

$$\left| (L_{h,\tau}y)^{j+1} \right| \le (L_{h,\tau}z)^{j+1} \quad \text{for } j = 0, \ldots, N-1,$$

$$|y| \le z \text{ on } S_x \cup S_0 \cup S_1.$$

Then $|y| \le z$ on $Q_{h,\tau}$.

Proof. Use induction on j to show that $|y^j| \le z^j$ for each j. First, $|y^0| \le z^0$ by hypothesis.

Suppose that $|y^j| \le z^j$ for some $j \in \{0, \ldots, N-1\}$. Then

$$A(\hat{z}^{j+1} - \hat{y}^{j+1}) = B(z^j - y^j) + (L_{h,\tau}z)^{j+1} - (L_{h,\tau}y)^{j+1} \ge 0,$$

since $B \ge 0$, $z^j - y^j \ge 0$ and $(L_{h,\tau}z)^{j+1} \ge (L_{h,\tau}y)^{j+1}$. But A is an M-matrix, so $A^{-1} \ge 0$ and it follows that $\hat{z}^{j+1} - \hat{y}^{j+1} \ge 0$. One can show similarly that $\hat{z}^{j+1} + \hat{y}^{j+1} \ge 0$. Hence $|\hat{y}^{j+1}| \le \hat{z}^{j+1}$. As $|y| \le z$ on $S_0 \cup S_1$, we have $|y^{j+1}| \le z^{j+1}$. This completes the induction and the proof. $\square$

It is easy to see that essentially the same argument demonstrates that, under the hypotheses of the lemma, $L_{h,\tau}$ satisfies a discrete maximum principle.

For all meshes in some family $\mathcal{F}$, the scheme $L_{h,\tau}u_{h,\tau} = \tilde{f}$ is L_∞ *stable* with respect to the data of the problem if there exists a constant K such that

$$\|u_{h,\tau}\|_{\infty,d} \leq K\big[\|\tilde{f}\|_{\infty,d} + \max\{|u_i^j| : (x_i, t_j) \in S_x \cup S_0 \cup S_1\}\big]. \qquad (3.15)$$

If a difference scheme satisfies the hypotheses of Lemma 3.12, then it usually satisfies (3.15). To make this deduction we take $y = u_{h,\tau}$ and $z = K\max\{|u_i^j| : (x_i, t_j) \in S_x \cup S_0 \cup S_1\}$, then try to choose K to satisfy the hypotheses of the lemma. Here for any positive *constant* mesh function ζ one needs to know that $(L_{h,\tau}\zeta)^{j+1} > 0$ for each j. This inequality holds true (for sufficiently small h and τ) for most reasonable schemes that approximate (1.1), since $d > 0$.

The method of Example 3.11 fails to satisfy the hypotheses of Lemma 3.12, because of the central difference used to approximate u_x.

Example 3.13. Take b and d to be constant and $f \equiv 0$ in (1.1). We modify the simple upwind scheme of Example 3.4 by introducing an extra difference quotient to approximate the diffusion term $-\varepsilon u_{xx}$ of (1.1). Thus the *simple upwind scheme* for this convection-diffusion problem is

$$(L_{h,\tau}u_{h,\tau})_i^j := \frac{u_i^{j+1} - u_i^j}{\tau} - \varepsilon\frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + b\frac{u_i^j - u_{i-1}^j}{h} + du_i^j$$
$$= 0,$$

for $i = 1, \ldots, M$ and $j = 0, \ldots, N - 1$. It is a generalization of the scheme (I.2.12) for two-point boundary value problems.

Rewrite the simple upwind scheme as

$$u_i^{j+1} - \left[\left(\frac{b\tau}{h} + \frac{\varepsilon\tau}{h^2}\right)u_{i-1}^j + \left(1 - d\tau - \frac{b\tau}{h} - \frac{2\varepsilon\tau}{h^2}\right)u_i^j + \frac{\varepsilon\tau}{h^2}u_{i+1}^j\right] = 0.$$

This is in the form (3.14), with A an M-matrix and $B \geq 0$, provided that

$$d\tau + \frac{b\tau}{h} + \frac{2\varepsilon\tau}{h^2} \leq 1. \qquad (3.16)$$

Thus the scheme satisfies a discrete maximum principle if (3.16) holds true. Note that for $\varepsilon \leq h^2$ and $\tau \ll 1$, condition (3.16) is almost identical to the CFL condition $b\tau/h \leq 1$ that was shown in Section 3.1.1 to pertain to the simple upwind scheme for the reduced problem. ♣

Example 3.14. One can modify the box scheme (3.10) by adding a term that approximates $-\varepsilon u_{xx}$, analogously to the modification of the simple upwind scheme in Example 3.13. Nevertheless, irrespective of how one chooses this difference approximation of $-\varepsilon u_{xx}$, it is impossible to satisfy the hypotheses of Lemma 3.12. For when ε is very small relative to h and τ, the scheme is essentially (3.10):

$$\frac{b}{2h}[(u_{i+1}^j + u_{i+1}^{j+1}) - (u_i^j + u_i^{j+1})] + \frac{1}{2\tau}[(u_i^{j+1} + u_{i+1}^{j+1}) - (u_i^j + u_{i+1}^j)] = f_i^j,$$

for $i = 0, \ldots, M - 1$ and $j = 0, \ldots, M - 1$ (here b in (3.10) is taken to be constant, with $d \equiv 0$). This is not of the form (3.14), with A an M-matrix and $B \geq 0$, except for the special case when $b\tau = h$. The box scheme does not in general satisfy a discrete maximum principle, and as a consequence its computed solutions often exhibit oscillations. ♣

3.2.2 Convergence

Let u be the solution of (1.1) and $u_{h,\tau} = \{u_i^j\}$ the solution of a difference scheme that approximates (1.1), where all meshes considered come from some family $\mathcal{F}$. We say that $u_{h,\tau}$ *converges to u in the L_p sense* if

$$\max_{t_j \in [0,T]} \|u(\cdot, t_j) - u_{h,\tau}(\cdot, t_j)\|_{p,d} \to 0 \quad \text{as } h, \tau \to 0, \tag{3.17}$$

where for any mesh function $v = (v_0, \ldots, v_M)$ the discrete norms are defined by

$$\|v\|_{p,d} = \begin{cases} \left(h \sum_{i=0}^{M} |v_i|^p \right)^{1/p} & \text{when } 1 \leq p < \infty, \\ \max_i |v_i| & \text{when } p = \infty. \end{cases}$$

This definition generalizes the L_2 convergence property of Section 3.1.3. It is often described as "convergence in $L_\infty(L_p)$", where L_∞ refers to the $\max_{t_j \in [0,T]}$ operator in (3.17), but as our analysis always uses L_∞ in the time variable we can discard this part of that notation.

Suppose that the scheme $L_{h,\tau} u_{h,\tau} = \tilde{f}$ is consistent with (1.1). If the scheme is L_2 stable, then the Lax-Richtmyer theorem shows that (3.17) holds true with $p = 2$. If the scheme is L_∞ stable, then (3.17) holds true with $p = \infty$.

For non-singularly perturbed problems (i.e., problems whose solutions u do not have layers), satisfactory convergence results can be obtained using the arguments of the previous paragraph. One shows that

$$|(Lu - L_{h,\tau} u)_i^j| + |(f - \tilde{f})_i^j| \leq g(h, \tau), \tag{3.18}$$

for some function $g(h, \tau)$ that satisfies $g(h, \tau) \to 0$ as $h, \tau \to 0$. If the scheme is L_2 stable, it follows that (for h and τ sufficiently small)

$$\left\{ h \sum_{i=0}^{M} |u(x_i, t_j) - u_i^j|^2 \right\}^{1/2} \leq Kg(h, \tau) \tag{3.19}$$

for each j, where the value of K depends on the initial-boundary data. If instead the scheme is L_∞ stable, then

$$\max\{|u(x_i, y_j) - u_i^j| : (x_i, t_j) \in Q_{h,\tau}\} \leq Kg(h, \tau). \tag{3.20}$$

For a convection-diffusion problem such as (1.1), however, a sharp analysis usually shows that

$$|(Lu - L_{h,\tau}u)_i^j| + |(f - \tilde{f})_i^j| \le \sigma(\varepsilon, h, \tau), \tag{3.21}$$

where, when ε is fixed, $\sigma(\varepsilon, h, \tau) \to 0$ as $h, \tau \to 0$, but if $\varepsilon \to 0$ with h and τ fixed, then $\sigma(\varepsilon, h, \tau) \to \infty$. As we are interested in robust schemes that work well even when ε is near 0, it is misleading to claim that a method has good convergence properties on the basis that its error (in L_2 or L_∞) is bounded by $K\sigma(\varepsilon, h, \tau)$ for some constant K. Nevertheless, one frequently encounters this line of argument, which is equivalent to treating ε as a medium-sized constant!

Remark 3.15. While the argument we have just discussed does not give satisfactory theoretical error bounds, it often has some practical value. For suppose that we can prove L_2 stability or L_∞ stability for all $\varepsilon \in (0, 1]$, with possibly some restriction on the mesh. Suppose also that in the consistency analysis we pretend that ε is bounded away from zero, and obtain

$$|(Lu - L_{h,\tau}u)_i^j| + |(f - \tilde{f})_i^j| \le K(h^{\alpha_1} + \tau^{\alpha_2})$$

for some positive constants α_1 and α_2. Such a bound on the consistency error is often described as "*formal consistency of* $\mathcal{O}(h^{\alpha_1} + \tau^{\alpha_2})$". We cannot deduce for all values of ε that $\|u - u_{h,\tau}\|_d \le K(h^{\alpha_1} + \tau^{\alpha_2})$, where $\|\cdot\|_d$ is the discrete norm (L_2 or L_∞) that corresponds to the stability result proved. In practice, however, this order of convergence in the computed solution is often observed on those parts of Q that are not "near" the location of any layer in the analytic solution u. ♣

Example 3.16. The simple upwind scheme of Example 3.13 is often called "first-order upwinding". It is indeed formally first-order consistent, that is, formally consistent of $\mathcal{O}(h + \tau)$.

Suppose that (3.16) holds true. Then the scheme is L_∞ stable, as we saw in Example 3.13. Nevertheless, the nodal errors in the solution computed by this scheme cannot be described as "first-order", even in the simpler case of a two-point boundary value problem: consider the sharp error bounds of Theorem I.2.12 for the upwind scheme (I.2.12). ♣

To obtain an error bound for a difference scheme, one begins by establishing a consistency error bound such as (3.18), then a comparison principle is used in tandem with a careful choice of barrier function to complete the analysis. If the barrier function is sufficiently well behaved, then the scheme satisfies an error bound that is independent of ε. We already witnessed an illustration of this approach in the proof of Theorem I.2.18. Further pertinent examples occur later in this chapter.

3.3 Polynomial Schemes

In this section we continue to use the rectangular equidistant grid of Section 3.2 and examine schemes whose coefficients are polynomial or rational

functions of the differential equation coefficients and of h and τ. Such schemes can be motivated and derived in several ways. The simplest approach is to use Taylor expansions of the analytical solution u, just as for non-singularly perturbed differential equations. (See Section I.2.1.) But if no form of upwinding nor of artificial diffusion is incorporated into the scheme, then stability analyses for singularly perturbed problems will in general lead to excessively stringent conditions on the mesh. (For (1.1), "upwinding" means taking backward differences in space, as in Examples 3.2 and 3.8.) One such mesh restriction appeared already in Example 3.10, where the condition $\tau < 2\varepsilon/b^2$ was needed to ensure the L_2 stability of the scheme (3.12), which used central differences in space and no artificial diffusion.

To construct a difference scheme that approximates (1.1), we ignore the initial-boundary conditions. If necessary, one can later adjust the scheme near $S_x \cup S_0 \cup S_1$ to take account of the data there.

The derivation of various schemes using the Taylor expansion idea is standard material in basic numerical analysis courses. In particular it is thoroughly covered in [Hir88] and [Str04]. We shall describe an alternative approach (see, e.g., [MS93]) that shows that many schemes can be derived from an integral representation of the solution of an initial-value problem associated with (1.1).

Consider

$$w_t(x,t) - \varepsilon w_{xx}(x,t) + b w_x(x,t) = 0, \qquad (3.22a)$$

where $(x,t) \in \hat{Q} := (-\infty, \infty) \times (0,T]$, with

$$w(x,0) = \hat{s}(x) \quad \text{on } (-\infty, \infty), \qquad (3.22b)$$

and $\hat{s} \in L_2(-\infty, \infty)$ is given. For simplicity we have taken b constant and $d \equiv f \equiv 0$. The advantage of (3.22) over (1.1) is that the solution of (3.22) has the well-known integral representation

$$w(x,t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \hat{s}(x - bt + 2y\sqrt{\varepsilon t})\, e^{-y^2}\, dy \quad \text{for } -\infty < x < \infty. \qquad (3.23)$$

Place an equidistant tensor-product mesh (x_i, t_j) on $\hat{Q}$, where $x_0 = 0$, $x_i = ih$ for $-\infty < i < \infty$, and $t_0 = 0$, $t_j = j\tau$ for $j = 0, \dots, N$.

Apply (3.23) to the strip $(-\infty, \infty) \times [t_j, T]$, with initial data $w(\cdot, t_j)$. On setting $t = t_{j+1}$ and recalling that $t_{j+1} - t_j = \tau$, this gives

$$w(x, t_{j+1}) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} w(x - b\tau + 2y\sqrt{\varepsilon\tau}, t_j) e^{-y^2}\, dy \qquad (3.24)$$

for $-\infty < x < \infty$.

We then use (3.24) to generate difference schemes for (3.22). For example, one can use a low-degree polynomial to interpolate to $w(x, t_j)$ for x near x_i, then compute from (3.24) the evolution of this polynomial at time t_{j+1}. This yields a difference scheme that relates nodal values at time t_{j+1} to nodal values at time t_j.

Example 3.17. Fix an integer i. Let $\tilde{w}(x, t_j)$ denote the quadratic polynomial in x that interpolates to w_k^j for $k = i-1, i$ and $i+1$. That is,

$$\tilde{w}(x, t_j) = w_i^j + \frac{w_{i+1}^j - w_{i-1}^j}{2h}(x - x_i) + \frac{w_{i+1}^j - 2w_i^j + w_{i-1}^j}{2h^2}(x - x_i)^2.$$

Replace w in the right-hand side of (3.24) by $\tilde{w}$, then evaluate the integral exactly, using

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} y^m e^{-y^2}\, dy = \begin{cases} 1 & \text{if } m = 0, \\ 0 & \text{if } m = 1, \\ 1/2 & \text{if } m = 2. \end{cases}$$

This yields

$$w(x, t_{j+1}) = w_j^i + \frac{w_{i+1}^j - w_{i-1}^j}{2h}(x - b\tau - x_i)$$
$$+ \frac{w_{i+1}^j - 2w_i^j + w_{i-1}^j}{2h^2}[2\varepsilon\tau + (x - b\tau - x_i)^2].$$

Set $x = x_i$ here. We obtain

$$u_i^{j+1} = u_i^j - \frac{b\tau}{2h}(u_{i+1}^j - u_{i-1}^j) + \left(\frac{\varepsilon\tau}{h^2} + \frac{b^2\tau^2}{2h^2}\right)(u_{i+1}^j - 2u_i^j + u_{i-1}^j). \quad (3.25)$$

This explicit one-step scheme is the well-known *Lax-Wendroff scheme* for (1.1). Evidently it can also be derived by applying a central difference in space and a forward difference in time to the differential equation

$$\tilde{w}_t(x, t) - \left(\varepsilon + \frac{b^2\tau}{2}\right)\tilde{w}_{xx}(x, t) + b\tilde{w}_x(x, t) = 0. \quad (3.26)$$

We see from Lemma 3.12 that the scheme satisfies a discrete maximum principle if $2\nu \leq \nu^2 + 2\mu \leq 1$, where $\nu = b\tau/h$ and $\mu = \varepsilon\tau/h^2$. These inequalities imply that $\nu \leq 1/2$ and $\nu \leq 4\mu/3$. That is, $3bh \leq 4\varepsilon$, which is too restrictive in practice.

A von Neumann stability analysis of (3.25) yields

$$\xi = 1 - (\nu^2 + 2\mu)(1 - \cos\theta) - \sqrt{-1}\,\nu\sin\theta.$$

Hence

$$|\xi|^2 = 1 - 2\mu\sin^2\theta - 4(\nu^2 + 2\mu)(1 - \nu^2 - 2\mu)\sin^4(\theta/2).$$

Consequently the Lax-Wendroff scheme is L_2 stable if $\nu^2 + 2\mu \leq 1$. When $\varepsilon \ll 1$, this condition is not a serious restriction on the mesh. We observe from (3.26) that (3.25) has in effect achieved L_2 stability by adding artificial diffusion of magnitude $b^2\tau/2$ to (3.22a); compare the scheme (2.14) of Part I.

An analysis of the consistency error shows that the Lax-Wendroff scheme is formally consistent of $\mathcal{O}(\tau + h^2)$. ♣

If in Example 3.17 the quadratic interpolant is replaced by a linear interpolant to w_{i-1}^j and w_i^j, this reproduces the simple upwind scheme of Example 3.4. If instead we use a linear interpolant to w_{i-1}^j and w_{i+1}^j, this generates the *Lax-Friedrichs scheme*

$$u_i^{j+1} = \frac{u_{i+1}^j + u_{i-1}^j}{2} - \frac{b\tau}{2h}(u_{i+1}^j - u_{i-1}^j).$$

No term corresponding to $-\varepsilon u_{xx}$ appears in these two schemes because if the initial data in (3.22b) is a linear function, then the solution of (3.22a) is also linear, so $-\varepsilon w_{xx} = 0$.

If the quadratic interpolant of Example 3.17 is replaced by a cubic interpolant to w_k^j for $k = i - 2, i - 1, i$ and $i + 1$, one gets the QUICKEST scheme of [Leo79a]:

$$u_i^{j+1} = u_i^j - \frac{\nu}{2}(u_{i+1}^j - u_{i-1}^j) + \left(\frac{\nu^2}{2} + \mu\right)(u_{i+1}^j - 2u_i^j + u_{i-1}^j)$$
$$+ \frac{\nu}{6}\left(1 - \nu^2 - 6\mu\right)(u_{i+1}^j - 3u_i^j + 3u_{i-1}^j - u_{i-2}^j),$$

where ν and μ are as in Example 3.17. This scheme is explicit and formally consistent of $\mathcal{O}(\tau + h^3)$. It is L_2 stable if

$$\nu^2 + \frac{6\mu(1 - 2\nu)}{3 - 2\nu} \leq 1.$$

Clearly we could choose other types of approximation to $w(\cdot, t_j)$. A piecewise polynomial finite element approach is also considered in [MS93], but in this case the integral in (3.24) cannot be evaluated exactly and further approximations must be introduced.

An error analysis of any scheme that is derived from (3.24) can in principle be deduced from the approximations introduced during the derivation; see [MS93].

Remark 3.18. Many stable schemes for (1.1) can be generated by Taylor expansions of u and a careful choice of the difference quotient coefficients. For example, an implicit one-step scheme that is formally consistent of $\mathcal{O}(\tau^2 + h^4)$ is given in [BSC+80]. This scheme is *compact*, that is, it uses three grid points on each of two time levels to obtain optimal fourth-order formal consistency in the x-direction. The derivation of the scheme is based on the HODIE approach of Section I.2.1.4.

In [DRH98], Donea et al. systematically consider various ways of constructing finite difference methods for (1.1) that are high-order accurate in time: Taylor-Galerkin methods, multistage methods based on Padé approximation of the exponential function, Runge-Kutta methods, and implicit methods based on Newton-Cotes quadrature approximation of the integrated time derivative. The accuracy and stability of each method is analysed.   ♣

3.4 Uniformly Convergent Methods

Uniformly convergent methods for ordinary differential equations were introduced in Section I.2.1.3. We now extend this idea to time-dependent problems. Methods of this type are constructed with the aim of computing an accurate approximation of the solution of (1.1) at all mesh points (recall Remark 3.15).

On a family of rectangular equidistant grids as in Section 3.2, a scheme is said to be *uniformly convergent in the discrete maximum norm* if its solution $\{u_i^j\}$ satisfies

$$|u(x_i, t_j) - u_i^j| \le C(h^{\alpha_1} + \tau^{\alpha_2}) \quad \text{for all } i \text{ and } j, \tag{3.27}$$

where α_1 and α_2 are positive constants that are independent of ε and of the mesh, and u is the solution of (1.1).

Analogously to Theorem I.2.17, the coefficients of uniformly convergent difference schemes on an equidistant mesh must satisfy a special condition. This result is stated precisely in the next theorem, whose proof is similar to that of Theorem I.2.17 and can be found in [Guo93].

Theorem 3.19. *Let b be a positive constant and set $d \equiv 0$ in (1.1). Take $s \equiv q_0 \equiv q_1 \equiv 0$ and assume that $f \in C^2(\bar{Q})$. Take $h = \tau$. Assume that the difference scheme can be written in the form*

$$\sum_{n=0}^{1} \sum_{m=-1}^{1} \alpha_{m,n} u_{i+m}^{j+n} = h \tilde{f}_i^j \tag{3.28}$$

for $i = 1, \ldots, M-1$ and $j = 0, \ldots, N-1$, where the coefficients $\alpha_{m,n}$ depend only on m, n and the ratio h/ε, and $|\tilde{f}_i^j| \le C$ for all i and j as M and N vary.

If the scheme is uniformly convergent in the discrete maximum norm, then

$$\sum_{n=0}^{1} \sum_{m=-1}^{1} \alpha_{m,n} = 0, \tag{3.29a}$$

$$\sum_{n=0}^{1} \sum_{m=-1}^{1} \alpha_{m,n} \exp(-bmh/\varepsilon) = 0. \tag{3.29b}$$

Remark 3.20. Condition (3.29a) is mild and is satisfied by any reasonable scheme, but (3.29b) shows that, just as in Theorem I.2.17, only those schemes that possess a certain exponential character can be uniformly convergent.

The hypotheses of Theorem 3.19 are reasonable. Any scheme that claims to solve (1.1) accurately on all of $\bar{Q}$ should be able to handle the special case of (1.1) considered in the theorem. The assumption that the scheme can be written in the form (3.28), with each $\alpha_{m,n}$ depending only on m, n and h/ε,

is satisfied by all schemes of which we are aware, whether or not they are uniformly convergent.

The theorem can easily be generalized to multi-step schemes and to methods that use more than three nodes on each time level.   ♣

3.4.1 Exponential Fitting in Space

In Section I.2.1 we discussed various forms of exponential fitting for second-order ordinary differential equations. Some of these yielded uniformly convergent difference schemes. It is infeasible to follow exactly the same approach for (1.1) because the solution of a partial differential equation with constant coefficients cannot in general be expressed in terms of functions that are easily evaluated. One can compromise by using exponential fitting only for the space derivatives then approximating the u_t term by a polynomial difference, and in this section we describe two methods of this type. In recent years, interest in using this approach to solve (1.1) has diminished, since, *inter alia*, it is difficult to generalize satisfactorily to time-dependent problems in more than one space dimension.

Titov and Shishkin [TS76] use an equidistant tensor-product mesh. They begin with a backward Euler scheme for the time derivative, then apply complete exponential fitting to the singularly perturbed two-point boundary value problem generated at each time level. The solution is computed at successive time levels.

To describe their scheme, assume for simplicity that d is constant. Suppose that the u_i^{j-1}, for $i = 0, \ldots, M$, are known. Then the u_i^j, for $i = 0, \ldots, M$, are computed from

$$-\gamma_{1,i}^j \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + \gamma_{2,i}^j \frac{u_{i+1}^j - u_{i-1}^j}{2h} + u_i^j = \frac{1}{1 + d\tau}(\tau f_i^j + u_i^{j-1}),$$

where

$$\gamma_{1,i}^j = -\frac{h^2}{4}\left[\coth\left(\frac{\lambda_{1,i}^j h}{2}\right)\coth\left(\frac{\lambda_{2,i}^j h}{2}\right) + 1\right]$$

and

$$\gamma_{2,i}^j = \frac{h}{2}\left[\coth\left(\frac{\lambda_{1,i}^j h}{2}\right) + \coth\left(\frac{\lambda_{2,i}^j h}{2}\right)\right]$$

are the roots of the equation

$$-\frac{\varepsilon\tau}{1 + d\tau}\lambda^2 + \frac{b_i^j \tau}{1 + d\tau}\lambda + 1 = 0.$$

The following result is proved in [TS76] *without making compatibility assumptions on the data*, unlike the vast majority of papers dealing with uniform convergence results for (1.1).

Theorem 3.21. *Assume that all the data of (1.1) are C^3. Then there exists a constant C such that the solution $\{u_i^j\}$ of the Titov-Shishkin scheme satisfies*

$$|u(x_i, t_j) - u_i^j| \le C(h^\alpha \tau^{-1} + \tau^{1/3})$$

for all i and j, where $\alpha \in (0, 2/5)$ is a certain positive constant.

A particularly ingenious example of exponential fitting in space is due to Stoyan [Sto82], whose scheme is designed to suit various types of problem including convection-diffusion. The scheme modifies its coefficients in a continuous manner as ε, h, τ and the given functions in (1.1a) vary. In its general form on nonequidistant tensor-product meshes, it includes a variable coefficient $r(x, t)$ of u_t in (1.1), *permits ε or r (but not both) to vanish* and can handle Dirichlet, Neumann and Robin boundary conditions.

Stoyan begins by constructing the exact three-point difference scheme (cf. Section I.2.1.3) for the constant coefficient ordinary differential equation $-\varepsilon y''(x) + by'(x) = g(x)$, where g is an arbitrary linear function. He simplifies this scheme slightly, generalizes it in a simple and obvious way for variable b, then moves to the parabolic problem by formally replacing $y(x)$ by $u(x, t)$ and $g(x)$ by $f(x, t) - (ru_t)(x, t)$. Finally, an analysis of stability motivates the introduction of a weighting parameter in the coefficients of the scheme.

For the details we refer the reader to [Sto82], where there is a lengthy discussion of the choice of free parameters in the method. Although no uniform convergence result has been proved for this scheme, it is included here as it is in the spirit of this section.

Section 4.1.2 will discuss some uniformly convergent finite element schemes that are exponentially fitted in space.

3.4.2 Layer-Adapted Tensor-Product Meshes

Assume that $b > 0$ and that problem (1.1) has smooth data that are compatible at the corners $(0, 0)$ and $(0, 1)$. Then the only layer in the solution u is an exponential boundary layer along the side $x = 1$ of $\bar{Q}$, as we learned in Section 2.2. To obtain a uniformly convergent finite difference method, an alternative to exponential fitting is (as in Section I.2.4) to construct a mesh that becomes very fine in the x-direction as x nears 1, together with a simple polynomial scheme.

In this section we consider tensor-product meshes that are equidistant in the t-direction, with spacing τ. In the x-direction, these meshes are nonequidistant with N subintervals; they may for example be piecewise equidistant Shishkin meshes as in Section I.2.4.2 or graded Bakhvalov-type meshes as in Section I.2.4.1.

Thus in the case of a Shishkin mesh, given an even positive integer N, set

$$\sigma = 1 - (k\varepsilon/\beta) \ln N,$$

where one typically takes $k \geq 2$ (see Remarks I.2.99 and I.2.104), then divide each of $[0, \sigma]$ and $[\sigma, 1]$ into $N/2$ equal intervals. The resulting mesh on $\bar{Q}$ is coarse on $[0, \sigma] \times [0, T]$ and fine in the x-direction on $[\sigma, 1] \times [0, T]$. The Bakhvalov-type tensor-product mesh is defined similarly, invoking a formula such as (2.135) from Section I.2.4.1 to define the mesh in the x-direction.

If one uses a tensor-product mesh that is a Shishkin mesh in the x-direction, then approximates $-\varepsilon u_{xx} + bu_x + du$ by one-dimensional upwind differencing and u_t by backward differencing, this implicit scheme satisfies a discrete maximum principle; if one has sufficient smoothness and compatibility of the data of the problem, then a consistency and barrier function argument [Kop97, Shi92b] shows that the solution $\{u_i^j\}$ of the scheme satisfies

$$|u(x_i, t_j) - u_i^j| \leq C(N^{-1} \ln N + \tau) \tag{3.30}$$

for all i and j and some constant C. Boglaev [Bog01, Bog06] analyses a parallelizable domain decomposition algorithm for this method; see [MOS96, Chapter 10] for a related study in the steady-state case.

A similar method on a similar mesh is considered in [HSS03]. Defect correction is then applied in both space and time to enhance the accuracy of the computed solution. The final computed solution $\{u_i^j\}$ satisfies

$$|u(x_i, t_j) - u_i^j| \leq C(N^{-2} \ln^2 N + \tau^3)$$

for all i and j and some constant C, provided that certain smoothness and compatibility conditions are satisfied by the data of the problem.

An alternative improvement of (3.30) is given in [Shi08], where it is shown in an n-width setting that, for the same number of degrees of freedom as in (3.30), one can obtain the optimal bound

$$|u(x_i, t_j) - u_i^j| \leq C N^{-1} \ln^{1/2} N \quad \text{for all } i \text{ and } j$$

by using $\mathcal{O}(N \ln^{1/2} N)$ points in space and $\mathcal{O}(N \ln^{-1/2} N)$ points in time in the tensor-product mesh.

On the same mesh, Kopteva [Kop97] defines the discrete approximation $L_x^N u_i^j$ of $(-\varepsilon u_{xx} + bu_x + du)(x_i, t_j)$ using central differencing to approximate bu_x, then investigates the scheme

$$\frac{u_i^{j+1} - u_i^j}{\tau} + \theta L_x^N u_i^{j+1} + (1 - \theta)L_x^N u_i^j = \theta f_i^{j+1} + (1 - \theta)f_i^j,$$

where $\theta \in [0.5, 1]$ is a user-chosen parameter ($\theta = 1$ yields the backward Euler method, $\theta = 0.5$ produces Crank-Nicolson). Under the assumption that one has the S-decomposition of Remark 2.9, it is shown that for the solution $\{u_i^j\}$ of the scheme one has

$$|u(x_i, t_j) - u_i^j| \leq C(N^{-2} \ln^2 N + \tau) \quad \text{if } \theta > 0.5$$

$$\text{and} \quad \left| u\left(x_i, t_j + \frac{\tau}{2}\right) - \frac{u_i^{j+1} + u_i^j}{2} \right| \leq C(N^{-2} \ln^2 N + \tau^2) \quad \text{if } \theta = 0.5,$$

for all i and j and some constant C.

In [Kop01b], the problem (1.1) is considered in the conservation form $u_t - \varepsilon u_{xx} + (bu)_x = f$. A four-point conservative discrete approximation of $-\varepsilon u_{xx} + (bu)_x$ from [GP69] is used on a class of nonequidistant meshes in the x-direction that includes both Shishkin and Bakhvalov-type meshes, and a backward difference approximates u_t. The four-point space difference operator has only non-oscillatory solutions, but it does not yield an M-matrix and this increases the complexity of the analysis. The inequality (I.2.145) is proved for the space difference operator and a related stability inequality is deduced for the full scheme on the (x,t) mesh. It is finally shown that the solution $\{u_i^j\}$ of the scheme satisfies

$$|u(x_i, t_j) - u_i^j| \leq C(N^{-2} \ln^k N + \tau)$$

for all i and j and some constant C, with $k = 0$ on a Bakhvalov-type mesh and $k = 2$ on a Shishkin mesh.

Clavero et al. [CGJ05] construct a HODIE difference scheme (see Section I.2.1.4) to discretize the space derivatives on a Shishkin mesh and approximate the time derivative by the classical Crank-Nicolson approximation. Under the assumption that $N^{-q} \leq C\tau^{3/2}$ for some C and some constant $q \in (0,1)$, they prove that the solution $\{u_i^j\}$ of the scheme satisfies

$$|u(x_i, t_j) - u_i^j| \leq C'(N^{-2+q} \ln^2 N + \tau^{3/2})$$

for all i and j and some constant C'.

The approximation of first-order derivatives of u, using simple upwinding on a Shishkin mesh to approximate u_x together with backward differencing of u_t, is discussed at length in [Shi04a].

We know of no paper that explicitly proves uniform convergence of a numerical method for (1.1) when an interior layer caused by incompatibility of the data emanates from the corner $(0,1)$, though related problems are analysed in [FHS96c, HS93, Shi88, Shi03]. A variant of (1.1), where $\bar{Q}$ is decomposed into two subdomains and the functions b, c and f are discontinuous across the interface, is examined in [SSH04]; a numerical method that handles the interior layer lying along the interface is constructed via a coordinate transformation and almost uniform convergence of the computed solution is proved.

3.4.3 Reaction-Diffusion Problems

Recall that one has a reaction-diffusion problem when $b \equiv 0$ in (1.1). That is, again setting $Q = (0,1) \times (0,T]$,

$$u_t(x,t) - \varepsilon u_{xx}(x,t) + d(x,t)u(x,t) = f(x,t) \quad \text{for } (x,t) \in Q, \qquad \text{(3.31a)}$$
$$u(x,0) = s(x) \quad \text{on } S_x := \{(x,0) : 0 \leq x \leq 1\}, \qquad \text{(3.31b)}$$
$$u(0,t) = q_0(t) \quad \text{on } S_0 := \{(0,t) : 0 < t \leq T\}, \qquad \text{(3.31c)}$$
$$u(1,t) = q_1(t) \quad \text{on } S_1 := \{(1,t) : 0 < t \leq T\}. \qquad \text{(3.31d)}$$

The solution u of (3.31) has in general *parabolic boundary layers* along the sides $x = 0$ and $x = 1$ of $\bar{Q}$; see Remark 2.11. The presence of these layers engenders the following remarkable result that alerts us to a fundamental difference between parabolic boundary layers and exponential boundary layers.

Remark 3.22. (Shishkin's obstacle result for parabolic layers) Suppose that the grid is equidistant in both x and t. Consider an arbitrary difference scheme that uses a fixed number of grid points in the space direction on each of a fixed number of time levels, satisfies a discrete maximum principle, and whose coefficients may depend on the coefficients of the differential operator, but must not depend on the boundary data. (This final hypothesis regarding the coefficients of the scheme is perfectly natural.) Then *it is impossible to achieve uniform convergence in the discrete maximum norm for the class of reaction-diffusion problems*, i.e., there are no positive α_1 and α_2 for which inequality (3.27) holds true as u ranges over a class of problems. This extraordinary conclusion was reached by Shishkin [Shi89]; see [MOS96, Chapter 14], [GRS07, pp.405–6] and [Shi97b] for detailed discussions that include, *inter alia*, the fact that the discrete maximum principle hypothesis can be discarded.

The essential reason for this negative result is that (cf. Example III.1.16) the behaviour of the solution u inside a parabolic boundary layer is influenced by *all* the boundary data along that side of Q (i.e., $x = 0$ or $x = 1$), and this data can be taken from the infinite set of functions $\mathcal{S} := \{t, t^2, t^3, \dots\}$; by a difficult extension of the proof of Theorem I.2.17, on equidistant meshes each function from $\mathcal{S}$ imposes a different condition on the coefficients of any uniformly convergent difference scheme. No difference scheme with a fixed stencil can satisfy all these conditions.   ♣

For the convection-diffusion problem (1.1) with $b > 0$, where boundary layers are exponential, methods that are uniformly convergent in the maximum norm on equidistant meshes certainly exist: see, e.g., Theorems 3.21 and 4.4. On the other hand, for reaction-diffusion problems where $b \equiv 0$, Remark 3.22 warns us that nonequidistant grids should be used if one is to attain uniform convergence in the maximum norm. In the remainder of this section we consider tensor-product meshes that are equidistant in the t-direction with spacing τ, while in the x-direction these meshes are nonequidistant with N subintervals. (It isn't necessary to modify the mesh spacing in the t-direction to achieve the desired uniform convergence.) Each of the bounds stated below assumes regularity of the data and some compatibility at the corners (0,0) and (1,0).

In [Shi83], the reaction-diffusion problem is solved using a tensor-product grid that is equidistant in the t-direction and of Bakhvalov type in the x-direction, so that the mesh becomes fine for x near 0 and x near 1, with the standard difference approximation (2.132) of u_{xx} and a backward difference approximation of u_t. The resulting implicit one-step scheme satisfies a discrete maximum principle, as can be seen from Lemma 3.12. Then, assuming sufficient smoothness of the solution u away from the boundary layers, a

consistency and barrier function argument yields

$$|u(x_i, t_j) - u_i^j| \le C(N^{-1} + \tau) \text{ for all } i \text{ and } j,$$

where $\{u_i^j\}$ is the computed solution. Shishkin returns to this scheme in [Shi84] and shows that Richardson extrapolation can be used to accelerate the convergence.

To choose a Shishkin mesh in the x-direction for our reaction-diffusion problem, one follows the recipe of Remark I.2.106, with perhaps a modified constant multiplier in the formula for the transition points – recall Remark I.2.104. The remaining schemes discussed in this section, except those of [LM07], use Shishkin meshes.

Using the standard approximation of u_{xx} and a backward difference approximation of u_t, a consistency error and barrier function argument shows [MOSS98] that

$$|u(x_i, t_j) - u_i^j| \le C(N^{-2} \ln^2 N + \tau) \text{ for all } i \text{ and } j,$$

where as usual $\{u_i^j\}$ is the computed solution. (In [LM07] a transparent analysis on general layer-adapted meshes includes this result as a special case.) This scheme is modified in [HSS00] by means of a defect correction technique that improves the difference approximation of u_t; two schemes are constructed for which

$$|u(x_i, t_j) - u_i^j| \le C(N^{-2} \ln^2 N + \tau^k) \text{ for all } i \text{ and } j,$$

with $k = 2$ and 3 respectively. If instead one applies defect correction to enhance the difference approximation of u_{xx}, a scheme is obtained [Shi96b] for which

$$|u(x_i, t_j) - u_i^j| \le C(N^{-6} \ln^6 N + \tau) \text{ for all } i \text{ and } j.$$

In [CG05] a HODIE difference scheme (see Section I.2.1.4) is used to discretize $-\varepsilon u_{xx} + cu$ while a third-order two-stage SDIRK method is applied to the time derivative; for the resulting method one has

$$|u(x_i, t_j) - u_i^j| \le C(N^{-3} + \tau^3) \text{ for all } i \text{ and } j.$$

A system of two reaction-diffusion equations, coupled through their zero-order reaction terms, is solved in [GL07] using central differencing on a Shishkin mesh and an error bound is derived. A more incisive analysis of the same problem and method in [Lin08b] gives the sharp error bound

$$|u(x_i, t_j) - u_i^j| \le C(N^{-2} \ln^2 N + \tau) \text{ for all } i \text{ and } j.$$

Remark 3.23. If the initial data $s(x)$ has a discontinuity at (say) $x = \hat{x}$, this will cause a parabolic interior layer in Q along the line $x = \hat{x}$. Although this

layer is asymptotically similar to a parabolic boundary layer, nevertheless the negative result of Remark 3.22 no longer applies: one can construct a fitted scheme on an equidistant mesh that yields a uniformly convergent solution. This is done in [HS93], where it is shown that

$$|u(x_i, t_j) - u_i^j| \leq C(h^{2/3} + \tau^{1/3}) \text{ for all } i \text{ and } j,$$

provided that $h \geq C'\tau^{7/6}$ for some constant C'; here h is the mesh diameter in the x-direction. See also [FHS96c]. The essential difference from Remark 3.22 is that here the only piece of boundary data that influences the parabolic interior layer is the multiplier effect of the magnitude of the jump in $s(x)$ at $x = \hat{x}$, and consequently the scheme is no longer required to satisfy infinitely many conditions.   ♣

In [Shi04b] the difficult problem of a concentrated source on a moving boundary $x = \phi(t)$ between two subdomains of $[0,1]$ is considered; the solution u has an interior layer along the curve $x = \phi(t)$, and the theory of Kolmogorov n-widths is employed to show that the orientation of this curve must be taken into account when constructing a uniformly convergent method.

4

Finite Element Methods

From Part I we know that standard Galerkin finite element methods on equidistant meshes yield inaccurate approximate solutions of singularly perturbed two-point boundary value problems unless a large number of mesh points are used. The same disappointing behaviour occurs arises when dealing with parabolic convection-diffusion problems, because such methods have no built-in upwinding. Finite element methods will now be developed specifically for the convection-diffusion situation, either by choosing special basis functions or by working on meshes designed for these problems.

The problem considered in Chapter 4 is

$$u_t(x,t) - \varepsilon u_{xx}(x,t) + b(x,t)u_x(x,t) + d(x,t)u(x,t) = f(x,t) \qquad (4.1a)$$

where $(x,t) \in Q := (0,1) \times (0,T]$, and

$$u(x,0) = s(x) \quad \text{on } S_x := \{(x,0) : 0 \le x \le 1\}, \qquad (4.1b)$$
$$u(0,t) = 0 \quad \text{on } S_0 := \{(0,t) : 0 < t \le T\}, \qquad (4.1c)$$
$$u(1,t) = 0 \quad \text{on } S_1 := \{(1,t) : 0 < t \le T\}. \qquad (4.1d)$$

Again assume that $d(x,t) \ge \gamma > 0$ and $b(x,t) > \beta > 0$. Unlike (1.1) homogeneous boundary conditions are used here, which is equivalent to solving (1.1) for the unknown function

$$u(x,t) - (1-x)q_0(t) - xq_1(t),$$

so there is no loss of generality. By changing the dependent variable in (4.1a) as in Section I.2.2.5, one can also assume that

$$d(x,t) - \frac{1}{2}b_x(x,t) \ge \omega > 0 \quad \text{on } Q. \qquad (4.2)$$

Some of the above assumptions will not hold true in certain examples below, but in each case the reader will see that nevertheless the integrity of the argument is preserved.

We concentrate on equidistant meshes. The simplest finite element approach handles the space derivatives using ideas from Section I.2.2, and approximates u_t by finite differences; see Section 4.1. The alternative possibility of treating separately the diffusive $(-\varepsilon u_{xx})$ and convective $(u_t + bu_x)$ operators is discussed in Section 4.2. A layer-adapted mesh is examined on page 210.

4.1 Space-Based Methods

Although Q is a two-dimensional domain, when solving the initial-boundary value problem (4.1) it's common to use finite element methods whose trial and test functions depend only on the single variable x, leaving until later the discretization of the time derivative. This technique is an example of the *method of lines*. For a lengthy discussion of this method, with many examples and much analysis, see [HV03].

Partition [0,1] by the equidistant grid $\{x_i\}$, where $x_i = i/M = ih$ for $i = 0, \ldots, M$. Then choose a basis $\{\phi_i : i = 1, \ldots, M-1\}$ of finite element functions for an $(M-1)$-dimensional subspace of the Sobolev space $H_0^1(0,1)$. For example, each ϕ_i may be the standard piecewise linear "hat" function that satisfies $\phi_i(x_j) = \delta_{ij}$ for all i and j.

We seek an approximate solution $u_h(x,t)$ of (4.1) in the form

$$u_h(x,t) = \sum_{i=1}^{M-1} u_i(t)\phi_i(x), \qquad (4.3)$$

where the u_i are at present unknown functions.

Next, choose a basis $\{\psi_i(x) : i = 1, \ldots, M-1\}$ of finite element functions for an $(M-1)$-dimensional test space. One often takes $\psi_i = \phi_i$ for all i in problems that are not convection-dominated, but for (4.1) the ψ_i are usually upwinded versions of the ϕ_i. Since the trial and test functions are not identical, this is a Petrov-Galerkin finite element method.

Armed with our ψ_i, we write down a *semidiscrete form* of (4.1a):

$$\big(\varepsilon(u_h)_x, \psi_i'\big) + \big((u_h)_t + b(u_h)_x + du_h, \psi_i\big) = (f, \psi_i), \quad \text{for each } t \in (0,T], \ (4.4)$$

where $i = 1, \ldots, M-1$, and $(\cdot, \cdot)$ denotes the $L_2(0,1)$ inner product. The initial condition (4.1b) is usually approximated either by interpolation, i.e.,

$$u_h(x_i, 0) = s(x_i) \quad \text{for } i = 0, \ldots, M,$$

or by L_2 projection, viz.

$$\int_{x=0}^{1} [u_h(x,0) - s(x)]\psi_i(x)\, dx = 0 \quad \text{for } i = 1, \ldots, M-1,$$

or by some other projection.

Integrating (4.4), we obtain a system of first-order ordinary differential equations in the unknowns $u_i(t)$, where $i = 1, \dots, M - 1$. The approximation of (4.1b) yields initial data for this problem. One arrives at a fully discrete approximation of (4.1) by discretizing the system (4.4) with respect to t.

Alternatively, one might discretize first in time, which yields a family of steady-state problems that must in turn be discretized. This is known as *Rothe's method* or the *horizontal method of lines*. See [AAS07] for a comparison of this approach with the method of lines that we consider here.

In Section 4.1.1, stable methods are generated by using polynomials to upwind the ψ_i. Section 4.1.2 examines upwinding using exponentials, in order to obtain uniformly convergent methods. Finally, Section 4.1.3 gives local convergence results that make minimal assumptions on the data of the problem.

4.1.1 Polynomial Upwinding

Suppose that the ϕ_i are piecewise polynomials. Then the ψ_i can be generated by adding a suitable piecewise polynomial of higher degree to each ϕ_i.

Example 4.1. Take each ϕ_i to be the standard piecewise linear "hat" function centred on x_i, with

$$\psi_i = \phi_i + \alpha v_i, \tag{4.5}$$

where α is a real parameter and v_i is piecewise quadratic, as in Section I.2.2.2. Mitchell and Griffiths [MG79] discuss such a method, with

$$v_i(x) = \begin{cases} 3(x - x_{i-1})(x_i - x)/h^2 & \text{if } x_{i-1} \le x \le x_i, \\ -3(x - x_i)(x_{i+1} - x)/h^2 & \text{if } x_i \le x \le x_{i+1}, \\ 0 & \text{if } |x - x_i| > h. \end{cases}$$

For the case where b is constant and $d \equiv f \equiv 0$, the semidiscrete form (4.4) becomes the system of equations

$$-\varepsilon S U_h + h^2 M \frac{\partial U_h}{\partial t} = 0.$$

Here $U_h := [u_1(t), \dots, u_{M-1}(t)]^T$, $S := (1 + \alpha R)A - RB$, where $R := bh/(2\varepsilon)$ is the *cell Reynolds number*, and the matrices M, A and B are given by

$$M = \frac{1}{6} \begin{pmatrix} 4 & 1 - 3\alpha/2 & 0 & 0 & \cdots \\ 1 + 3\alpha/2 & 4 & 1 - 3\alpha/2 & 0 & \cdots \\ 0 & & \ddots & & \ddots & \ddots \end{pmatrix},$$

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 & \cdots \\ 1 & -2 & 1 & 0 & \cdots \\ 0 & & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

and

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 1 & 0 & 1 & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

The matrix M, which is the (scaled) coefficient of $\partial U_h/\partial t$, is known as the *mass matrix*.

For two-point boundary value problems where the differential equation has constant coefficients, α can be chosen in the analogue of this method so as to get the exact solution at every node [MG80]. For (4.1), no value of α will guarantee an exact nodal solution; various choices are investigated in [MG79], but no recommendation for an "optimal" α is made. ♣

Westerink and Shea [WS89] consider piecewise linear ϕ_i and

$$\psi_i = \phi_i + \alpha v_i + \sigma w_i, \tag{4.6}$$

where α and σ are real parameters, v_i is piecewise quadratic and w_i piecewise cubic. This permits better control of truncation error than when (4.5) is used. Furthermore, a Fourier analysis [WS89] indicates that the choice (4.6) controls phase errors without excessive damping of the solution, whereas (4.5) cannot simultaneously achieve both these aims.

An alternative is to use a piecewise quadratic ϕ_i with $\psi_i = \phi_i + \sigma_1 w_i + \sigma_2 q_i$, where σ_1 and σ_2 are real parameters, w_i is piecewise cubic and q_i piecewise quartic.

Numerical results in [WS89], for the case of the constant coefficient reduced problem (i.e., $\varepsilon = 0$) with a Gaussian pulse as initial data, bear out the above theoretical results for the linear/cubic schemes and show that the quadratic/quartic scheme is more accurate for this problem. Nevertheless, it is unclear how well these schemes would perform in the presence of an outflow boundary layer.

Yu and Heinrich [YH86] analyse a related approach with space-time finite elements for a problem where $d \equiv f \equiv 0$ and b is constant. They use a tensor-product uniform grid $\{(x_i, t_j) : i = 0, \ldots, M, \ j = 0, \ldots, N\}$, with piecewise bilinear trial functions $w_{ij}(x, t)$ satisfying $w_{ij}(x_k, t_n) = \delta_{ik}\delta_{jn}$ for each node (x_k, t_n). Their test functions have the form

$$\sigma_{ij} + \sigma_1 \frac{\partial \sigma_{ij}}{\partial x} + \sigma_2 \frac{\partial^2 \sigma_{ij}}{\partial x \partial t}, \tag{4.7}$$

where σ_1 and σ_2 are real parameters. To discretize (4.1a), they use the weak form

$$\int_{t=t_{n-1}}^{t_n} \left\{ \int_{x=0}^{1} \left[\varepsilon(u_h)_x w_x + ((u_h)_t + b(u_h)_x)w \right] dx - \varepsilon(u_h)_x w|_{x=0}^{1} \right\} dt = 0 \tag{4.8}$$

for $n = 1, \ldots, N$, where u_h is the computed solution and w is an arbitrary test function. This scheme satisfies the von Neumann stability condition on

every equidistant grid. It is formally consistent of $\mathcal{O}(h^3 + \tau)$, where τ is the grid spacing in the t-direction.

To obtain better formal consistency, one can replace the bilinear trial functions here by functions that are linear in space and quadratic in time. The resulting scheme is formally consistent of $\mathcal{O}(h^3 + \tau^2)$, and satisfies the von Neumann condition if the Courant number $b\tau/h$ is less than one. In fact [YH86] there is an optimal choice of $b\tau/h$ that yields formal consistency of $\mathcal{O}(h^4 + \tau^2)$.

Several numerical experiments in [YH86] deal with the convection and diffusion of a Gaussian pulse, but outflow boundary layers are not present.

Remark 4.2. (SDFEM in space) The streamline diffusion method (SDFEM) for two-point boundary value problems was already discussed in Section I.2.2.3. To discretize (4.1), one could combine the SDFEM in space with a subsequent time discretization; see Remark III.4.3. ♣

4.1.2 Uniformly Convergent Schemes

Theorem 3.19 implies that none of the schemes encountered so far in Section 4.1 is uniformly convergent. Since the condition (3.29b) of Theorem 3.19 resembles strongly the corresponding condition (2.19) of Part I for two-point boundary value problems, one can hope to generate uniformly convergent schemes for (4.1) by combining a uniformly convergent finite element method from Section I.2.2.5 with a simple approximation of u_t.

Place the usual equidistant tensor-product grid on $\bar{Q}$. The nodes are (x_i, t_j), where $x_i = i/M = ih$ for $i = 0, \ldots, M$ and $t_j = jT/N = j\tau$ for $j = 0, \ldots, N$.

Example 4.3. We begin by defining trial and test functions on each line segment $[0, 1] \times \{t_j\}$. The trial basis functions $\{\phi^{ij}(x, t_j) : i = 1, \ldots, M - 1\}$ are standard piecewise linears that satisfy $\phi^{ij}(x_k, t_j) = \delta_{jk}$. Set $w_i^j = w(x_i, t_j)$ for all $w \in C(\bar{Q})$. Define the piecewise constant approximation $\bar{w}(\cdot, t_j)$ of $w(\cdot, t_j)$ by

$$\bar{w}(x, t_j) = \begin{cases} w_i^j & \text{if } x_{i-1} < x < x_i \text{ for } i = 1, \ldots, M, \\ 0 & \text{otherwise.} \end{cases}$$

The test functions $\{\psi^{ij}(x, t_j) : i = 1, \ldots, M - 1\}$ satisfy

$$-\varepsilon\psi_{xx}^{ij}(x, t_j) - \bar{b}(x, t_j)\psi_x^{ij}(x, t_j) = 0 \quad \text{for } x \in \cup_{i=1}^M (x_{i-1}, x_i), \qquad (4.9a)$$

$$\psi^{ij}(x_k, t_j) = \delta_{ik} \quad \text{for } k = 0, \ldots, M. \qquad (4.9b)$$

Each $\psi^{ij}(\cdot, t_j)$ has support $[x_{i-1}, x_{i+1}]$. These functions are similar to the L^*-splines of Section I.2.2.5.

We introduce a parameter θ whose value determines the approximation of y_t. The values $\theta = 0 \,/\, 0.5 \,/\, 1$ correspond to forward Euler / Crank-Nicolson / backward Euler differencing respectively.

Set

$$(y, z)_j = \int_{x=0}^{1} y(x, t_j) z(x, t_j) \, dx$$

for all piecewise continuous $y(\cdot, t_j)$ and $z(\cdot, t_j)$. Our approximation of (4.1) is

$$(1 - \theta)\{h^{-1}[\varepsilon((u_{h,\tau})_x, \psi_x^{ij})_j + (\bar{b}_j(u_{h,\tau})_x, \psi^{ij})_j] - d_i^j u_i^j\}$$
$$+ \theta\{h^{-1}[\varepsilon((u_{h,\tau})_x, \psi_x^{i,j+1})_{j+1} + (\bar{b}_{j+1}(u_{h,\tau})_x, \psi^{i,j+1})_{j+1}] - d_i^{j+1} u_i^{j+1}\}$$
$$+ (u_i^{j+1} - u_i^j)/\tau = (1 - \theta) f_i^j + \theta f_i^{j+1}, \tag{4.10}$$

for $i = 1, \ldots, M - 1$ and $j = 0, \ldots, N - 1$, where $(\cdot, \cdot)_n$ is the $L_2(0, 1)$ inner product for $t = t_n$, and the computed solution

$$u_{h,\tau}(x, t_j) = \sum_{i=0}^{M-1} u_i^j \phi^{ij}(x, t_j) \quad \text{for } j = 0, \ldots, N$$

is required to interpolate to the initial condition (4.1b).

The one-step difference scheme equivalent to (4.10) is given explicitly in [NSOS88, SO89]. The nodal values u_i^{j+1} at each time level t_{j+1} are computed cheaply from the solution at time t_j by simple tridiagonal Gaussian decomposition (in fact when $\theta = 0$, the scheme is explicit).

If

$$\tau(1 - \theta) \left\{ \frac{4\|b\|_{L_\infty(Q)}}{h(1 - e^{-\beta h/\varepsilon})} + \|d\|_{L_\infty(Q)} \right\} \le 1, \tag{4.11}$$

then the scheme satisfies the hypotheses of Lemma 3.12 and hence a discrete maximum principle. Inequality (4.11) is a generalized CFL condition.   ♣

Theorem 4.4. *Suppose that (4.11) holds true. Assume also that, for $k \le 1$ and $k + m \le 2$,*

$$\left| \frac{\partial^{k+m} u(x, t)}{\partial x^k \partial t^m} \right| \le C(1 + \varepsilon^{-k} e^{-\beta(1-x)/\varepsilon}) \quad \text{for all } (x, t) \in Q. \tag{4.12}$$

Then for all i and j, the solution $u_{h,\tau}$ of (4.10) satisfies

$$|u(x_i, t_j) - u_i^j| \le C(h + \tau). \tag{4.13}$$

If $\theta = 1/2$ and (4.12) also holds true for higher-order derivatives of u, then one can improve (4.13) to

$$|u(x_i, t_j) - u_i^j| \le C(h + \tau^2).$$

Proof. A consistency and stability argument of finite difference flavour [SO89] establishes the result. □

If the data of (4.1) are smooth and compatible, then by Remark 2.8 inequality (4.12) will hold true – so interior layers are excluded – and Theorem 4.4 now guarantees that (4.10) is a uniformly convergent method.

Example 4.5. (A nonlumped scheme) Consider now a semidiscrete formulation, as described in Section 4.1, with piecewise linear trial functions ϕ_i and test functions ψ_i that are defined essentially by (4.9). Replacing terms of the form $(u_h)_t(x, t_j)$ by

$$\frac{u_h(x, t_{j+1}) - u_h(x, t_j)}{\tau},$$

and applying the quadrature rule

$$(u_h, \psi_i)_m \approx u_h(x_i, t^m)(1, \psi_i)_m = hu_i^m \tag{4.14}$$

for each i and m, we can derive (4.10).

The simplification (4.14) is called *mass lumping*. In [NSOS88] the left-hand side of (4.14) is retained, producing a *nonlumped* scheme that is a variant of (4.10), and an error bound like (4.13) is proved for this scheme under the hypotheses that $\theta = 1$, $b\tau/h \geq 1$ and (4.12) all hold true. ♣

The remaining examples are uniformly convergent in the sense of L_2 or energy norms, unlike the L_∞ setting of (3.27).

Example 4.6. (L_2 convergence) In [GS93] two lumped schemes for (4.1) are examined on fairly general tensor-product meshes. If b is constant and the mesh is equidistant, then these schemes are identical to (4.10) with $\theta = 1$, and the computed solution $\{u_i^j\}$ satisfies

$$\left\{ h \sum_{i=1}^{M-1} \left(u(x_i, t_j) - u_i^j \right)^2 \right\}^{1/2}$$

$$\leq Ch^{1/2} \left\{ \tau \sum_{n=1}^{N} \left(\int_{s=0}^{1} \left(|u_x(s, t_n)| + |(f - u_t - bu)_x(s, t_n)| \right) ds \right)^2 \right\}^{1/2}$$

$$+ C\tau^{1/2} \left\{ h \sum_{i=1}^{M-1} \left(\int_{y=0}^{T} |u_{tt}(x_i, y)| \, dy \right)^2 \right\}^{1/2} \tag{4.15}$$

for $j = 1, \ldots, N$. For each j, this is a discrete L_2-norm error estimate. So far, unlike Theorem 4.4, no assumptions have been made regarding the behaviour of u.

In practice each integral in (4.15) may be bounded, uniformly in ε, by a fixed constant. Then we get the uniform L_2 convergence bound

$$\left\{ h \sum_{i=1}^{M-1} \left(u(x_i, t_j) - u_i^j \right)^2 \right\}^{1/2} \leq C(h^{1/2} + \tau^{1/2}) \tag{4.16}$$

for $j = 1, \ldots, N$. If we assume that (4.12) holds true, then the right-hand side of (4.16) can be sharpened [GS93] to $C(h + \tau)$, for $j = 1, \ldots, N$.

A variant of the nonlumped scheme of Example 4.5 is analysed in [GS94] and L_2-convergence results similar to those just described are proved. ♣

For all $w \in H_0^1(0,1)$, define the energy norm

$$\|w\|_{1,\varepsilon} = \left\{ \int_{x=0}^1 [\varepsilon(w'(x))^2 + w^2(x)] \, dx \right\}^{1/2}.$$

Example 4.7. (Energy norm convergence) In (4.1) suppose that $b = b(x)$, $d = d(x)$, $f = f(x)$ and $s \equiv 0$. We generate a semidiscrete solution of the form (4.3) by means of a Galerkin approach where the trial space is "enriched" [HK82] by the insertion of a boundary layer function.

Let $\phi_i(x)$ be the usual piecewise linear function, with $\phi_i(x_j) = \delta_{ij}$, for $i = 1, \ldots, M-1$ and $j = 0, \ldots, M$. Also define

$$\phi_M(x) = e^{-b(1)(1-x)/\varepsilon} - 1 - (1-x)(e^{-b(1)/\varepsilon} - 1) \qquad (4.17)$$

for $x \in [0,1]$. All these $\phi_i(x)$ vanish at $x = 0$ and $x = 1$. Define the trial space V to be the span of $\{\phi_i : i = 0, \ldots, M\}$ and choose the test space to be V also. Then the semidiscrete solution

$$u_h(x,t) = \sum_{i=1}^M u_i(t)\phi_i(x)$$

is required to satisfy (4.4): for each $t \in (0,T]$, and $i = 1, \ldots, M$,

$$(\varepsilon(u_h)_x, \phi_i') + ((u_h)_t + b(u_h)_x + du_h, \phi_i) = 0. \qquad (4.18)$$

Set $\zeta = u - u_h$. Now (4.1) and (4.18) imply that

$$(\varepsilon\zeta_x, \phi') + (\zeta_t + b\zeta_x + d\zeta, \phi) = 0$$

for all $\phi \in V$. Hence

$$(\zeta_t, \zeta) + (\varepsilon\zeta_x, \zeta_x) + (b\zeta_x + d\zeta, \zeta)$$
$$= (\varepsilon\zeta_x, (\zeta - \phi)_x) + (b\zeta_x + d\zeta + \zeta_t, \zeta - \phi). \qquad (4.19)$$

But an integration by parts and (4.2) yield

$$(\varepsilon\zeta_x, \zeta_x) + (b\zeta_x + d\zeta, \zeta) \geq \min\{1, \omega\} \|\zeta(\cdot, t)\|_{1,\varepsilon}^2 \qquad (4.20)$$

for each $t \in (0,T]$. Substitute (4.20) into (4.19), then integrate in time to get

$$\int_{x=0}^1 \frac{1}{2}\zeta^2(x,t) \, dx + \min\{1, \omega\} \int_{s=0}^t \|\zeta(\cdot, t)\|_{1,\varepsilon}^2 \, ds$$
$$\leq \int_{s=0}^t |(\varepsilon\zeta_x, (\zeta - \phi)_x) + (b\zeta_x + d\zeta + \zeta_t, \zeta - \phi)| \, ds \qquad (4.21)$$

for $0 \leq t \leq T$, where each inner product is evaluated at time s, and $\phi \in V$ is arbitrary.

To deduce an energy norm error bound from (4.21), one needs adequate approximation theory estimates for the difference $\zeta - \phi$. In [HK82] smoothness and compatibility conditions are assumed for the data and the solution u is decomposed in an asymptotic expansion. This leads to the bound

$$\int_{x=0}^{1} \frac{1}{2} \zeta^2(x,t) \, dx + \min\{1, \omega\} \int_{s=0}^{t} \|\zeta(\cdot, t)\|_{1,\varepsilon}^2 \leq C h^{3/2}, \qquad (4.22)$$

which is uniform in ε.   ♣

In [YJS99] the bounds (2.20) are assumed and an arbitrary tensor product grid is used with nodes (x_i, t_j) for $i = 0, \ldots, M$ and $j = 0, \ldots, N$. Set $\tau_j = t_j - t_{j-1}$ for each j and $h = \max_i(x_i - x_{i-1})$. The method bears some resemblance to Example 4.3; compare the discussion of Petrov-Galerkin methods in Section I.2.2.5. To discretize in space, piecewise linear test functions are used; the trial functions are piecewise linear outside the "layer region" (i.e., for those x_i such that $x_i \leq 1 - (2\varepsilon/\beta)|\ln \varepsilon|$) and are $\bar{L}$-splines (as in (4.9), but with the sign of $\bar{b}$ changed) inside the layer region. Backward Euler differencing is used to approximate u_t. It is proved that

$$\sum_{j=1}^{N} \tau_j \|u^j - U^j\|_{1,\varepsilon,d}^2 + \max_{j=0,\ldots,N} \|u^j - U^j\|_{0,d} \leq C(\tau^2 + h^2 |\ln h|),$$

where $w^j(x) := w(x, t^j)$ for each function $w \in C(\bar{\Omega})$, U is the computed solution, $\|\cdot\|_{0,d}$ is a discrete $L_2[0,1]$ norm and $\|\cdot\|_{1,\varepsilon,d}$ is the energy norm $\|\cdot\|_{1,\varepsilon}$ with its L_2 component replaced by $\|\cdot\|_{0,d}$.

Finally, a uniform convergence estimate will be described in (4.60) for the case where the boundary data in (4.1) is periodic, which excludes a boundary layer.

4.1.3 Local Error Estimates

All proofs that methods are uniformly convergent assume a bound such as (4.12) on the derivatives of the solution. This is a strong assumption that often fails to be satisfied in realistic problems; in particular, it excludes interior layers. To operate in a more practical framework, yet still develop proofs that numerical methods give accurate results even when ε is very small, we replace the target of global uniform convergence by a less demanding objective: the convergence of each method only at those nodes that lie outside layers.

That is, *local error estimates* will examined on regions in Q where u is "smooth": the solution u of (4.1) is said to be *smooth on a subdomain Q' of Q* if *a certain number of its derivatives are bounded, uniformly in ε, on Q'*. This terminology is commonly used in the literature. When dealing with functions that do not depend on ε, we use "smooth" in its classical sense of having sufficient regularity.

Our analysis must somehow distinguish nodes "inside" the layers from those "outside". This is done using cut-off functions as in the local analysis of the streamline diffusion method in Section III.3.2.1. We omit the details and content ourselves here with a statement of the results obtained.

Example 4.8. Two schemes from [GS93] were discussed in Example 4.6. The following local error estimate holds true for each of these schemes.

Suppose that $b \equiv d \equiv 1$ on Q and the mesh is equidistant. Let (x_i, t_j) be a node in Q. Define Q_1 by

$$Q_1 := \{(x, t) \in Q : 0 \leq x \leq x_i + C_1 \varepsilon^* |\ln(h\tau)|,$$
$$|x - t - (x_i - t_j)| \leq C_1 \sqrt{\varepsilon^*} \, |\ln(h\tau)|\}, \qquad (4.23)$$

where $\varepsilon^* = \max\{\varepsilon, h, \tau\}$ and C_1 is a constant chosen in the proof. The subdomain Q_1 is (see Figure 4.1) a long thin region centred on the subcharacteristic $t = x$ through (x_i, t_j) and extending from the inflow boundary of Q to slightly downstream of (x_i, t_j).
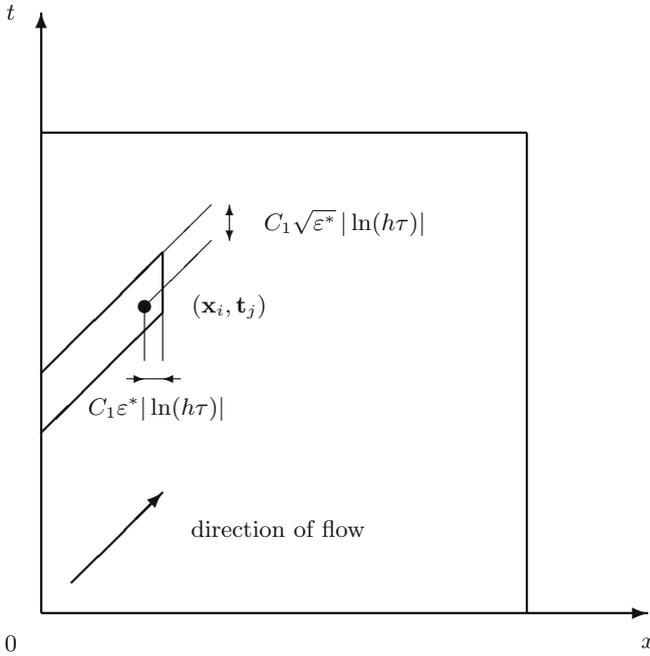

Fig. 4.1. The region Q_1 of Example 4.8

Assume that
$$\|u\|_{C^2(Q_1)} + \|f\|_{C^1(Q_1)} \leq C, \qquad (4.24)$$

where

$$\|w\|_{C^k(Q_1)} := \max\left\{\left\|\frac{\partial^{\ell+m}w}{\partial x^\ell \partial t^m}\right\|_{L_\infty(Q_1)} : \ell + m \le k\right\}.$$

In other words, we assume that (x_i, t_j) is not close to any layer. Assume also that there exists a constant C such that

$$\|s\|_{L_\infty(0,1)} + \tau \sum_{j=1}^N \left[\|u(\cdot, t_j)\|_{L_\infty(0,1)} + \|f(\cdot, t_j)\|_{L_\infty(0,1)}\right]$$

$$+ \tau^2 \sum_{j=1}^N \|u_t(\cdot, t_j)\|_{L_1(0,1)} \le C. \qquad (4.25)$$

One can reasonably expect (4.25) to hold true in many practical problems, even when interior layers are present.

Then, writing u_i^j for the solution computed by either of the schemes from [GS93], from that paper we have the local error estimate

$$|u(x_i, t_j) - u_i^j| \le C(h + \tau). \qquad (4.26)$$

This bound shows that the schemes are convergent away from layers. Furthermore, the schemes are also convergent at any node inside the boundary layer, if the boundary layer has "typical" behaviour in a neighbourhood of that node. For replace the bound on u in (4.24) by the assumption that, for $k \le 1$ and $k + m \le 2$,

$$\left|\frac{\partial^{k+m}u(x,t)}{\partial x^k \partial t^m}\right| \le C(1 + \varepsilon^{-k}e^{-\beta(1-x)/\varepsilon}) \quad \text{for all } (x,t) \in Q_1. \qquad (4.27)$$

(This is (4.12) restricted to Q_1.) Assume again that (4.25) holds true. Then (4.26) is still valid.

Corresponding results hold good for the nonlumped scheme of [GS94]. ♣

Remark 4.9. The property of convergence, uniformly in ε, on regions where u is smooth is much less stringent than uniform convergence on all of Q. For example, for two-point boundary value problems, Theorem I.2.12 shows that the upwind scheme (2.12) of Part I is first-order convergent, uniformly in ε, at all nodes sufficiently far from the boundary layer at $x = 1$, while Theorem I.2.17 implies that this method fails to be uniformly convergent on [0,1].

The proof of Theorem 3.19 relies on following the scheme into the boundary layer; it fails for methods that achieve uniform in ε convergence only where u is smooth. ♣

4.2 Subcharacteristic-Based Methods

The approach of Section 4.1 essentially splits the operator L into $(\cdot)_t$ and $-\varepsilon(\cdot)_{xx} + b(\cdot)_x + d(\cdot)$. In this section we shall develop several finite element

methods that are based on a decomposition of L into $-\varepsilon (\cdot)_{xx}$ and $(\cdot)_t + b (\cdot)_x + d (\cdot)$. This choice of operator splitting implies that space-time finite elements must be used. Furthermore, the subcharacteristics of (4.1) will play a significant role.

Throughout Section 4.2, set

$$w_z := w_t + bw_x$$

for various functions w. We call w_z the *material derivative* of w.

This section is divided into three subsections, beginning with the streamline diffusion finite element method in Section 4.2.1. While the test functions in this method are related to the material derivative, the subcharacteristics themselves are not explicitly used. The discontinuous Galerkin, continuous Reed-Hill-Richter, and Eulerian-Lagrangian methods of the remaining two subsections do however require some knowledge of the subcharacteristics of (4.1).

4.2.1 SDFEM in Space-Time

Standard Galerkin methods that use continuous space-time basis functions are usually unstable when applied to (4.1). They can be stabilized by artificially increasing the diffusion coefficient ε in (4.1a), but the computed solution will then betray the fact that we have modified (4.1) as its layers will be excessively diffuse.

The streamline diffusion finite element method (SDFEM), which is also known as the *streamline upwinding Petrov-Galerkin method (SUPG)*, achieves stability by adding diffusion to the problem only in the direction of the subcharacteristics of (4.1). This produces much less diffusion in the numerical solution than the artificial diffusion method of the previous paragraph. Moreover, the increased diffusion in the method is generated in a credible manner using a Petrov-Galerkin finite element framework that is described below. The approach here is of course related to the investigation of the SDFEM for two-point boundary value problems in Section I.2.2.3.

To discretize (4.1) one could apply the SDFEM in the spatial variable, combined with an arbitrary discretization in time, but we shall not discuss this approach here; see Remarks 4.2 and III.4.3.

Partition $[0, T]$ by an equidistant mesh $\{t_j : j = 0, \dots, N\}$, where for each j we set $t_j = jT/N = j\tau$. Our trial functions will be continuous on each space-time strip

$$Q_j := [0, 1] \times (t_{j-1}, t_j).$$

This avoids possible technical difficulties in the interpretation of $-\varepsilon v_{xx}$ when v is a trial function.

One could work with trial functions that were continuous on Q, but this would lead later to multi-step schemes. Consequently we allow the trial functions to be discontinuous across each line $t = t_j$.

In general, one uses a space of standard finite element functions on an arbitrary subdivision of each Q_j into triangles or quadrilaterals. Here we shall consider a simple specific case. Divide [0,1] by the equidistant mesh $\{x_i : i = 0, \ldots, M\}$, with $x_i = i/M = ih$ for each i. Then partition each strip Q_j by the lines $t = t_{j-1}$, $t = t_j$ and $x = x_i$ for $i = 0, \ldots, M$, together with each northwest to southeast diagonal of the rectangles so formed. This gives a uniform structured triangulation of Q_j.

Let V_j denote the space of standard piecewise linear functions on this triangulation of Q_j that vanish at $x = 0$ and $x = 1$. Our solution $u_{h,\tau}$ will satisfy $u_{h,\tau}|_{Q_j} \in V_j$ for each j.

So far, we have described a typical space-time finite element space whose functions are continuous in space but possibly discontinuous in time as one moves from one strip Q_j to the next. The use of such spaces is not peculiar to the streamline diffusion method.

Set $\hat{u}^j = u_{h,\tau}|_{Q_j}$. Each $\hat{u}^j \in V_j$ is defined by a Petrov-Galerkin method. In this method, each test function is constructed from a corresponding trial function ϕ by the mapping

$$\phi \mapsto \phi + \delta\phi_z, \tag{4.28}$$

where δ is a sufficiently small positive constant (see below).

Thus in (4.1a) we change u to $\hat{u}^j$, then multiply both sides of the equation by $\phi + \delta\phi_z$, where ϕ is an arbitrary member of V_j, and finally integrate over Q_j. This gives

$$\varepsilon(\hat{u}^j_x, \phi_x)_{Q_j} - \varepsilon \sum_{T \subset Q_j} (\hat{u}^j_{xx}, \delta\phi_z)_T + (\hat{u}^j_z + d\hat{u}^j, \phi + \delta\phi_z)_{Q_j}$$
$$= (f, \phi + \delta\phi_z)_{Q_j}, \tag{4.29}$$

where $(\cdot, \cdot)_G$ is the $L_2(G)$ inner product for any measurable $G \subset \bar{Q}$, each T is an open triangle in Q_j, and the term $-\varepsilon(\hat{u}^j_{xx}, \phi)_{Q_j}$ was integrated by parts.

In the present case, where $\hat{u}^j$ is piecewise linear, the contribution from $\hat{u}^j_{xx}$ to (4.29) is zero.

Of course (4.29) alone cannot determine $\hat{u}^j$; some initial condition is needed at $t = t_{j-1}$. As the solution may be discontinuous across $t = t_{j-1}$, the information provided by $\hat{u}^{j-1}$ can be used only in a weak sense. This procedure is also standard in the discontinuous Galerkin method [Joh87, Section 8.4.3].

Thus modify (4.29) to the complete *streamline diffusion formulation*:

$$\varepsilon(\hat{u}^j_x, \phi_x)_{Q_j} - \varepsilon \sum_{T \subset Q_j} (\hat{u}^j_{xx}, \delta\phi_z)_T + (\hat{u}^j_z + d\hat{u}^j, \phi + \delta\phi_z)_{Q_j} + \langle \hat{u}^j_+, \phi_+ \rangle_{j-1}$$
$$= (f, \phi + \delta\phi_z)_{Q_j} + \langle \hat{u}^{j-1}_-, \phi_+ \rangle_{j-1}, \tag{4.30}$$

for $j = 1, \ldots, N$. Here $\phi \in V_j$ is arbitrary, the $L_2(0,1)$ inner product for $t = t_{j-1}$ is denoted by $\langle \cdot, \cdot \rangle_{j-1}$, and

$$w_{\pm}(x, t_{j-1}) := \lim_{k \to 0^+} w(x, t_{j-1} \pm k)$$

for $x \in [0, 1]$. We replace $\langle \hat{u}_-^{j-1}, \phi_+ \rangle_{j-1}$ in (4.30) by $\langle s, \phi_+ \rangle_0$ when $j = 1$.

One can write (4.30) as a linear system of equations in the unknowns $\hat{u}_-^j(x_i, t_j)$, where $i = 1, \ldots, M-1$. For let $\phi_{m,k}$ be the canonical basis function in V_j that satisfies

$$\phi_{m,k}(x_i, t_\ell) = \delta_{mi} \delta_{k\ell} \quad \text{for } i = 0, \ldots, M \text{ and } \ell = j - 1, j. \tag{4.31}$$

Take $\phi = \phi_{i,j-1}$ in (4.30). On considering the support of $\phi_{i,j-1}$ in Q_j, we find that (4.30) yields a linear equation in

$$(\hat{u}_-^{j-1})_{i-1}^{j-1}, \ (\hat{u}_-^{j-1})_i^{j-1}, \ (\hat{u}_-^{j-1})_{i+1}^{j-1}, \ (\hat{u}_+^j)_{i-1}^{j-1}, \ (\hat{u}_+^j)_i^{j-1}, \ (\hat{u}_+^j)_{i+1}^{j-1},$$
$$(\hat{u}_-^j)_{i-1}^j, \ \text{and} \ (\hat{u}_-^j)_i^j.$$

Here, for example,

$$(\hat{u}_-^{j-1})_{i-1}^{j-1} := \hat{u}_-^{j-1}(x_{i-1}, t_{j-1}).$$

Of the above eight quantities, the values of the $(\hat{u}_-^{j-1})_{(\cdot)}^{j-1}$ are known from the previous time step. The $(\hat{u}_-^j)_{(\cdot)}^j$ will be computed as the numerical solution advances forward in time. The $(\hat{u}_+^j)_{(\cdot)}^{j-1}$ are of relatively minor interest and we now eliminate them.

In turn take $\phi = \phi_{i,j-1}, \phi_{i+1,j-1}, \phi_{i-1,j}$, and $\phi_{i+1,j}$ in (4.34). These four equations can be combined to eliminate the intrusive terms $(\hat{u}_+^j)_m^{j-1}$, where $i - 1 \le m \le i + 2$.

Example 4.10. Suppose that $b \equiv 1$, $d \equiv 0$, $\delta = h$, and we set $\varepsilon = 0$. Then the resulting difference scheme is [Joh87, Näv82]

$$\left(\frac{1}{4} + \frac{2\nu}{3} + \frac{4\nu^2}{9} \right) u_{i-2}^{j+1} + \left(\frac{1}{6\nu} - \frac{23}{18} - \frac{101\nu}{36} - \frac{49\nu^2}{36} \right) u_{i-1}^{j+1}$$
$$+ \left(\frac{2}{3\nu} + \frac{9}{4} + \frac{63\nu}{18} + \frac{3\nu^2}{2} \right) u_i^{j+1} + \left(\frac{1}{6\nu} - \frac{5}{6} - \frac{57\nu}{36} - \frac{25\nu^2}{36} \right) u_{i+1}^{j+1}$$
$$+ \left(\frac{1}{9} + \frac{2\nu}{9} + \frac{\nu^2}{9} \right) u_{i+2}^{j+1}$$
$$= \left(\frac{1}{6\nu} + \frac{2}{9} + \frac{\nu}{36} \right) u_{i-1}^j + \left(\frac{2}{3\nu} + \frac{3}{4} + \frac{\nu}{12} \right) u_i^j$$
$$+ \left(\frac{1}{6\nu} - \frac{1}{3} - \frac{\nu}{12} \right) u_{i+1}^j - \left(\frac{5}{36} + \frac{\nu}{36} \right) u_{i+2}^j,$$

where $\nu = \tau/h$ and $u_i^j := (\hat{u}_-^j)_i^j$.   ♣

Does (4.30) have a unique solution? Write its left-hand side as $a(\hat{u}^j, \phi)$. Then for arbitrary $\phi \in V_j$, one has

$$a(\phi, \phi) = \varepsilon \|\phi_x\|^2_{L_2(Q_j)} + (\phi_z + d\phi, \phi)_{Q_j} + \delta \|\phi_z\|^2_{L_2(Q_j)}$$
$$+ \delta(d\phi, \phi_z)_{Q_j} + \langle \phi_+, \phi_+ \rangle_{j-1}$$
$$\geq \varepsilon \|\phi_x\|^2_{L_2(Q_j)} + \omega \|\phi\|^2_{L_2(Q_j)} + \frac{1}{2} \langle \phi_-, \phi_- \rangle_j + \delta \|\phi_z\|^2_{L_2(Q_j)}$$
$$+ \delta(d\phi, \phi_z)_{Q_j} + \frac{1}{2} \langle \phi_+, \phi_+ \rangle_{j-1},$$

where we integrated by parts and used (4.2). But

$$|\delta(d\phi, \phi_z)_{Q_j}| \leq \frac{\delta}{2} \|d\|^2_{L_\infty(Q_j)} \|\phi\|^2_{L_2(Q_j)} + \frac{\delta}{2} \|\phi_z\|^2_{L_2(Q_j)}.$$

Hence, if $\delta \leq \omega / \|d\|^2_{L_\infty(Q_j)}$, then

$$a(\phi, \phi) \geq \varepsilon \|\phi_x\|^2_{L_2(Q_j)} + \frac{\delta}{2} \|\phi_z\|^2_{L_2(Q_j)} + \frac{\omega}{2} \|\phi\|^2_{L_2(Q_j)}$$
$$+ \frac{1}{2} \langle \phi_-, \phi_- \rangle_j + \frac{1}{2} \langle \phi_+, \phi_+ \rangle_{j-1}.$$

It follows that (4.30) has a unique solution.

Assume that $\tau \leq Ch$ for some $C > 0$ in the analysis that follows.

For higher order piecewise polynomial trial spaces, the term

$$-\varepsilon \sum_{T \subset Q_j} (\hat{u}^j_{xx}, \delta \phi_z)_T$$

is not zero and must be bounded in the above analysis. This leads to the stability requirement [Joh87, Näv82] that $\delta \leq Ch$ when $\varepsilon < h$. On the other hand, if $h \leq \varepsilon$, then the ordinary Galerkin method is stable so one can take $\delta = 0$. To obtain accurate results using the streamline diffusion method, one must choose δ carefully. See Section III.3.2.1 for a discussion of this choice in the elliptic case.

We next state a simplified version of the global and local error estimates of Närvert [Näv82] for the streamline diffusion solution $u_{h,\tau}$. These estimates are proved using standard but detailed finite element techniques. To prove the local estimates one uses a cut-off function (cf. Section III.3.2.1) and approximation theory arguments.

For each set $\hat{Q}$ that is the closure of a union of open triangles T, and each w that lies in $H^1(T)$ for all $T \subset \hat{Q}$, define

$$|||w|||_{\hat{Q}} := \left\{ \varepsilon \sum_{T \subset \hat{Q}} \|\nabla w\|^2_{L_2(T)} + \sum_{T \subset \hat{Q}} \delta \|w_z\|^2_{L_2(T)} + \|w\|^2_{L_2(\hat{Q})} \right\}^{1/2}.$$

Theorem 4.11. *(Global error bound) Assume that we have $\varepsilon \leq h, \tau \leq Ch$, and $\delta \leq C'h$ for some sufficiently small constant C'. Then there exists a constant C such that for all sufficiently small h (independently of ε),*

$$|||u - u_{h,\tau}|||_Q \leq Ch^{3/2} |u|_{H^2(Q)}. \tag{4.32}$$

Theorem 4.12. *(Local error bound) Assume the hypotheses of Theorem 4.11. Suppose that b is constant. Let $Q_2 \subset Q$ be a union of open mesh triangles. For each mesh node (x_i, t_j) in $\bar{Q}_2$, define Q_2^{ij} (cf. Example 4.8) by*

$$
Q_2^{ij} := \{(x,t) \in Q : x + bt \le x_i + bt_j + C_2 h |\ln h|, \\
|x - bt - (x_i - bt_j)| \le C_2 \sqrt{h} \, |\ln h|\}, \tag{4.33}
$$

where C_2 is a fixed constant chosen in the proof. Set $Q_3 = \cup_{(x_i, t_j) \in Q_2} Q_2^{ij}$. Then for any union of mesh triangles Q_4 for which $Q_3 \subset \bar{Q}_4$, there exists a constant C such that

$$
|||u - u_{h,\tau}|||_{Q_2} \le C \{ h^{3/2} |u|_{H^2(Q_4)} + h^2 [\|f\|_{L_2(Q)} + \|s\|_{L_2(0,1)}] \}. \tag{4.34}
$$

In Theorem 4.12, the region Q_2^{ij} has width $2 C_2 \sqrt{h} \, |\ln h|$ about the subcharacteristic through (x_i, t_j), and extends from the upstream boundary of Q to $\mathcal{O}(h |\ln h|)$ downstream of (x_i, t_j).

Global bounds such as (4.32) do not guarantee that $u_{h,\tau}$ is close to u, since typically $|u|_{H^2(Q)}$ is $\mathcal{O}(\varepsilon^{-3/2})$ (recall Theorem 2.6) and in practice $\varepsilon < h$. The local bound (4.34) is much more useful, but also harder to prove. It ensures that $u_{h,\tau}$ is an $\mathcal{O}(h^{3/2})$ L_2 approximation to u on regions in Q that are not "near" layers. Numerical experiments in [Näv82], for a problem whose subcharacteristics are straight lines, show that $\|u - u_{h,\tau}\|_{L_2}$ is $\mathcal{O}(h^2)$ away from layers, but it is known that for elliptic problems with smooth solutions one sometimes achieves only $\mathcal{O}(h^{3/2})$ accuracy in L_2 [Zho97].

SDFEM on a Layer-adapted Mesh

It seems natural to combine the SDFEM with some layer-adapted mesh yet the literature contains only one paper [GS97] that analyses this combination for (4.1). Its contents will now be described.

Theorem 3.19 implies that the streamline diffusion method cannot be uniformly convergent on the rectangular equidistant mesh of Section 3.2. In particular it fails to converge inside layers. We can alleviate this situation inside the outflow boundary layer by replacing our original equidistant mesh by the following Shishkin mesh, which is similar to the mesh of Section 3.4.2.

Assume that $d \equiv 1$ and that b is a positive constant. Let the mesh lines $t = t_j$ be equidistantly spaced. For the x-direction, assume that $\varepsilon \le M^{-1}$ and that M is even. Set $\sigma = 2b^{-1} \varepsilon \ln M$; the parameter σ is the distance from the boundary $x = 1$ at which the mesh switches from coarse to fine. Let

$$
x_i = \begin{cases} 2i(1-\sigma)M^{-1} & \text{for } i = 0, \ldots, M/2, \\ 1 - \sigma + 2\sigma(i - M/2)M^{-1} & \text{for } i = M/2 + 1, \ldots, M. \end{cases}
$$

Construct a triangulation of Q as before, based on the points (x_i, t_j), and use the SDFEM scheme (4.30). In the next result, recall that all constants C are generic.

Theorem 4.13. *[GS97] Assume that*

$$\|f\|_{L_2(Q)} + \|s\|_{L_2(0,1)} + \|u_t + bu_x\|_{L_1(Q)} + \varepsilon\|u_{xx}\|_{L_1(Q)} \le C. \qquad (4.35)$$

Assume also that $\max\{M/N, N/M\} \le C$. *Let* (x_i, t_j) *be a node in* Q. *Set*

$$Q_3 := \big\{(x,t) \in Q : x \le x_i + C_3 M^{-1} \ln M,$$
$$|x - bt - (x_i - bt_j)| \le C_3 M^{-1/2} \ln M \big\},$$

where C_3 *is a constant chosen in the proof.*

If $\|u\|_{C^2(Q_3)} \le C$, *where* $\|\cdot\|_{C^2(Q_3)}$ *is defined as in Example 4.8, then*

$$|u(x_i, t_j) - u_i^j| \le CM^{-5/4} \ln M. \qquad (4.36)$$

If, for $k + m \le 2$,

$$\left|\frac{\partial^{k+m} u(x,t)}{\partial x^k \partial t^m}\right| \le C(1 + \varepsilon^{-k} e^{-\beta x/\varepsilon}) \quad \text{for all } (x,t) \in Q_3,$$

then

$$|u(x_i, t_j) - u_i^j| \le CM^{-3/4} \ln^{5/2} M. \qquad (4.37)$$

The condition (4.35) is frequently true in practice and can indeed sometimes be proved. The estimate (4.36) is useful away from layers, while (4.37) indicates that the method converges inside typical parts of the boundary layer. Numerical results in [GS97] indicate that the method converges at a rate of approximately $\mathcal{O}(M^{-0.8})$ inside a representative boundary layer.

When considering elliptic problems in Section III.3.2.1, it will be shown that local error estimates such as (4.36) can be improved by the addition of a judicious amount of artificial crosswind diffusion, and for certain meshes are close to $\mathcal{O}(M^{-2})$.

Several variants of the SDFEM are compared in [HD05] via a Fourier stability analysis and a computation of phase error, and many references to the literature are given.

We shall continue the discussion of the streamline diffusion method for parabolic problems in Chapter 5, where adaptive implementations of two of its variants are discussed.

4.2.2 Explicit Galerkin Methods

In this Section we examine two methods originally due to Reed and Hill [RH73]. The first is a Galerkin method that generates a discontinuous approximation to u, the other a Petrov-Galerkin method that yields a continuous approximation. The methods were originally designed for first-order problems such as the reduced problem (2.5) and were later extended to convection-diffusion problems by Richter [Ric90, Ric92].

We begin with a version of the *discontinuous Galerkin method* that uses space-time elements for parabolic problems. The discontinuous Galerkin method for elliptic problems that will be discussed at length in Section III.3.4 incorporates several extra features that do not appear in the present section.

Consider the subdivision of Q by a family $\mathcal{F}$ of quasi-uniform triangular meshes. (Quasi-uniform means that there exists positive constants C_0 and C_1 such that the minimum angle of each triangle is bounded below by C_0 for all meshes in $\mathcal{F}$ and on each mesh in $\mathcal{F}$ the ratio of maximum triangle diameter to minimum triangle diameter is bounded above by C_1.) Let k be a non-negative integer. For each triangle T, let $P_k(T)$ denote the space of polynomials of degree at most k defined on T.

Our discontinuous Galerkin method computes a piecewise continuous solution on successive triangles by following the subcharacteristics of (4.1a). The computed solution $u_{h,\tau}$ will satisfy $u_{h,\tau}|_T \in P_k(T)$ for each T. This solution is not necessarily continuous on Q; consequently the total number of degrees of freedom is larger than for functions in $C(Q)$ that are piecewise polynomials.

Assume that $d \equiv 0$ and b is constant. Let $z = (b, 1)$ be a vector parallel to the subcharacteristic direction. Assume that no triangle side lies on or near the direction z, i.e., that

$$|z \cdot n| \geq K \text{ for some } K > 0, \tag{4.38}$$

where n represents all possible unit vectors perpendicular to triangle sides.

With respect to the reduced problem (2.5), Lesaint and Raviart [LR74] show that the triangles in the mesh can always be explicitly ordered in such a way that the domain of dependence of each triangle is contained in the union of all earlier triangles with the inflow boundary $S_0 \cup S_x$.

Using this ordering, the solution $u_{h,\tau}$ for (4.1) will be computed triangle by triangle. Begin by choosing $u_{h,\tau}|_{S_0 \cup S_x}$ as the interpolant to the initial-boundary conditions.

Let T be a typical triangle. We require $u_{h,\tau} \in P_k(T)$ to satisfy

$$(-\varepsilon(u_{h,\tau})_{xx} + (u_{h,\tau})_z, \psi)_T - \int_{\partial_- T} \left[(u_{h,\tau})_+ - (u_{h,\tau})_-\right] \psi \, z.n \, ds$$

$$+ \varepsilon \int_{\partial_- T'} \left[((u_{h,\tau})_x)_+ - ((u_{h,\tau})_x)_-\right] \psi n_1 \, ds = (f, \psi)_T, \tag{4.39}$$

for all $\psi \in P_k(T)$. Here $(\cdot, \cdot)_T$ is the $L_2(T)$ inner product, $\partial_- T$ is the inflow boundary of T, $\partial_- T' = \partial_- T \setminus (S_0 \cup S_x)$, and $n = (n_1, n_2)$ is the outward-pointing unit normal on the boundary of T. The functions w_+ and w_- are the downstream and upstream limits respectively of w on $\partial_- T$, viz.,

$$w_\pm(p) := \lim_{k \to 0^+} w(p \pm kz),$$

for every $p \in \partial_- T$. All boundary integrals are with respect to arclength. If we set $\varepsilon = 0$ in (4.39), we recover the original discontinuous Galerkin method for (2.5).

This algorithm for solving (4.1) does not use any data from the outflow boundary S_1, so its solution cannot exhibit boundary layers.

Example 4.14. Take $k = 1$. Recall the triangulation of Q used in Section 4.2.1. We order the triangles by moving along each strip Q_j from left to right, starting with Q_0, then continuing with $Q_1, Q_2, \ldots, Q_{N-1}$.

Let the triangle T have vertices $(x_{i-1}, t_{j-1}), (x_i, t_{j-1})$ and (x_{i-1}, t_j). Denote by u_m^ℓ nodal values of $u_{h,\tau}|_T$, by $(u_L)_m^\ell$ nodal values on the triangle to the left of T, and by $(u_B)_m^\ell$ nodal values on the triangle below T.

We take in turn $\psi = \phi_{i-1,j-1}, \phi_{i,j-1}$, and $\phi_{i-1,j}$ in (4.39), where each $\phi_{m,k}$ is defined in (4.31). This gives three independent equations. For constant f, these equations are

$$\frac{b}{h}(u_i^{j-1} - u_{i-1}^{j-1}) + \frac{1}{\tau}(u_{i-1}^j - u_{i-1}^{j-1}) + \frac{b}{h}[2u_{i-1}^{j-1} - 2(u_L)_{i-1}^{j-1} + u_{i-1}^j - (u_L)_{i-1}^j]$$
$$+ \frac{1}{\tau}[2u_{i-1}^{j-1} - 2(u_B)_{i-1}^{j-1} + u_i^{j-1} - (u_B)_i^{j-1}]$$
$$- \frac{3\varepsilon}{h^2}[u_i^{j-1} - u_{i-1}^{j-1} + (u_L)_{i-1}^j - (u_L)_{i-2}^j] = f,$$

$$\frac{b}{h}(u_i^{j-1} - u_{i-1}^{j-1}) + \frac{1}{\tau}(u_{i-1}^j - u_{i-1}^{j-1})$$
$$+ \frac{1}{\tau}[2u_i^{j-1} - 2(u_L)_i^{j-1} + u_{i-1}^{j-1} - (u_L)_{i-1}^{j-1}] = f,$$

and

$$\frac{b}{h}(u_i^{j-1} - u_{i-1}^{j-1}) + \frac{1}{\tau}(u_{i-1}^j - u_{i-1}^{j-1}) + \frac{b}{h}[2u_{i-1}^j - 2(u_L)_{i-1}^j + u_{i-1}^{j-1} - (u_L)_{i-1}^{j-1}]$$
$$- \frac{3\varepsilon}{h^2}[u_i^{j-1} - u_{i-1}^{j-1} + (u_L)_{i-1}^j - (u_L)_{i-2}^j] = f.$$

Now take $k = 0$. Applying (4.39) on T and on the triangles immediately to the left of T and below T, then combining these three equations, one obtains the *simple upwind scheme* (cf. Example 3.4)

$$\frac{1}{\tau}(u_{i-1}^j - u_{i-1}^{j-1}) + \frac{b}{h}(u_{i-1}^j - u_{i-2}^j) = f, \tag{4.40}$$

where u_k^m denotes the value of $u_{h,\tau}$ on the triangle with vertices at (x_k, t_m), (x_k, t_{m-1}) and (x_{k+1}, t_{m-1}). ♣

Write $a(u_{h,\tau}, \psi)$ for the left-hand side of (4.39). Richter [Ric92] assumes that $\varepsilon \le C_1 h$ for some constant C_1, and shows by a careful analysis that $a(\cdot, \cdot)$ satisfies a complicated coercivity inequality. Then he deduces the following error estimate for the discontinuous Galerkin method:

Theorem 4.15. *There exists a constant C such that*

$$\|u - u_{h,\tau}\|_{L_2(Q)} + \left(\sqrt{\varepsilon} + h\right) \left\{ \sum_{T \subset Q} \|(u - u_{h,\tau})_x\|_{L_2(T)}^2 \right\}^{1/2}$$

$$+ h^{1/2} \left\{ \sum_{T \subset Q} \|(u - u_{h,\tau})_z\|_{L_2(T)}^2 \right\}^{1/2} \leq Ch^{k+1/2}\|u\|_{H^{k+1}(Q)}, \quad (4.41)$$

where each sum is over all open triangles T in Q.

Remark 4.16. These global error bounds are not convincing at first sight, because usually $\|u\|_{H^{k+1}(Q)}$ is $\mathcal{O}(\varepsilon^{-k-1/2})$. Recall, however, that the solution $u_{h,\tau}$ is computed triangle by triangle; hence (4.41) also holds true on any collection Q' of triangles in Q for which

$$\partial_- Q' \subset \partial_- Q, \quad (4.42)$$

where ∂_- denotes the inflow boundary. ♣

If Q' is chosen so that $\|u\|_{H^{k+1}(Q')} \leq C$, then (4.41) shows that the L_2 error of the derivative of the computed solution in the subcharacteristic direction is of optimal order. If also ε/h is bounded below by a positive constant, then (4.41) yields

$$\left\{ \sum_{T \subset Q'} \|\nabla(u - u_{h,\tau})\|_{L_2(T)}^2 \right\}^{1/2} \leq Ch^k,$$

which is an optimal order bound on the gradient of the error.

When $\|u\|_{H^{k+1}(Q')} \leq C$, the bound

$$\left\{ \sum_{T \subset Q'} \|u - u_{h,\tau}\|_{L_2(T)}^2 \right\}^{1/2} \leq Ch^{k+1/2} \quad (4.43)$$

of (4.41) is order $1/2$ less than optimal. Peterson [Pet91] has shown that (4.43) is the best possible general result, but Richter [Ric88] proves for first-order problems such as (2.5) that on certain triangulations the optimal order is achieved. Numerical results in [Ric92] include a convection-diffusion example for which piecewise linears on a uniform structured mesh yield the optimal rate of convergence, viz.,

$$\|u - u_{h,\tau}\|_{L_2(Q)} \leq Ch^2.$$

We now move on to the *continuous Reed-Hill-Richter method*. This provides a continuous piecewise polynomial numerical solution for (4.1). It constructs this solution triangle by triangle, as in the discontinuous Galerkin

method, but it is a Petrov-Galerkin method. The analyses of the two methods are quite similar.

Once again assume that the mesh is quasiuniform, b is constant, $d \equiv 0$, and (4.38) holds true.

For each triangle T of the mesh, let $m(T)$ denote the number of inflow sides. Inequality (4.38) implies that the value of $m(T)$ – i.e., 1 or 2 – is well defined.

The solution $u_{h,\tau}$ is computed on successive triangles. We require that $u_{h,\tau}|_T \in P_k(T)$ for each T, and that $u_{h,\tau} \in C(Q)$. On $S_0 \cup S_x$, take $u_{h,\tau}$ to be the interpolant to the initial-boundary data.

On each triangle T, the solution satisfies

$$(-\varepsilon(u_{h,\tau})_{xx} + (u_{h,\tau})_z, \psi)_T + \varepsilon \int_{\partial_- T} \left[((u_{h,\tau})_x)_+ - ((u_{h,\tau})_x)_- \right] \psi n_1 \, ds$$

$$= (f, \psi)_T, \qquad (4.44)$$

for all $\psi \in P_{k-m(T)}(T)$, where the notation is that of (4.39). Take $k \geq 2$ so that $k - m(T) \geq 0$. As $u_{h,\tau}$ lies in $C(Q)$ and is already computed on $\partial_- T$, the number of degrees of freedom of $u_{h,\tau}$ on T equals the dimension of $P_{k-m(T)}(T)$.

This method, with $\varepsilon = 0$, is due to Reed and Hill [RH73]. Richter [Ric90] introduced the generalization (4.44). Note how (4.44) resembles (4.39); this partly explains why the analyses of the methods are alike.

Example 4.17. We use the same triangulation as in Example 4.14, with $k = 2$ for the simplest possible case. Then the value of $u_{h,\tau}$ must be determined at six points in each triangle: the vertices and the midpoints of the edges.

Let the triangle T have vertices $(x_{i-1}, t_{j-1}), (x_i, t_{j-1})$ and (x_{i-1}, t_j). Then $m(T) = 2$, so $k - m(T) = 0$. Since $\dim P_0(T) = 1$ and $u_{h,\tau}$ is already computed on $\partial_- T$, (4.44) yields a single equation in the single unknown

$$u_{i-1/2}^{j-1/2} := u\left(\frac{x_{i-1} + x_i}{2}, \frac{t_{j-1} + t_j}{2} \right).$$

Take $\psi = 1$ in (4.44) to get this equation:

$$\frac{b}{3h} u_i^{j-1} + \frac{4}{3}\left(\frac{b}{h} + \frac{1}{\tau} \right) u_{i-1/2}^{j-1/2} + \frac{1}{3\tau} u_{i-1}^j - \frac{4b}{3h} u_{i-1}^{j-1/2}$$

$$- \frac{1}{3}\left(\frac{b}{h} + \frac{1}{\tau} \right) u_{i-1}^{j-1} - \frac{4}{3\tau} u_{i-1/2}^{j-1} + \frac{2\varepsilon}{h^2}\left[-3u_i^{j-1} + 2u_{i-1/2}^{j-1/2} \right.$$

$$\left. - u_{i-1}^j - 2u_{i-1}^{j-1/2} - 3u_{i-1}^{j-1} + 4u_{i-1/2}^{j-1} + 2u_{i-3/2}^j - u_{i-2}^j + 2u_{i-3/2}^{j-1/2} \right]$$

$$= \frac{2}{h\tau} \int_T f. \qquad (4.45)$$

If we formally allow k to equal 1, then (4.44) is vacuous on T since $m(T) = 2$. But on the triangle T' to the left of T, one has $m(T') = 1$; if we set $\varepsilon = 0$ and $k = 1$ then (4.44) applied on T' yields the simple upwind scheme (4.40). ♣

Assume that $\varepsilon \leq C'h$ for some constant C'. Our assumptions on the mesh then imply that

$$\varepsilon \tau_T \leq C'h_T^2 \quad \text{for all } T, \tag{4.46}$$

where h_T and τ_T are the lengths of the projections of triangle T onto the x- and t-axes respectively. Condition (4.46) ensures that the method is stable [Ric90]. Under the reasonable assumption that the triangulation can be divided into J strips, where $J = \mathcal{O}(h^{-1})$ and each strip is one triangle deep, it is shown in [Ric90] that one has the following error estimates for the continuous Reed-Hill-Richter method:

Theorem 4.18. *Under the above assumptions, the method (4.44) has a unique solution $u_{h,\tau}$. There exists a constant C such that*

$$\|u - u_{h,\tau}\|_{L_2(Q)} + \sqrt{\varepsilon}\,\|(u - u_{h,\tau})_x\|_{L_2(Q)} \leq C\varepsilon^{-1/2}h^{n+1}\|u\|_{H^{n+1}(Q)}, \tag{4.47a}$$

and, when $\varepsilon \leq Ch^{3/2}$,

$$\|u - u_{h,\tau}\|_{L_2(Q)} + h^{1/4}\|(u - u_{h,\tau})_z\|_{L_2(Q)} + h^{3/4}\|(u - u_{h,\tau})_x\|_{L_2(Q)}$$
$$\leq Ch^{n+1/4}\|u\|_{H^{n+1}(Q)}. \tag{4.47b}$$

Remark 4.16 applies to these estimates also.

Richter [Ric90] considers a modification of (4.6) near S_1 in order to incorporate data from the outflow boundary, but concludes that the best strategy is to use a triangulation of Q that excludes S_1 from the domain of dependence of $u_{h,\tau}$ in the interior of Q. Numerical results in [Ric90] for the case $k = 2$ illustrate the effect of this strategy on the computed solution. One example, for which $\varepsilon = h$, exhibits the optimal convergence rate of $\mathcal{O}(h^3)$ for $\|u - u_{h,\tau}\|_{L_2(Q)}$, although the estimate (4.47a) predicts only $\mathcal{O}(h^{5/2})$.

Falk and Richter [FR92] prove the following *local error estimate* for the method, while assuming that $\varepsilon \leq Ch$ for some sufficiently small C.

Let (x_i, t_j) be a node of the triangulation. Let $u \in H^{n+1}(Q)$. Set

$$D = \{(x, t) \in Q : t \leq t_j, |x - bt - (x_i - bt_j)| \leq K\}$$

for some positive K, and

$$D' = \{(x, t) \in Q : t \leq t_j, |x - bt - (x_i - bt_j)| \leq K + C'\sqrt{h}\,|\ln h|\},$$

where C' is constant with $C' > 2n+7/2$. Let D_h be the union of all triangles T for which $T \cap D'$ is nonempty. Then

$$\|(u - u_{h,\tau})_z\|_{L_2(D)} + h^{1/2}\|\nabla(u - u_{h,\tau})\|_{L_2(D)}$$
$$\leq Ch^n \left\{\|u\|_{H^{n+1}(D_h)} + \delta \left[\|u\|_{L_2(Q)} + \|f\|_{L_2(Q)} + \|s\|_{L_2(0,1)}\right]\right\}, \tag{4.48}$$

where $\delta \to 0^+$ as $h \to 0$, independently of ε.

This local estimate of $\|(u - u_{h,\tau})_z\|_{L_2(Q)}$ is of optimal order.

The methods of this subsection have in recent years been superseded by more sophisticated discontinuous Galerkin finite element methods that incorporate additional features such as the stabilization of jumps across edges in the computed solution. In Section III.3.4 these methods are presented for elliptic problems; for their application in a parabolic context see [CCSS02] and its references, and the discussion in Section III.4.3.

4.2.3 Eulerian-Lagrangian Methods

We know from Chapter 2 that the solution of (4.1) closely approximates the solution of the reduced problem (2.5) on most of Q. The methods of Section 4.2.2 use this fact implicitly, since they choose triangles in an order that depends on the subcharacteristics of (4.1). Eulerian-Lagrangian finite element methods are also based on the premise that the computed solution of (4.1) should evolve along the subcharacteristics, just as (more or less) the true solution does; they are akin to the classical "method of characteristics" for first-order hyperbolic problems.

The basic methodology of this section appears in the literature under various names, such as the *characteristic Galerkin method*, the *Lagrange-Galerkin method* and the *modified method of characteristics*. A discussion of some of these methods and of their properties can be found in [EW01]; see also [BK02, CRHE90, DHP99, DR95, HS01a, Pir89, Pri94] and their references. Furthermore, the *characteristic streamline diffusion method* of Section 5.1 below falls into this category. These methods all bear a strong resemblance to each other. Consequently, instead of attempting to present an overview of all Eulerian-Lagrangian methods for (4.1), we shall consider only the *Eulerian-Lagrangian local adjoint method* (ELLAM) [CRHE90] since it has been applied to many types of problem [EW01], its analysis is well developed, and unlike some of the other methods in this family it conserves mass, which is often important in applications.

Assume for simplicity that b is constant and $d \equiv 0$. Then (4.1a) becomes

$$-\varepsilon u_{xx} + u_z = f. \tag{4.49}$$

Writing u_0 for the solution of the reduced problem (2.5), one has

$$u_0(x,t) = u_0(x - b\Delta t, t - \Delta t) + \int_{y=0}^{\Delta t} f\left(x + b(y - \Delta t), t + y - \Delta t\right) dy \tag{4.50}$$

for any $\Delta t > 0$, provided that (x,t) and $(x - b\Delta t, t - \Delta t)$ both lie in Q.

Our treatment of convection will use a Lagrangian frame of reference to mimic (4.50), while for diffusion the previous Eulerian frame is retained.

Place the usual equidistant rectangular grid of points $\{(x_i, t_j)\}$ on $\bar{Q}$, where $x_i = i/M = ih$ for $i = 0, \ldots, M$ and $t_j = j\,T/N = j\tau$ for $j = 0, \ldots, N$. We shall use trial and test functions that depend on both x and t, lie in $C(\bar{Q})$, and vanish on $S_0 \cup S_1$.

Set
$$Q_j := [0,1] \times (t_{j-1}, t_j).$$

Let ψ be a typical basis function from the space of test functions, so ψ has "small" support. Assume that the support of ψ does not intersect any sub-characteristic that passes through $(S_0 \cup S_1) \cap Q_j$. Then (4.49) implies that

$$\iint_{Q_j} f\psi \, dt \, dx = \iint_{Q_j} (-\varepsilon u_{xx} + u_z)\psi \, dt \, dx$$

$$= \int_{x=0}^{1} \int_{z'=t_{j-1}}^{t_j} (-\varepsilon u_{xx} + u_z)\psi \, dz' \, dx, \qquad (4.51)$$

where z' is a subcharacteristic coordinate. Now integration by parts gives

$$\int_{x=0}^{1} \int_{z'=t_{j-1}}^{t_j} (-\varepsilon u_{xx}\psi - u\psi_z) \, dz' \, dx$$

$$+ \int_{x=0}^{1} [u(x,t_j)\psi(x,t_j) - u(x - b\tau, t_{j-1})\psi(x - b\tau, t_{j-1})] \, dx$$

$$= \iint_{Q_j} f\psi \, dt \, dx. \qquad (4.52)$$

For each t, the support of $\psi(\cdot, t)$ lies in a small interval, so (4.52) is analogous to (4.50).

We intend to replace u in (4.52) by our computed solution $u_{h,\tau}$ in order to generate a difference scheme. Before doing this, formally integrate twice by parts so that

$$\iint (-\varepsilon u_{xx}\psi - u\psi_z) \, dz' \, dx = \iint u(-\varepsilon \psi_{xx} - \psi_z) \, dz' \, dx. \qquad (4.53)$$

The test functions ψ are chosen to yield an approximate solution of the adjoint equation $-\varepsilon \psi_{xx} - \psi_z = 0$ almost everywhere in Q. There are two obvious ways of doing this. The first is to choose $\psi(x,t)$ to satisfy (cf. Section 4.1.2)

$$-\varepsilon \psi_{xx} - b\psi_x = 0 \quad \text{and} \quad -\psi_t = 0.$$

In the ELLAM, one requires instead that ψ satisfy the local adjoint equation for the convective derivatives given by

$$-\psi_z = -b\psi_x - \psi_t = 0, \qquad (4.54)$$

which says that each test function is constant along the subcharacteristics of (4.1); thus convection is treated in a Lagrangian frame of reference. Diffusion is treated from the Eulerian point of view, so the test functions at each discrete time level t_j can be any standard piecewise polynomials in the x variable, and since ε is small one has $-\varepsilon \psi_{xx} - \psi_z \approx 0$.

Let us consider the ELLAM in detail when the trial functions are piecewise linear functions of x. Then for $j = 1, \ldots, N$ and $i = 1, \ldots, M - 1$, define

$$
\psi_{i,j}(x,t) := \begin{cases} (x - x_{i-1})/h + b(t_j - t)/h & \text{for } (x,t) \in Q_{i,j}, \\ (x_{i+1} - x)/h - b(t_j - t)/h & \text{for } (x,t) \in Q_{i+1,j}, \\ 0 & \text{otherwise}, \end{cases} \tag{4.55}
$$

where

$$
Q_{m,j} := \{(x,t) \in Q_j : x_{m-1} - bt_j \le x - bt \le x_m - bt_j\}.
$$

The test space on Q_j is the span of $\{\psi_{i,j}(x,t) : i = 1, \ldots, M - 1\}$.

Now substitute (4.53) into (4.52), then replace u by $u_{h,\tau}$ and ψ by $\psi_{i,j}$. This gives

$$
-\frac{\varepsilon}{h} \int_{z'=t_{j-1}}^{t_j} [u_{h,\tau}(x_{i-1} - b(t_j - z'), z')
$$
$$
- 2u_{h,\tau}(x_i - b(t_j - z'), z') + u_{h,\tau}(x_{i+1} - b(t_j - z'), z')]\, dz'
$$
$$
+ \frac{1}{h} \int_{x=x_{i-1}}^{x_i} [u_{h,\tau}(x, t_j) - u_{h,\tau}(x - b\tau, t_{j-1})](x - x_{i-1})\, dx
$$
$$
+ \frac{1}{h} \int_{x=x_i}^{x_{i+1}} [u_{h,\tau}(x, t_j) - u_{h,\tau}(x - b\tau, t_{j-1})](x_{i+1} - x)\, dx
$$
$$
= \iint_{Q_j} f\psi_{i,j}\, dt\, dx. \tag{4.56}
$$

In practice, one uses a quadrature rule to evaluate the first integral in (4.56). The simplest option is a one-point rule at $z' = t_j$, which yields

$$
-\frac{\varepsilon}{h^2} [u_{i-1}^j - 2u_i^j + u_{i+1}^j]
$$
$$
+ \frac{1}{h^2\tau} \int_{x=x_{i-1}}^{x_i} [u_{h,\tau}(x, t_j) - u_{h,\tau}(x - b\tau, t_{j-1})](x - x_{i-1})\, dx
$$
$$
+ \frac{1}{h^2\tau} \int_{x=x_i}^{x_{i+1}} [u_{h,\tau}(x, t_j) - u_{h,\tau}(x - b\tau, t_{j-1})](x_{i+1} - x)\, dx
$$
$$
= \frac{1}{h\tau} \iint_{Q_j} f\psi_{i,j}\, dt\, dx. \tag{4.57}
$$

Note that the diffusion term $-\varepsilon u_{xx}$ is approximated by a spatial difference, while the convective term $bu_x + u_t$ is in effect upwinded along the subcharacteristics.

Example 4.19. For each fixed t_k, suppose that the trial functions are piecewise linear on $[0, 1]$. Then (4.57) simplifies to [CRHE90]

$$-\frac{\varepsilon}{h^2}[u^j_{i-1} - 2u^j_i + u^j_{i+1}] + \frac{1}{6\tau}[u^j_{i-1} + 4u^j_i + u^j_{i+1}]$$

$$-\frac{1}{6\tau}\Big[(1 - 3\theta + 3\theta^2 - \theta^3)u^{j-1}_{i-\lfloor\nu\rfloor-2} + (4 - 6\theta^2 + 3\theta^3)u^{j-1}_{i-\lfloor\nu\rfloor-1}$$

$$+ (1 + 3\theta + 3\theta^2 - 3\theta^3)u^{j-1}_{i-\lfloor\nu\rfloor} + \theta^3 u^{j-1}_{i-\lfloor\nu\rfloor+1}\Big]$$

$$= \frac{1}{h\tau}\iint_{Q_j} f\psi_{i,j}\, dt\, dx, \tag{4.58}$$

where $\nu = b\tau/h$ is the Courant number, $\lfloor\nu\rfloor$ denotes its integer part, and $\theta := 1 - (\nu - \lfloor\nu\rfloor)$. ♣

Remark 4.20. (Boundary conditions) Near S_0 and S_1, a subcharacteristic passing through a point (x_m, t_k), where $k = j - 1$ or j, may leave $\bar{Q}$ before it reaches $t = t_{2j-k-1}$. In the derivation of (4.52), we took care to avoid such subcharacteristics. When they occur, (4.56) must be modified. The treatment of boundary conditions in Eulerian-Lagrangian methods is in general a non-trivial problem, but for the ELLAM it can be done in a systematic manner [WDE+99]. ♣

Remark 4.21. (Variable b) When b is non-constant, so the subcharacteristics of (4.1) are curved, then in (4.52) the test functions are specified as before at $t = t_j$ and are extended backwards in time to $t = t_{j-1}$ by requiring that they remain constant along each subcharacteristic. Thus (4.54) is satisfied. Russell and Trujillo [RT90] show how certain resulting integrals in the method should be evaluated using a forward-tracking algorithm.

Furthermore, one then needs to impose a time-stepping restriction of the form $\tau\|b\|_{L_\infty(0,T;W^1_\infty)} \le 1$ (this norm is defined in (III.4.8)) to ensure that subcharacteristics from neighbouring mesh points cannot cross during each time step. ♣

When piecewise linear trial functions are used, Wang et al. [WER95] show that for all j one has

$$\|u(\cdot, t_j) - u_{h,\tau}(\cdot, t_j)\|_{L_2(0,1)} \le K(h^2 + \tau). \tag{4.59}$$

This result is of optimal order; it should be noted that theoretical error bounds for alternative Eulerian-Lagrangian methods are sometimes suboptimal. Here, however, K depends on certain Sobolev norms of u and hence on ε. Numerical results in [WER95] show that (4.59) is sharp when u has no layers and that the scheme is stable even for large Courant numbers. See [WW] for a more recent related result.

In [WW07] the authors consider (4.1a) with periodic boundary data, which excludes a boundary layer. A complicated analysis demonstrates that for all j one has

$$\|u(\cdot, t_j) - u_{h,\tau}(\cdot, t_j)\|_{L_2(0,1)} \le C[h^2 + \tau + \min\{h, \tau\}], \tag{4.60}$$

where the constant C depends only on the data of the problem and so is completely independent of ε; this is the first ε-independent error bound for the ELLAM.

A finite volume analogue of the ELLAM appears in [HR93] and is analysed in [RSW08].

In Section III.4.3 we shall return to the ELLAM in the context of multi-dimensional parabolic problems.

5

Two Adaptive Methods

All methods considered so far in Part II use fixed meshes that are chosen *a priori*. Adaptive methods, which were applied to two-point boundary value problems in Section I.2.5, aim to produce accurate numerical solutions by refining the mesh in certain regions, using the current computed solution to (4.1) as a guide to this refinement. In the present Chapter we consider two such methods.

An adaptive method can be constructed by combining any of the numerical methods of Part II with some recipe that refines the mesh in regions where the computed solution appears to be inaccurate. Plausible mesh-refinement criteria abound in the literature, but it is difficult to give meaningful analyses (i.e., error bounds that are independent of ε) of adaptive methods in the context of convection-diffusion problems. We nevertheless acknowledge that approaches based on an incomplete theory can yield satisfactory numerical results; see, e.g., [AFMW92, CA92, FVZ90].

The discussion below does not attempt to give a comprehensive overview of adaptive methods; see also Sections I.2.5 and III.3.6. Our exposition is confined to *streamline diffusion* finite element methods and to *moving mesh* methods. These topics will provide an adequate exposure to the core ideas used in adaptive techniques.

5.1 Streamline Diffusion Methods

In the present section two variants of the streamline diffusion finite element method (SDFEM) of Section 4.2.1 are considered; the first of these adds a controlled amount of artificial diffusion to the problem, while the second uses a mesh that is oriented along the subcharacteristics of the problem.

Suppose for simplicity that $d \equiv 0$ and $b \equiv 1$ in (4.1), so the problem is

$$u_t(x,t) - \varepsilon u_{xx}(x,t) + u_x(x,t) = f(x,t) \text{ for } (x,t) \in Q, \qquad (5.1a)$$

$$u(x,0) = s(x) \quad \text{on } S_x := \{(x,0) : 0 \le x \le 1\}, \qquad (5.1b)$$

$$u(0,t) = 0 \quad \text{on } S_0 := \{(0,t) : 0 < t \le T\}, \qquad (5.1c)$$

$$\frac{\partial u}{\partial x}(1,t) = 0 \quad \text{on } S_1 := \{(1,t) : 0 < t \le T\}. \qquad (5.1d)$$

Note that on S_1 the Dirichlet boundary condition of (4.1d) has been replaced by a Neumann condition. Consequently no strong boundary layer in u forms at S_1 (cf. Remark I.1.5) but interior layers may still be present.

Consider the following modification of the SDFEM (4.30), with the same triangulation and choice of linear trial functions as in (4.30), while assuming that $\max\{\tau/h, h/\tau\} \le C$:

$$
\begin{aligned}
(\hat\varepsilon \hat u_x^j, \phi_x)_{Q_j} &+ (\hat\varepsilon \hat u_t^j, \phi_t)_{Q_j} + (\hat u_t^j + \hat u_x^j, \phi + \delta(\phi_t + \phi_x))_{Q_j} + \langle \hat u_+^j, \phi_+ \rangle_{j-1} \\
&= (f, \phi + \delta(\phi_t + \phi_x))_{Q_j} + \langle \hat u_-^{j-1}, \phi_+ \rangle_{j-1},
\end{aligned}
\qquad (5.2)
$$

for $j = 1, \ldots, N$ and all $\phi \in V_j$, where on each triangle $T \in Q_j$ one sets

$$\hat\varepsilon|_T := \max_T\{\varepsilon, C_2 h^2 |f - \hat u_t^j - \hat u_x^j|\} \quad \text{and} \quad \theta|_T := C_1 \min\{0, h - \hat\varepsilon\},$$

with C_1 and C_2 positive constants. Compared with (4.30), we have introduced a *shock-capturing* artificial diffusion $\hat\varepsilon$ that in general depends on the computed solution $u_{h,\tau}$. Thus (5.2) *is a nonlinear method*, even though (5.1) is a linear problem.

Eriksson and Johnson [EJ93a] analyse this method (cf. [EJ93b], which is similar), while assuming that $\hat\varepsilon$ is approximable by a smoothly varying function that we also call $\hat\varepsilon$. They argue heuristically that $\hat\varepsilon = \max\{\varepsilon, \mathcal{O}(h^{3/2})\}$ near interior layers and that $\hat\varepsilon = \max\{\varepsilon, \mathcal{O}(h^3)\}$ on regions where the solution is smooth. Thus in all cases $\hat\varepsilon$ is small.

Let $\tilde u$ be the solution of the problem obtained when (5.1a) is replaced by

$$\tilde u_t - (\hat\varepsilon \tilde u_t)_t - (\hat\varepsilon \tilde u_x)_x + \tilde u_x = f,$$

with the same initial-boundary conditions as in (5.1b)–(5.1d). Now

$$\|u - u_{h,\tau}\| \le \|u - \tilde u\| + \|\tilde u - u_{h,\tau}\|,$$

where $\|\cdot\|$ denotes the $L_2(Q)$ norm and $u_{h,\tau}$ is the solution computed by (5.2). One can regard $u - \tilde u$ as a perturbation error and $\tilde u - u_{h,\tau}$ as a discretization error. In [EJ93a] a rigorous *a posteriori* bound on $\|\tilde u - u_{h,\tau}\|$ is derived. The term $\|u - \tilde u\|$ is more troublesome: for it only a heuristic *a posteriori* bound is given. (A derivation of a bound on $\|u - \tilde u\|$ that depends only on $u_{h,\tau}$ and on the data of (5.1) is sketched in [EJ93a].)

The results outlined here still hold true, with appropriate changes in the notation, when the equidistant structured mesh of Section 4.2.1 is replaced by any triangular mesh that satisfies the classical minimum angle condition.

Eriksson and Johnson [EJ91] describe two local refinement methods whose objectives are to construct a mesh for which $\|u - u_{h,\tau}\|$ is less than some user-prescribed tolerance and whose *a posteriori* error estimators are approximately equally distributed over all triangles in the mesh. Their first approach begins with a coarse mesh which is refined locally until some tolerance is reached. The alternative method of [EJ91] is to create triangles of a suitable size at a "front" separating the triangulated and untriangulated parts of Q; a single sweep across Q generates the final mesh.

Remark 5.1. Either of these approaches could be implemented with any finite element method for which an *a posteriori* bound on $\|u - u_{h,\tau}\|$ is available. ♣

Remark 5.2. The *characteristic streamline diffusion method* [EJ93a, Han92] is a streamline diffusion method of Eulerian-Lagrangian type. In each time interval, the movement of the space mesh is approximately along the subcharacteristics. The analysis of this method shows that one may often use mesh triangles whose length in the subcharacteristic direction greatly exceeds their length in the x-direction. Consequently we can adaptively compute a mesh that has far fewer degrees of freedom than meshes based on rectangular grids in the (x, t)-plane. ♣

All methods of Section 5.1 extend to higher-dimensional problems.

5.2 Moving Mesh Methods (r-refinement)

The adaptive methods of Section 5.1 introduce extra grid nodes in subregions of Q where they are apparently useful. We now consider a popular alternative adaptive method for solving (4.1) that does not attack the problem by increasing the number of mesh points; instead, the number of spatial nodes remains the same as one moves from the discrete time level $t = t_{j-1}$ to the level $t = t_j$, but the locations of these nodes is altered so as to get a good approximation (in some norm) to $u(\cdot, t_j)$. This leads to the name *moving mesh method*. The technique is also called *r-refinement*. When the method is applied to convection-diffusion problems, some nodes will cluster around layers.

The literature contains many papers that implement this idea in both finite difference and finite element contexts. See Hawken et al. [HGH91] for a thorough survey up to 1988; more recent work can be found in the bibliographies of [BMRS02, Hua01].

There are two types of moving grid method. In the *static method* one computes a trial solution at discrete time levels t_j, using information from $t = t_{j-1}$. This solution is used to choose a mesh on $[0, 1] \times \{t_j\}$, then the solution at $t = t_j$ is recomputed using this new mesh. Thus there is no direct link between the mesh on $[0, 1] \times \{t_{j-1}\}$ and that on $[0, 1] \times \{t_j\}$.

We shall study the more widely used *dynamic method*. In this framework, one starts from a user-specified mesh on S_x, then for each t the mesh on $[0,1] \times \{t\}$ is constructed by continuously moving the nodes according to some prescription.

Should the nodes be explicitly forced to follow the subcharacteristics of (4.1)? This might appear attractive for linear convection-diffusion problems, but if the solution reaches steady-state one certainly does not want the nodes to keep following the subcharacteristics as this will move them all to the outflow boundary. Furthermore, Furzeland et al. [FVZ90] give a nonlinear reaction-diffusion example for which such subcharacteristic-tracking is incompatible with the primary aim of generating an optimal spatial approximation to $u(\cdot, t)$. Bank and Santos [BS93a] describe a finite element method where the user can directly control the location of the nodes at each discrete time level. An alternative approach (for a pure convection problem) is described in [MB01].

Many authors select the nodes at each time level so as to equidistribute (cf. Section I.2.5) some estimate of the solution. That is, the value of the estimate on each subinterval of $[0,1] \times \{t\}$ should be the same for all subintervals. This condition and (4.1) generate a system of nonlinear ordinary differential equations that control the computed solution and the movement of the mesh.

We now present a detailed description of the *moving finite element method*, which has received much attention. Other moving mesh methods can be found in [FVZ90, HGH91, HHG92, HR01, LBD$^+$02] and are applied to nonlinear variants of (4.1) such as Burgers' equation $-\varepsilon u_{xx} + u u_x + u_t = 0$ in [BKS98, BMRS01, HLP03, MQS97]; see also Section III.4.3. Huang et al. [HRR94] discuss the relationships between the movements of the mesh in some of these methods. For further reading, see Baines's monograph [Bai94].

For each $t \in [0, T]$, let

$$0 = X_0(t) < X_1(t) < \ldots < X_M(t) = 1 \tag{5.3}$$

denote the mesh on $[0,1] \times \{t\}$. Set

$$X(t) := (X_0(t), X_1(t), \cdots, X_M(t)).$$

Our computed solution u_h will be piecewise linear in x for each fixed t. The basis functions on the mesh $X(t)$ are

$$\phi_i(x, X(t)) = \begin{cases} (x - X_{i-1})/\Delta X_i & \text{when } X_{i-1} \le x \le X_i, \\ (X_{i+1} - x)/\Delta X_{i+1} & \text{when } X_i \le x \le X_{i+1}, \\ 0 & \text{otherwise}, \end{cases} \tag{5.4}$$

for $i = 1, \ldots, M-1$, where $\Delta X_i := X_i - X_{i-1}$. Then

$$u_h(x, t) = \sum_{i=1}^{M-1} u_i(t) \phi_i(x, X(t)) \quad \text{for } (x, t) \in \bar{Q}. \tag{5.5}$$

This representation generalizes (4.3).

From (5.4) and (5.5),

$$(u_h)_t(x,t) = \sum_{i=1}^{M-1} \left[u_i'(t)\phi_i(x, X(t)) + u_i(t) \sum_{j=i-1}^{i+1} \frac{\partial\phi_i(x, X(t))}{\partial X_j} X_j'(t) \right],$$

where each prime denotes differentiation with respect to t. That is,

$$
\begin{aligned}
(u_h)_t(x,t) = \sum_{i=1}^{M-1} \bigg[& u_i'(t)\phi_i(x, X(t)) \\
& + \sum_{j=1}^{M-1} X_j'(t) \sum_{i=j-1}^{j+1} u_i(t) \frac{\partial\phi_i(x, X(t))}{\partial X_j} \bigg].
\end{aligned}
\tag{5.6}
$$

Differentiating (5.4), we see that

$$\frac{\partial\phi_i(x, X(t))}{\partial X_i} = \begin{cases} -\phi_i/\Delta X_i & \text{when } X_{i-1} < x < X_i, \\ \phi_i/\Delta X_{i+1} & \text{when } X_i < x < X_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

Now $\phi_{i-1} + \phi_i = 1$ when $x \in [X_{i-1}, X_i]$, so (5.7) implies that

$$\frac{\partial\phi_{i-1}(x, X(t))}{\partial X_i} = \frac{\phi_i}{\Delta X_i} \quad \text{when } X_{i-1} < x < X_i. \tag{5.8}$$

Similarly,

$$\frac{\partial\phi_{i+1}(x, X(t))}{\partial X_i} = \frac{-\phi_i}{\Delta X_{i+1}} \quad \text{when } X_i < x < X_{i+1}. \tag{5.9}$$

Substitution of (5.7)–(5.9) into (5.6) gives

$$(u_h)_t(x,t) = \sum_{i=1}^{M-1} \left[u_i'(t)\phi_i(x, X(t)) + X_i'(t)\omega_i(x, X(t)) \right], \tag{5.10}$$

where

$$\omega_i(x, X(t)) = \begin{cases} -\Delta u_i\phi_i(x, X(t))/\Delta X_i & \text{when } X_{i-1} < x < X_i, \\ -\Delta u_{i+1}\phi_i(x, X(t))/\Delta X_{i+1} & \text{when } X_i < x < X_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$

and $\Delta u_i := u_i - u_{i-1}$. Each function ω_i is piecewise linear and is in general discontinuous at $x = X_i$.

We now generate the semidiscrete system of equations that determines the $2(M-1)$ unknown functions $u_i(t)$ and $X_i(t)$. Replace u in (4.1) by u_h, invoke (5.10), then for fixed t minimize the $L_2(0,1)$ norm of the residual (i.e., apply a

standard Galerkin procedure). This gives, for $j = 1, \ldots, M-1$ and $t \in [0, T]$,

$$\sum_{i=1}^{M-1} [u_i'(t)(\phi_i, \phi_j) + X_i'(t)(\omega_i, \phi_j)]$$

$$+ \varepsilon((u_h)_x, (\phi_j)_x) + (b(u_h)_x + cu_h, \phi_j) = (f, \phi_j) \qquad (5.11a)$$

and

$$\sum_{i=1}^{M-1} [u_i'(t)(\phi_i, \omega_j) + X_i'(t)(\omega_i, \omega_j)]$$

$$+ \varepsilon((u_h)_x, (\omega_j)_x) + (b(u_h)_x + cu_h, \omega_j) = (f, \omega_j), \qquad (5.11b)$$

where $(\cdot, \cdot)$ is the $L_2(0,1)$ inner product for fixed t. Here t was dropped from the notation for brevity. If the mesh were fixed, so that $X_i' = 0$ for all i, then (5.11a) would be the standard Galerkin equations for the method of lines.

The terms $\sum_i [\ldots]$ in (5.11) are given explicitly by

$$[u_{j-1}' \Delta X_j + 2u_j'(\Delta X_j + \Delta X_{j+1}) + u_{j+1}' \Delta X_{j+1}]/6$$

$$- [X_{j-1}' \Delta u_j + 2X_j'(\Delta u_j + \Delta u_{j+1}) + X_{j+1}' \Delta u_{j+1}]/6 \qquad (5.12a)$$

and

$$- [u_{j-1}' \Delta u_j + 2u_j'(\Delta u_j + \Delta u_{j+1}) + u_{j+1}' \Delta u_{j+1}]/6$$

$$+ [X_{j-1}'(\Delta u_j)^2/\Delta X_j + 2X_j'((\Delta u_j)^2/\Delta X_j + (\Delta u_{j+1})^2/\Delta X_{j+1}))$$

$$+ X_{j+1}'(\Delta u_{j+1})^2/\Delta X_{j+1}]/6 \qquad (5.12b)$$

respectively. Thus one can write (5.11) as the system of nonlinear ordinary differential equations

$$A(y)y' = g(y), \qquad (5.13)$$

where $y := [u_1, X_1, u_2, X_2, \ldots, u_{M-1}, X_{M-1}]^T$, the block tridiagonal matrix $A(y)$ is gleaned from (5.12), and the vector $g(y)$ contains all terms that depend on ε, b, d and f. The system (5.13) is valid for all $t > 0$. It is complemented by the initial condition $y(0)$, whose data come from $s(\cdot)$ and from the initial mesh on S_x chosen by the user.

Does (5.13), subject to $y(0)$, have a unique solution? The answer in general is no. We could have foreseen as early as (5.5) that such an unwelcome circumstance might occur. For if $u_h(\cdot, t)$ has the same slope on (X_{i-1}, X_i) and (X_i, X_{i+1}) for some i, then the representation (5.5) is not unique: one can move X_i to any other location in (X_{i-1}, X_{i+1}) and change u_i accordingly. The phenomenon is known as *parallelism*. A related difficulty may arise when two nodes X_{i-1} and X_i are very close to each other. If $u_h(X_{i-1}, t)$ and $u_h(X_i, t)$ are also very close, then a swop of ϕ_{i-1} and ϕ_i in (5.5) has little effect on u_h. Such virtual loss of uniqueness of representation ill-conditions $A(y)$.

These two deficiencies, *parallelism* and *node-crossing*, have attracted much attention. Various modifications of the basic method try to exclude them. For

example, Miller [Mil83] introduces penalty functions. Unlike our minimization of the L_2 norm of a residual (call it R) to get (5.11), he minimizes

$$(R, R) + \sum_{i=1}^{M-1} (\gamma_i \Delta(X_i') - \delta_i)^2$$

when deriving (5.11b), with

$$\gamma_i^2 := K_1^2/(\Delta X_i - K_3), \qquad \delta_i := K_2^2/[\gamma_i(\Delta X_i - K_3)^2],$$

and K_1, K_2, K_3 user-chosen small positive constants. In effect, K_3 usually acts like a lower bound on the distance between adjacent nodes. Furzeland et al. [FVZ90] discuss how K_1, K_2 and K_3 should be chosen in practice.

Even when one does not encounter the twin difficulties described above, the system (5.13) is stiff. In practice the implicit backward differentiation formula method is frequently used to solve (5.13). The stability of time-integration schemes for (5.13) is examined in [MM07].

Baines [Bai91] considers the reduced problem (2.5) and shows that in certain cases the moving finite element method for this problem is closely related to the classical method of characteristics. Zegeling and Blom [ZB92] apply the method to convection-diffusion problems and investigate the movement of the nodes. Their numerical experiments show that the moving nodes tend to follow the characteristics of the reduced problem, but that in more than one space dimension this property may cause severe distortion of the mesh.

The satisfactory extension of moving mesh methods to time-dependent problems with two or more space dimensions will be considered in Section III.4.3.

Part III

Elliptic and Parabolic Problems in Several Space Dimensions

Parts I and II contain results that are representative of the large body of theory dealing with singularly perturbed boundary value problems in one space variable. We now move to several space dimensions, where one encounters technical problems that are much more varied and challenging.

The first three Chapters of Part III will discuss the linear singularly perturbed boundary value problem

$$Lu := -\varepsilon \Delta u + b(x) \cdot \nabla u + c(x)u = f(x) \quad \text{for } x \in \Omega \subset \mathbb{R}^d, \tag{0.1a}$$
$$Bu = 0 \quad \text{on } \partial\Omega = \Gamma, \tag{0.1b}$$

where B is some operator that represents the boundary conditions. Here Ω is a bounded domain in $\mathbb{R}^d$ with $d \geq 2$ and, as usual, the parameter ε satisfies $0 < \varepsilon \ll 1$. We restrict ourselves to second-order differential equations and assume, for simplicity, that the *diffusion term* is $-\varepsilon\Delta$ and not some more general elliptic expression. When the vector b is not identically zero, $b \cdot \nabla u$ represents *convection*; this is the *convection-diffusion* case. It is our main focus of interest, but we shall also make some remarks on problems of *reaction-diffusion* type where $b \equiv 0$.

The last Chapter presents discretization methods for the unsteady problem

$$u_t + Lu = f(x,t) \quad \text{for } (x,t) \in Q := \Omega \times (0,T], \tag{0.2a}$$
$$u|_{t=0} = u_0, \quad Bu = 0 \quad \text{on } \partial\Omega \times (0,T]. \tag{0.2b}$$

Here, in contrast to Part II, the space dimension is greater than one.

1

Analytical Behaviour of Solutions

Notation: throughout Chapter 1, a generic point in $\Omega \subset \mathbb{R}^d$ is denoted by $x = (x_1, x_2, \ldots, x_d)$, but in the case $d = 2$ the notation (x, y) is sometimes used instead. The standard Sobolev spaces $H^k(\Omega)$ with associated norms $\|\cdot\|_k$ and seminorms $|\cdot|_k$ are often used, as their more general counterparts $W^{k,p}(\Omega)$ with seminorm $|\cdot|_{W^{k,p}(\Omega)}$.

1.1 Classical and Weak Solutions

In general, the boundary value problem (0.1) has a classical solution that is smooth in the closed domain $\bar{\Omega}$ only if

- b, c, f and the boundary data are sufficiently smooth
- the boundary $\partial\Omega = \Gamma$ is (at least piecewise) smooth
- the boundary data satisfy some extra conditions.

Example 1.1. Consider the boundary value problem

$$-\Delta u(x, y) = 0 \quad \text{in } \Omega = (0, 1) \times (0, 1),$$
$$u = x^2 \quad \text{on } \Gamma.$$

Then the exact solution u does not lie in $C^2(\bar{\Omega})$: for if $u \in C^2(\bar{\Omega})$, then $u_{xx}(0, 0) = 2$ and $u_{yy}(0, 0) = 0$, which contradicts the differential equation. ♣

To describe classical solutions that are smooth on $\bar{\Omega}$ one should use Hölder spaces related to those of Part II, but unlike Section II.2.1 one sets

$$\text{dist}\,(x, x') = \left(\sum_{i=1}^{d}(x_i - x_i')^2\right)^{1/2} \qquad \text{for } x, x' \in \mathbb{R}^d.$$

With this measure of distance, one obtains the Hölder space $C^{2,\alpha}(\bar{\Omega})$, which (unlike Part II) requires Hölder continuity of *all* second-order, first-order and

zero-order derivatives. The Hölder space $C^{k,\alpha}(\bar{\Omega})$ is defined analogously for each non-negative integer k.

Existence theorems guaranteeing that $u \in C^{2,\alpha}(\bar{\Omega})$ usually require that the boundary Γ belong at least to the class $C^{2,\alpha}$. From the practical point of view this condition is too restrictive, so such results are not described here; see instead [ADN59]. It is more realistic to look for classical solutions in $C(\bar{\Omega}) \cap C^2(\Omega)$. We say that Ω *is a domain with a regular boundary* if Γ belongs to the class $C^{0,1}$, that is, if Γ can be described locally by Lipschitz continuous functions. In two dimensions, a polygonal domain without slits belongs to this class. Roughly speaking, for problems with a regular boundary and continuous data, one expects a classical solution in $C(\bar{\Omega}) \cap C^2(\Omega)$. Let us quote, for instance, the following slightly weakened theorem from [Mic77]:

Theorem 1.2. *Consider the elliptic differential equation (0.1a) with homogeneous Dirichlet boundary conditions. If b, c and f are Hölder continuous on $\bar{\Omega}$, $c \geq 0$ and Ω is a domain with a regular boundary, then this problem has a unique classical solution $u \in C(\bar{\Omega}) \cap C^2(\Omega)$.*

The behaviour of classical solutions near any point on the boundary where different boundary conditions meet is described, e.g., in [Wig70].

If one is interested in solutions that are smooth up to the boundary, and the domain is not in the class $C^{2,\alpha}$ but only regular, then additional conditions – *compatibility conditions* – are necessary at corners.

Example 1.3. Consider the boundary value problem

$$-\varepsilon \Delta u + b(x,y) \cdot \nabla u + c(x,y)u = f(x,y) \quad \text{in } \Omega = (0,1) \times (0,1), \quad (1.1a)$$
$$u = 0 \quad \text{on } \Gamma. \quad (1.1b)$$

Assume that b_1, b_2, c are smooth. Let $\alpha \in (0,1)$.

(i) Suppose that $f \in C^{0,\alpha}(\bar{\Omega})$. Then the boundary value problem has a solution in $C^{1,\alpha}(\bar{\Omega}) \cap C^{2,\alpha}(\Omega)$.

(ii) Let $f \in C^{0,\alpha}(\bar{\Omega})$. Then $u \in C^{2,\alpha}(\bar{\Omega})$ if and only if

$$f(0,0) = f(1,0) = f(0,1) = f(1,1) = 0. \quad (1.2)$$

If in addition to (1.2) one has $f \in C^{1,\alpha}(\bar{\Omega})$, then $u \in C^{3,\alpha}(\bar{\Omega})$. These results appear in [Vol65, Gri85b, HK90]. In [HK90], higher-order compatibility conditions are discussed; see also [Azz80]. ♣

Second-order elliptic problems satisfy *maximum* and *comparison principles* if the coefficients b and c of the operator L are continuous and $c(x) \geq 0$. See [PW67] for a discussion of more general cases.

Let $w \in C(\bar{\Omega}) \cap C^2(\Omega)$, where $\Omega \subset \mathbb{R}^n$. The operator L is said to be *inverse-monotone* if the inequalities

$$Lw(x) \geq 0 \quad \text{for all } x \in \Omega \quad \text{and} \quad w(x) \geq 0 \quad \text{for all } x \in \Gamma$$

together imply that $w(x) \geq 0$ for all $x \in \bar{\Omega}$.

We say that L satisfies a *maximum principle* if

$$Lw(x) = 0 \quad \text{for all } x \in \Omega$$

implies that for every $x \in \bar{\Omega}$

$$\min_{x \in \Gamma}\{w(x), 0\} \leq w(x) \leq \max_{x \in \Gamma}\{w(x), 0\}.$$

For our purposes the comparison principle below is the most useful formulation. The maximum and comparison principles are direct consequences of the inverse monotonicity property; in fact each of these three properties is sometimes called a maximum principle.

Theorem 1.4. *(Classical comparison principle) Suppose that $c \geq 0$. Let $v, w \in C(\bar{\Omega}) \cap C^2(\Omega)$ satisfy the inequalities*

$$|(Lv)(x)| \leq (Lw)(x) \quad \text{for all } x \in \Omega,$$
$$|v(x)| \leq w(x) \quad \text{for all } x \in \Gamma.$$

Then, for every $x \in \bar{\Omega}$, we have

$$|v(x)| \leq w(x).$$

The solution v of a given boundary value problem is often bounded by invoking Theorem 1.4 with an appropriate *barrier function w*.

The results that were sketched above show that in domains with corners – which are important in many practical applications – the theory of classical solutions is not entirely satisfactory. An alternative possibility is the use of *weak solutions*. The basic existence and uniqueness result for such solutions is Theorem I.2.31, the Lax-Milgram lemma. With a view to using this theorem, let us demonstrate here how to transform an elliptic boundary value problem into its weak formulation.

Suppose that the boundary Γ is divided into three disjoint pieces called Γ_1, Γ_2 and Γ_3. We study the problem

$$-\varepsilon \Delta u + b(x) \cdot \nabla u + c(x)u = f(x) \quad \text{in } \Omega, \qquad (1.3a)$$

$$u = 0 \quad \text{on } \Gamma_1, \qquad (1.3b)$$

$$\varepsilon \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_2, \qquad (1.3c)$$

$$\varepsilon \frac{\partial u}{\partial n} + \mu u = g \quad \text{on } \Gamma_3. \qquad (1.3d)$$

Set

$$V = \left\{ v \in H^1(\Omega) : \ v = 0 \text{ on } \Gamma_1 \right\},$$

where $H^m(\Omega)$ is the usual Sobolev space of functions whose m^{th}-order generalized derivatives lie in $L_2(\Omega)$. Introduce the bilinear form

$$a(v, w) := \varepsilon(\nabla v, \nabla w) + (b \cdot \nabla v + cv, w) + \int_{\Gamma_3} \mu v w,$$

and the linear form

$$f(v) := (f, v) + \int_{\Gamma_3} g\, v.$$

The Lax-Milgram lemma tells us that if the linear form is continuous and the bilinear form is V-elliptic and continuous, then the given problem has a unique weak solution $u \in V$. To get V-ellipticity, it is standard to assume that

$$c - \frac{1}{2} \operatorname{div} b \geq \omega > 0, \quad \mu > 0, \quad b \cdot n \geq 0 \text{ on } \Gamma_2 \cup \Gamma_3 \quad \text{and } \Gamma_1 \neq \emptyset.$$

To guarantee a smoother solution (e.g., $u \in H^2(\Omega)$) one needs additional assumptions involving the given data, the structure of the boundary and the boundary conditions. For homogeneous boundary conditions in a polygonal domain $\Omega \subset \mathbb{R}^2$, to ensure $u \in H^2(\Omega)$ requires in general convexity of Ω; see [Gri85b].

Let us mention finally that, even for weak solutions, one has maximum and comparison principles. See [GT83, Tro87] for details.

1.2 The Reduced Problem

Our experience in Part II leads us to expect that the solution of the singularly perturbed boundary value problem (0.1) is, except near layers, close to the solution of the first-order *reduced equation*

$$b(x) \cdot \nabla w + c(x) w = f(x)$$

subject to some boundary conditions. But which boundary conditions of (0.1) should one use?

Assume that Ω is a *domain with a regular boundary* that has an outward-pointing unit normal vector n defined uniquely almost everywhere on its boundary Γ. Set

$$\Gamma_+ = \{x \in \Gamma \ : \ b \cdot n > 0\},$$
$$\Gamma_- = \{x \in \Gamma \ : \ b \cdot n < 0\},$$
$$\Gamma_0 = \{x \in \Gamma \ : \ b \cdot n = 0\}.$$

The *subcharacteristics* $\xi_x(\tau)$ *of the reduced equation* are defined to be the solutions of

$$\frac{d\xi}{d\tau} = b(\xi(\tau)), \quad \xi(0) = x.$$

These subcharacteristics are transverse to the boundary at Γ_+ and Γ_-, while at Γ_0 the subcharacteristics and the boundary are parallel. As in problems

of fluid dynamics, we use the terminology *inflow boundary* for Γ_-, *outflow boundary* for Γ_+ and *characteristic boundary* for Γ_0. One often thinks of τ as a time-like variable.

It seems adequate to augment the reduced equation by boundary conditions on $\bar{\Gamma}_-$ or $\bar{\Gamma}_+$, but the cancellation law from Part I tells us that $\bar{\Gamma}_-$ is the correct choice. From now on, we shall assume homogeneous Dirichlet boundary conditions on $\bar{\Gamma}_-$. Thus, define the *reduced problem* for (0.1) to be

$$L_0 u_0 := b(x) \cdot \nabla u_0 + c(x) u_0 = f(x) \quad \text{in } \Omega, \tag{1.4a}$$

$$u_0 = 0 \quad \text{on } \bar{\Gamma}_-. \tag{1.4b}$$

The solution u_0 of the reduced problem may behave in a very complicated way, which makes it difficult to treat singularly perturbed problems in several dimensions because for small ε the solution of (0.1) is close to the solution of (1.4). A simpler version of (1.4) was examined in Section II.3.1.

We state without proof (see [BBB73, Rau72]) some basic results on existence, uniqueness and regularity of the solution u_0 of (1.4).

Lemma 1.5. *Let* $b, c \in C^1(\bar{\Omega})$, $f \in L_2(\Omega)$ *and*

$$c - \frac{1}{2} \operatorname{div} b \geq \omega > 0. \tag{1.5}$$

(i) There exists a unique solution $u_0 \in L_2(\Omega)$ *of the reduced problem (1.4).*
(ii) In the special case $\Omega = (0,1)^d$ *and* $b_k \geq \beta_k > 0$, *the solution lies in the graph space* $\{v \in L_2(\Omega) : L_0 v \in L_2(\Omega)\}$ *even for arbitrary* $u_0 \in L_2(\Gamma_-)$.

Remark 1.6. If in addition $f \in W^{1,1}(\Omega) \cap L_\infty(\Omega)$, then the solution u_0 of the reduced problem lies in $L_\infty(\Omega)$. In general, however, we can expect neither $u_0 \in W^{1,1}(\Omega)$ nor $u_0 \in H^1(\Omega)$; see [BBB73]. ♣

The properties of solutions of hyperbolic problems such as (1.4) are very different from those of elliptic problems. Thus $\varepsilon > 0$ is a general assumption in this book to exclude the case $\varepsilon = 0$. (See [Kro97, LeV90] for numerical methods for hyperbolic problems.) Let us mention two of these properties:

- Unlike elliptic operators, first order hyperbolic operators have no intrinsic smoothing property in isotropic Sobolev spaces: while for elliptic operators $Lu \in L_2(\Omega)$ implies $u \in H^s(\Omega)$ for some $s > 0$ – e.g., $\Delta u \in L_2(\Omega)$ implies $u \in H^2(\Omega)$ for convex domains Ω – this is not true of hyperbolic operators.
- In two dimension, the Green's function of an elliptic problem is smooth except for a point logarithmic singularity. But for hyperbolic problems the Green's function (or Green's measure, to be more precise) has nonlocal singularity concentration along a characteristic and does not, in general, decay along the characteristic. In this context see our investigation of the Green's function of the convection-diffusion operator in Section 1.4.

Even when the geometrical behaviour of the subcharacteristics is relatively simple – for instance, if all subcharacteristics leave the domain in a finite time – the function u_0 may not be very smooth. We give two typical examples.

Example 1.7. Consider the first-order problem

$$b(x,y) \cdot \nabla u_0 + c(x,y)u_0 = f(x,y) \quad \text{in } \Omega = (0,1) \times (0,1),$$
$$u_0 = 0 \quad \text{on } \bar{\Gamma}_-,$$

with $b = (b_1, b_2)$ and $b_1 > 0$, $b_2 > 0$.

An inspection of u_0 shows that it lies only in $W^{1,\infty}(\Omega)$; its first-order derivatives have jumps along the subcharacteristic through the corner $(0,0)$. See Figure 1.1. ♣
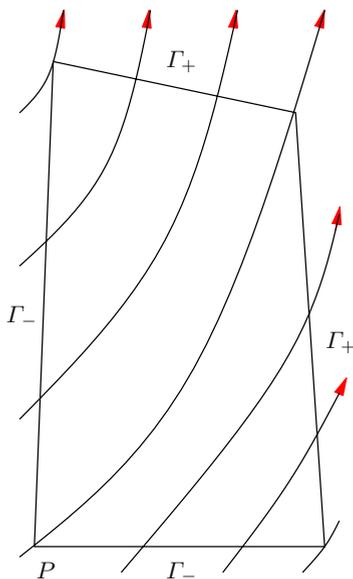
Fig. 1.1. A subcharacteristic through a corner

Example 1.8. For the problem

$$\frac{\partial u_0}{\partial y} + cu_0 = f \quad \text{in } \Omega = \left\{ (x,y) : 1 < x^2 + 2y^2 < 4 \right\},$$
$$u_0 = 0 \quad \text{on } \bar{\Gamma}_-,$$

the solution u_0 will in general be discontinuous along the line segments $\{(-1,y) : 0 \le y \le \sqrt{3/2}\}$ and $\{(1,y) : 0 \le y \le \sqrt{3/2}\}$ that pass through the points C and E in Figure 1.2. ♣
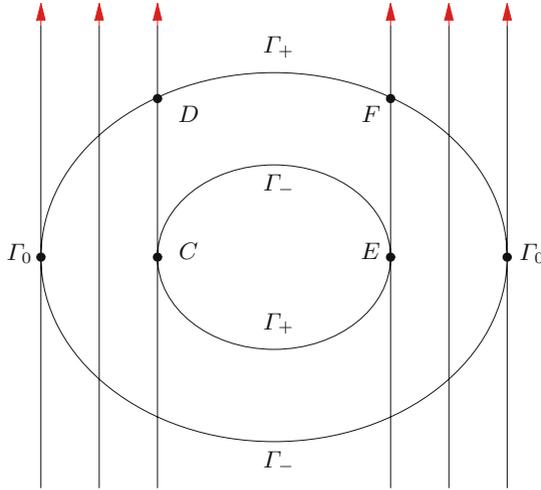
Fig. 1.2. A discontinuity at an inner subcharacteristic

The situation is simplest when

- the boundary Γ is smooth
- Γ_- is closed in Γ (hence $\Gamma_- = \Gamma$ or $\Gamma_- = \phi$ if Γ is simply connected; in general a similar statement holds true on each connected component of Γ)
- c is sufficiently large.

Lemma 1.9. *Let Ω be a domain with a smooth boundary Γ, where Γ_- is closed in Γ. Let k be a positive integer. Then for $c > c_0(k, \Omega)$ and smooth data, the solution of (1.4) lies in $C^{k,\alpha}(\bar{\Omega})$.*

See [OR73] for the proof. A simple example shows that the condition on c is necessary. Consider, for constant positive c, the problem

$$-x\frac{\partial u_0}{\partial x} - y\frac{\partial u_0}{\partial y} + cu_0 = 1 \quad \text{in } \Omega = \left\{(x,y) : 0 < x^2 + y^2 < 1\right\},$$

$$u_0 = 0 \quad \text{on } \bar{\Gamma}_- = \Gamma.$$

Then

$$u_0(x,y) = \frac{1}{c}\left[1 - \left(x^2 + y^2\right)^{c/2}\right];$$

the smoothness of u_0 does indeed depend on the magnitude of c.

If the subcharacteristics do not all leave the domain Ω in a finite time, then their geometrical behaviour is more complicated. For each $x \in \Omega$, define

$$\beta(x) := \bigcap_{0 \le s < \infty} \left\{\xi_x(\tau) : s \le \tau \le \infty\right\} \cap \bar{\Omega},$$

$$\alpha(x) := \bigcap_{-\infty < s \le 0} \left\{\xi_x(\tau) : -\infty \le \tau \le s\right\} \cap \bar{\Omega}.$$

It is usual to say that the subcharacteristic $\xi_x(\tau)$ through x *originates* at $\alpha(x)$ and *dies* at $\beta(x)$. Let $I(\bar{\Omega})$ denote the union of all sets of the above types. In particular, if $b(x^*) = 0$ then x^* belongs to $I(\bar{\Omega})$; such points are called *stationary points* of the field $b(x)$ or *turning points* of the original singular perturbation problem.

Closed subcharacteristics – that is, the images of periodic solutions of the characteristic equations – are another example of subsets of $I(\bar{\Omega})$. They can appear, for instance, as a continuum or as an isolated *limit cycle*.

Example 1.10. Let us consider the equation

$$[x(r^2 - 1) - y(r^2 + 1)]\frac{\partial v}{\partial x} + [y(r^2 - 1) + x(r^2 + 1)]\frac{\partial v}{\partial y} + cv = f$$

where $r^2 = x^2 + y^2$, in $\Omega = \{(x, y) \ : \ 1/4 < r^2 < 4\}$. Then the circle $r^2 = 1$ is an isolated limit cycle of the field of subcharacteristics. There are no stationary points in Ω.   ♣

If $\partial\Omega$ is not smooth or Γ_- is not closed in Γ, or both, then we cannot expect to prove regularity results for the solution of the reduced problem that are globally valid in $\bar{\Omega}$. We therefore examine regularity on subsets of $\bar{\Omega}$. For each point $x \in \Omega \cup \bar{\Gamma}_-$, let

$$\tau_+(x) = \inf\{\tau > 0 : \xi_x(\tau) \notin \Omega\}$$

be the exit time. The *domain of influence* of each $\Gamma^* \subset \bar{\Gamma}_-$ is defined to be

$$D_{infl}(\Gamma^*) = \{\, \xi_x(\tau) : x \in \Gamma^*, 0 \le \tau \le \tau_+(x)\}.$$

It is generated by the characteristics through points of Γ^*. We have

Lemma 1.11. *[GFL$^+$83] Assume that Σ is a connected, smooth compact set with $\bar{\Sigma} \subset \Gamma_-$. Then the problem*

$$b(x) \cdot \nabla u_0^* + c(x)u_0^* = f(x) \quad in \ D_{infl}(\Sigma),$$
$$u_0^* = 0 \quad on \ \bar{\Sigma},$$

has a unique smooth solution u_0^* in $D_{infl}(\Sigma) \setminus U(I(\bar{\Omega}))$, where $U(I)$ is some open neighbourhood of I.

Let us remark finally that even if $I(\bar{\Omega})$ belongs to the domain of influence, it may not be possible to extend a smooth solution to all of $I(\bar{\Omega})$. This is demonstrated by the example

$$-x\frac{\partial v}{\partial x} - y\frac{\partial v}{\partial y} = f \quad in \ \Omega = \{(x, y) : \ x^2 + y^2 < 1\},$$
$$v = 0 \quad on \ \Gamma.$$

1.3 Asymptotic Expansions and Boundary Layers

Consider the boundary value problem

$$Lu := -\varepsilon \Delta u + b(x) \cdot \nabla u + c(x)u = f(x) \quad \text{in } \Omega \subset \mathbb{R}^d, \qquad (1.6a)$$
$$u = 0 \quad \text{on } \partial\Omega = \Gamma, \qquad (1.6b)$$

where the boundary Γ is regular. The discussion of properties of the solution u_0 of the reduced problem in the previous section leads us to expect that, as $\varepsilon \to 0$, the solution of (1.6) tends to the solution of the reduced problem only in some weak sense.

Theorem 1.12. *[Lio73] Assume that the hypotheses of Lemma 1.5 are satisfied. Then $u \rightharpoonup u_0$ weakly in $L_2(\Omega)$ as $\varepsilon \to 0$.*

Lemma 1.11 shows nevertheless that the solution u_0 of the reduced problem is smooth in certain subdomains of Ω; thus, away from boundary layers and from the union of all limit sets of subcharacteristics, we expect a better result. In [GFL+83] the following statement is proved:

Theorem 1.13. *Let Ω be a domain with smooth boundary and Σ a connected, compact set with $\Sigma \subset \Gamma_-$. Assume that $c > 0$ on $\bar{\Omega}$. Then there exists a constant C (independent of x and ε) such that*

$$|u(x) - u_0(x)| \leq C\varepsilon \quad \text{for } x \in D_{infl}(\Sigma) \setminus (U_\gamma(\Gamma_+) \cup U_\gamma(I(\bar{\Omega}))).$$

Furthermore, if all characteristics through points of $D_{infl}(\Sigma)$ leave the domain Ω at points of Γ_+ in finite time, then

$$|u(x) - u_0(x)| \leq C\varepsilon \quad \text{in } D_{infl}(\Sigma) \setminus U_{\gamma(\varepsilon)}(\Gamma_+),$$

where $\gamma(\varepsilon)$ satisfies

$$\lim_{\varepsilon \to 0} \frac{\varepsilon \ln \varepsilon}{\gamma(\varepsilon)} = 0.$$

Here $U_\gamma(G)$ denotes an open neighbourhood of G with $dist(G, \Omega \setminus U_\gamma(G)) \geq \gamma$.

If no limit sets exist, then all subcharacteristics through $\bar{\Gamma}_-$ leave $\bar{\Omega}$ in a finite time. Choose a fixed $\delta > 0$, independently of ε. Then the global domain

$$\Omega_\delta = \{x \in \Omega : dist(x, \Gamma_0 \cup \Gamma_+) > \delta\}$$

is far from possible boundary layers at Γ_0 and Γ_+ when ε is small. Theorem 1.13 now tells us that

$$|u(x) - u_0(x)| \leq C\varepsilon \quad \text{for } x \in \Omega_\delta. \qquad (1.7)$$

But if Γ is not smooth, then (1.7) is no longer true because subcharacteristics through corners cause additional difficulties. For this case Felgenhauer [Fel84] (see also [BCG87]) used a weak maximum principle to prove

$$|u(x) - u_0(x)| \leq C\varepsilon^{1/2} \quad \text{for all } x \in \Omega_\delta. \tag{1.8}$$

To improve the results mentioned so far in this chapter, local corrections are needed at the layers. These follow the same principles as in Parts I and II. Near Γ_+ (or connected smooth parts of Γ_+), let us introduce the local coordinates

$$(\rho, \varphi) = (\rho, \varphi_1, ..., \varphi_{d-1}),$$

where $\rho(x) := \text{dist}(x, \Gamma_+)$ and $0 < \rho < \rho_0$ corresponds to a strip parallel to Γ_+. In these new coordinates L is transformed into $\tilde{L}$, with

$$\tilde{L}u := -\varepsilon L_2 u + B_0(\rho, \varphi)\frac{\partial u}{\partial \rho} + \sum_{\mu=1}^{d-1} B_\mu(\rho, \varphi)\frac{\partial u}{\partial \varphi_\mu} + c(\rho, \varphi),$$

where L_2 is an elliptic operator and

$$B_0(0, \varphi) = b \cdot \nabla\rho|_{\rho=0} < 0$$

from the definition of Γ_+. Set $\zeta = \rho/\varepsilon$. The first term v_0 of a local correction satisfies the equation

$$A_0(0, \varphi)\frac{d^2 v_0}{d\zeta^2} + B_0(0, \varphi)\frac{dv_0}{d\zeta} = 0.$$

Thus there is an *exponential boundary layer* at Γ_+ as the ellipticity of L_2 forces $A_0(0, \varphi) < 0$. Written out explicitly, the boundary layer function is

$$v_0(\rho, \varphi) = -u_0|_{\Gamma_+} \exp\left(-\frac{B_0(0, \varphi)\rho}{A_0(0, \varphi)\varepsilon}\right).$$

At Γ_0 the procedure is analogous, but the boundary layer equations are more complicated because $B_0(0, \varphi) = 0$. Introducing

$$\xi = \frac{\rho}{\varepsilon^{1/2}},$$

one then obtains for the local correction v_0 the equation

$$A_0(0, \varphi)\frac{\partial^2 v_0}{\partial \xi^2} + \frac{\partial B_0}{\partial \rho}(0, \varphi)\xi\frac{\partial v_0}{\partial \xi} + \sum_{\mu=1}^{d-1} B_\mu(0, \varphi)\frac{\partial v_0}{\partial \varphi_\mu} + c(0, \varphi)v_0 = 0. \tag{1.9}$$

This is a parabolic partial differential equation so we say that there is a *parabolic boundary layer* or *characteristic layer* at Γ_0. See [GFL+83] for a general discussion of equation (1.9); Examples 1.16 and 1.24 below perform a detailed investigation of particular parabolic layers.

Remark 1.14. Parabolic layers may also occur as interior layers, e.g., along the exceptional subcharacteristics of Figures 1.1 and 1.2. For each interior layer is located along a subcharacteristic; consequently one gets $b \cdot \nabla\rho|_{\rho=0} = 0$ on introducing local coordinates in a neighbourhood of this subcharacteristic. ♣

For smooth domains with $\Gamma = \Gamma_+$ or $\Gamma = \Gamma_-$ or $\Gamma = \Gamma_0$ (unusual situations that rarely occur in practical applications), one can construct full asymptotic expansions and prove their validity [Eck79, GFL$^+$83]. In other cases each problem must be studied individually.

This chapter described the *principles* underpinning the construction of asymptotic expansions in several dimensions; even when the validity of these expansions cannot be proved, one nevertheless gains an insight into the behaviour of the solution of (1.6) when ε is small. This information is extremely valuable in devising effective numerical methods for (1.6).

To round off our exposition, we discuss two model problems in more detail.

Example 1.15. (Exponential boundary layers) Consider the boundary value problem

$$-\varepsilon \Delta u + b(x,y) \cdot \nabla u + c(x,y)u = f(x,y) \quad \text{in } \Omega = (0,1) \times (0,1),$$
$$u = 0 \quad \text{on } \Gamma.$$

Assume that the data are smooth and that $c \geq 0$ with

$$b = (b_1, b_2) \quad \text{where } b_1 > 0 \text{ and } b_2 > 0.$$

Then the subcharacteristics behave as in Figure 1.1 and the reduced problem is defined by
$$b \cdot \nabla u_0 + c u_0 = f, \quad u_0|_{x=0} = u_0|_{y=0} = 0.$$
We expect exponential boundary layers at $x = 1$ and at $y = 1$. The asymptotic approximation

$$u_{as}^*(x,y) := u_0(x,y) - u_0(1,y) \exp\left[-b_1(1,y)\frac{1-x}{\varepsilon}\right]$$
$$- u_0(x,1) \exp\left[-b_2(x,1)\frac{1-y}{\varepsilon}\right]$$

is inaccurate near the corner $(1,1)$ because the boundary layer terms overlap there. Consequently one adds a *corner layer* correction, which [Eck79] is given by a solution of

$$-\left(\frac{\partial^2 w}{\partial \xi^2} + \frac{\partial^2 w}{\partial \eta^2}\right) - b_1(1,1)\frac{\partial w}{\partial \xi} - b_2(1,1)\frac{\partial w}{\partial \eta} = 0 \quad \text{on } (0,\infty) \times (0,\infty),$$

where $\xi := (1-x)/\varepsilon$ and $\eta := (1-y)/\varepsilon$. One then obtains

$$u_{as}(x,y) := u_{as}^*(x,y) + u_0(1,1) \exp\left[-b_1(1,1)\frac{1-x}{\varepsilon}\right] \exp\left[-b_2(1,1)\frac{1-y}{\varepsilon}\right].$$

If $u_0 \in C^2(\Omega) \cap C(\bar{\Omega})$, the classical comparison principle yields

$$\|u - u_{as}\|_\infty \leq C\varepsilon,$$

where $\| \cdot \|_\infty$ is the maximum norm on $C(\bar{\Omega})$. But, as we already know, the assumption that $u_0 \in C^2(\Omega) \cap C(\bar{\Omega})$ is not always satisfied. Without this assumption, we get only (1.8); and in the ε-weighted H^1 norm defined by

$$\|v\|_\varepsilon^2 := \varepsilon |v|_1^2 + \|v\|_0^2$$

one can prove that

$$\|u - u_{as}\|_\varepsilon \leq C\varepsilon^{1/2}; \tag{1.10}$$

see [Sch84, Sch86] for related estimates. ♣

Example 1.16. (A parabolic boundary layer) Let us consider the boundary value problem

$$-\varepsilon \Delta u + \frac{\partial u}{\partial y} = f \quad \text{in } \Omega = (0,1) \times (0,1),$$

$$u = 0 \quad \text{on } \Gamma.$$

This popular example appears in most textbooks on asymptotic expansions (see, e.g., [Eck79]) as it is the simplest problem that exhibits a parabolic layer. Define u_0 by

$$\frac{\partial u_0}{\partial y} = f \quad \text{in } \Omega, \quad u_0|_{y=0} = 0.$$

We expect an exponential boundary layer at the outflow boundary $y = 1$, and parabolic boundary layers at the characteristic boundaries $x = 0$ and $x = 1$. Consider the boundary $x = 0$ (the boundary $x = 1$ is analogous). Introduce the variable $\xi := x/\varepsilon^{1/2}$. The first term of a local correction at $x = 0$ satisfies

$$-\frac{\partial^2 v_0}{\partial \xi^2} + \frac{\partial v_0}{\partial y} = 0 \quad \text{in } (0, \infty) \times (0,1),$$

$$v_0(0, y) = g(y) := -u_0(0, y),$$

$$v_0|_{y=0} = 0.$$

The solution of this standard parabolic initial-boundary value problem is

$$v_0(\xi, y) = \sqrt{\frac{2}{\pi}} \int_{\xi/\sqrt{2y}}^\infty \exp\left(-\frac{t^2}{2}\right) g\left(y - \frac{\xi^2}{2t^2}\right) dt. \tag{1.11}$$

One can hence deduce some typical features of parabolic layers:

- the *thickness* (corresponding to the stretching exponent α in the local variable $\xi = x/\varepsilon^\alpha$) of the parabolic layer is $\mathcal{O}(\varepsilon^{1/2})$, in contrast to $\mathcal{O}(\varepsilon)$ for an exponential layer;
- $|u|_{1,\Omega^*} = \mathcal{O}(\varepsilon^{-1/4})$ near a parabolic layer, while $|u|_{1,\Omega^0} = \mathcal{O}(\varepsilon^{-1/2})$ near an exponential layer (here Ω^* and Ω^0 denote small neighbourhoods of the respective layers);

- parabolic layers have a more complicated analytical structure than exponential layers.

A detailed analysis shows that derivatives of v_0 have singularities at $(0,0)$ and, furthermore, that at $(0,1)$ and $(1,1)$ the overlap of a parabolic and an exponential layer causes some difficulties. See [KS07a, Lel76, SK87] for more information.

The approximation

$$u_{as}(x,y) := u_0(x,y) - u_0(x,1) \exp\left(-\frac{1-y}{\varepsilon}\right),$$

which completely neglects the parabolic layers, nevertheless achieves [Sch86]

$$\|u - u_{as}\|_\varepsilon = \mathcal{O}(\varepsilon^{1/4}).$$

If the exponential layer were neglected similarly in Example 1.15, one would obtain only $\|u - u_0\|_\varepsilon \le C$. Thus, when measured in the ε-weighted H^1-norm, parabolic layers are less significant asymptotically.

If in this example one replaces $\dfrac{\partial u}{\partial y}$ by $\dfrac{\partial u}{\partial x}$, then the exponential layer moves to $x = 1$ and the parabolic boundary layers are at $y = 0$ and $y = 1$.   ♣

Remark 1.17. (Neumann outflow boundary conditions) As we saw already in Section I.1.1.1, the strength of the layer depends on the boundary conditions. With a homogeneous Neumann condition $\partial u/\partial n = 0$ at the outflow boundary Γ_+, one expects $\partial u/\partial n$ but not u to have large first-order derivatives. Therefore (assuming that Γ_0 is empty) if one uses the asymptotic expansion $\tilde{u}_{as} = u_0 + \varepsilon v_0$ with an exponential boundary layer correction v_0, one expects that

$$|u|_1 = \mathcal{O}(\varepsilon^{1/2}) \quad \text{and} \quad \|u - u_0\|_1 = \mathcal{O}(\varepsilon^{1/2}) \tag{1.12}$$

(cf. (1.10)), in contrast to the estimates

$$|u|_1 = \mathcal{O}(\varepsilon^{-1/2}) \quad \text{and} \quad \|u - u_0\|_1 = \mathcal{O}(\varepsilon^{-1/2})$$

that are typical of Dirichlet outflow boundary conditions. But we do not know of any rigorous proof of (1.12) for homogeneous Neumann outflow boundary conditions.   ♣

1.4 A Priori Estimates and Solution Decomposition

The method of matched asymptotic expansions, when applied to singularly perturbed problems of convection-diffusion type in several dimensions, does not in general provide enough information for the analysis of numerical methods, because a rigorous proof of the validity of an asymptotic approximation is available only in exceptional cases. Such proofs are especially difficult if one

works with a norm that includes derivatives. Consequently we regard asymptotic expansions as only an auxiliary technique that communicates to us some understanding of the nature of the problem. The derivative bounds that are needed for numerical analysis are derived directly by other means, as will now be demonstrated.

Consider the boundary value problem

$$Lu := -\varepsilon\Delta u + b(x) \cdot \nabla u + c(x)u = f(x) \quad \text{in } \Omega, \qquad (1.13a)$$
$$u = 0 \quad \text{on } \Gamma. \qquad (1.13b)$$

Let us introduce the ε-weighted norm

$$\|v\|_{\varepsilon,p} := \left\{ \int_\Omega (p-1)\varepsilon(\nabla v)^2 |v|^{p-2} + |v|^p \right\}^{1/p} \quad \text{for } 2 \le p < \infty,$$

while $\|\cdot\|_{\varepsilon,\infty}$ is defined to be the usual $L_\infty(\Omega)$ norm. This is a generalization of the norm $\|\cdot\|_\varepsilon$ (obtained by setting $p = 2$ here) that was used already.

Lemma 1.18. *Assume that*

$$c - \frac{1}{p}\operatorname{div} b \ge \omega_p > 0 \quad \text{on } \bar{\Omega}.$$

Then the solution u of the boundary value problem (1.13) satisfies

$$\|u\|_{\varepsilon,p} \le C\|f\|_{L_p} \quad \text{for } 2 \le p \le \infty. \qquad (1.14)$$

If in addition $u \in H^2(\Omega)$, then

$$\varepsilon^{3/2}\|u\|_2 + \varepsilon^{1/2}\|u\|_1 + \|u\|_0 \le C\|f\|_0. \qquad (1.15)$$

Proof. Multiply (1.13a) by u^{p-1}, integrate by parts and invoke Hölder's inequality to get (1.14). In particular for $p = 2$ one has

$$\varepsilon^{1/2}\|u\|_1 + \|u\|_0 \le C\|f\|_0.$$

This inequality and (1.13a) yield

$$\varepsilon\|\Delta u\|_0 \le C\varepsilon^{-1/2}\|f\|_0.$$

Then since $\|u\|_2 \le C(\|\Delta u\|_0 + \|u\|_0)$ (see, e.g., [LU68, Chapter I]), inequality (1.15) follows. □

For problems with exponential boundary layers, the estimate (1.15) is sharp, as can be inferred from the exposition of Section 1.3.

Remark 1.19. Using techniques similar to the proof of Lemma 1.18, further special estimates have been established in the literature:

(i) For the problem

$$-\varepsilon\Delta u + u_x = f(x,y) \quad \text{in } \Omega = (0,1)^2,$$
$$u = 0 \quad \text{on } \Gamma,$$

one has, in addition to (1.14) for $p = 2$, the bound

$$\|\phi^{1/2}u_x\|_0 + \left(\int_{\Gamma}\varepsilon\phi|\nabla u|^2|n|\,d\Gamma\right)^{1/2} \le C\|f\|_0.$$

Here ϕ is a cutoff function that vanishes on Γ_+ and has certain other properties; see [EJ93b, Lemma 1.2].

(ii) The solution of the problem

$$-\varepsilon\Delta u + u_x = f(x) \quad \text{in } \Omega = (0,1)^2,$$
$$u = 0 \quad \text{on } \Gamma_- \cup \Gamma_0,$$
$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma_+,$$

has a weak exponential outflow layer and parabolic boundary layers. Here, as well as the bounds of Lemma 1.18, one has [EJ93b, Lemma 1.1]

$$\|u_x\|_0 + \varepsilon|u|_2 + \left(\int_{\Gamma_+}u^2\,n\,d\Gamma\right)^{1/2} + \left(\int_{\Gamma}\varepsilon|\nabla u|^2|n|\,d\Gamma\right)^{1/2} \le C\|f\|_0.$$

(iii) If neither an outflow nor an inflow boundary is present – i.e., if $\Gamma_-\cup\Gamma_+$ is empty – then one can prove [AL90] that

$$\varepsilon\|u\|_2 + \|b\cdot\nabla u\|_0 + \|u\|_0 \le C(\|b\cdot\nabla f\|_0 + \|f\|_0).$$

(iv) Pointwise gradient estimates at the boundary and L_p estimates for the gradient can be found in [Doe98, Doe99a].

(v) Further *a priori* bounds on u and on its smooth and layer components are derived in [KS01a]. ♣

Lemma 1.18 bounds the L_2 norm of the gradient of the exact solution. In the one-dimensional case, we know from Theorem I.1.13 that in many cases the L_1 norm of this gradient is often bounded uniformly in ε. A similar result holds true in several dimensions.

Lemma 1.20. *[BBB73] Assume that the hypotheses of Lemma 1.5 are satisfied and that $f \in W^{1,1}(\Omega)$. Then there exists a constant C such that the solution u of (1.13) satisfies*

$$\|u\|_{W^{1,1}(\Omega)} \le C\|f\|_{W^{1,1}(\Omega)}.$$

In the one-dimensional case it was quite fruitful to study the properties of the Green's function associated with the differential operator in order to prove sharp stability results and *a priori* estimates. In two dimensions the behaviour of the Green's function is much more complicated:

Remark 1.21. (The Green's function in two dimensions) Let us consider (1.13) with constant coefficients while assuming that

$$b_1^2 + b_2^2 > 0 \quad \text{and} \quad c > 0.$$

Fix $(x, y) \in \Omega$. Assume that the domain Ω allows existence of a Green's function $G(x, y; \xi, \eta)$ and that $f \in L_p(\Omega)$. The representation

$$u(x, y) = \int_\Omega G(x, y; \xi, \eta) f(\xi, \eta) \, d\Omega$$

will be used to derive *a priori* estimates for u. Let $\tilde{G}$ be the free space Green's function that is defined on all of $\mathbb{R}^2$. A weak maximum principle argument shows that $0 \le G \le \tilde{G}$. This is helpful since $\tilde{G}$ is known explicitly: on setting $r^2 = (\xi - x)^2 + (\eta - y)^2)$ and $\lambda^2 = (b_1^2 + b_2^2 + 4\varepsilon c)/(2\varepsilon)$, one has

$$\tilde{G}(x, y; \xi, \eta) = \frac{1}{2\pi\varepsilon} \exp\left\{[b_1(\xi - x) + b_2(\eta - y)]/(2\varepsilon)\right\} K_0(\lambda r), \qquad (1.16)$$

where K_0 denotes the modified Bessel function of the second kind. The behaviour of K_0 is well known and enables us to deduce some properties of $\tilde{G}$:

- logarithmic singularity at $r = 0$
- fast exponential decay downwind of the point (x, y)
- slower $(\approx r^{-1/2})$ decay upwind of (x, y)
- symmetry with respect to the wind direction b.

Our Green's function G has the same properties. Using (1.16) one can show that

$$\|G\|_{L_p} \le C(p, b_1, b_2, c)\varepsilon^{-(p-1)/p} \quad \text{for } 1 \le p < \infty.$$

Then for (1.13) with constant coefficients the following estimates are valid:

$$\|u\|_\infty \le C\varepsilon^{-1/q}\|f\|_{L_q} \quad \text{for } 1 < q \le \infty,$$
$$\|u\|_{L_p} \le C\varepsilon^{-(p-1)/p}\|f\|_{L_1} \quad \text{for } 1 \le p < \infty;$$

see [DR90]. Unfortunately, it seems difficult to get similar information about the derivatives of G and thereby improved stability estimates.   ♣

Using an approximate Green's function, interesting anisotropic stability estimates are proved by Dörfler [Doe99a]. For simplicity, assume that $b_1 > \beta_1 > 0$. Then we call the x-direction the global stream direction. (In the general case, Dörfler introduces a new coordinate system representing the global stream direction and a direction orthogonal to it; as $b_1 > 0$ we can use the (x, y) coordinate system.)

We say that $w \in L_\mu \otimes L_{\underline{\nu}}$ if $q \mapsto \|w(q, \cdot)\|_{L_\nu} \in L_\mu$. That is, the underlined $\underline{\nu}$ indicates that the norm associated with the second variable is evaluated first.

Theorem 1.22. *Assume that $b_1 > \beta_1 > 0$. Then the solution of (1.13) satisfies the following anisotropic stability estimates:*

$$\|u\|_{\infty \otimes \underline{1}} \leq C\|f\|_1 \tag{1.17}$$

and

$$\|u\|_\infty \leq C\|f\|_{1 \otimes \infty}. \tag{1.18}$$

We sketch the proof of this result. Fix $(x, y) \in \Omega$. Let $\{\delta_n(\cdot; x, y)\}_{n=1}^\infty$ be a Dirac sequence associated with (x, y). This is a sequence of smooth positive functions, each with L_1 norm equal to 1, such that

$$|\text{supp } \delta_n| \to 0 \quad \text{and} \quad \int_\Omega u(\xi, \eta)\delta_n(\xi, \eta; x, y)) \, d\Omega \to u(x, y) \quad \text{as } n \to \infty.$$

For each $n \geq 1$, define an approximate Green's function G_n by

$$L^* G_n(x, y; \cdot) = \delta_n(\cdot, x, y) \text{ in } \Omega, \quad G_n(x, y; \cdot) = 0 \text{ on } \Gamma. \tag{1.19}$$

An integration by parts gives

$$\int_\Omega u(\xi, \eta)\delta_n(\xi, \eta; x, y) \, d\Omega = \int_\Omega G_n(x, y; \xi, \eta) \, f(\xi, \eta) \, d\Omega.$$

Assume for the moment that (1.17) is proved. Applying this result to (1.19), one gets

$$\sup_{x,y} \sup_\xi \int G_n(x, y; \xi, \eta) \, d\eta \leq C.$$

Letting $n \to \infty$, it follows that

$$|u(x, y)| \leq C\|f\|_{1 \otimes \infty}$$

and (1.18) is established. It remains to prove (1.17). Roughly speaking, this estimate follows if we first integrate the differential equation with respect to the second variable and then apply an (L_∞, L_1) stability result for the subsequent one-dimensional problem; see [Doe99a] for a full description.

If one is interested in the construction or analysis of uniformly convergent numerical methods, then the above *a priori* estimates do not yield enough information – more precise bounds on the derivatives are required. For first-order derivatives, taking a derivative of the differential equation then invoking a comparison principle and barrier function often delivers the desired bound.

Example 1.23. (Bounds for first-order derivatives: exponential layers)
We continue the discussion of Examples 1.3 and 1.15, assuming that all hypotheses of these examples are satisfied and in particular that the compatibility condition that f vanishes at each of the four corners of $\bar{\Omega}$ is fulfilled. Then for the boundary value problem (1.13),

- a differentiable classical solution exists;
- this solution has exponential boundary layers at $x = 1$ and $y = 1$.

One can prove the following *a priori* estimate [RAF96]: suppose that $c \geq \gamma$ with γ sufficiently large, and that

$$(b_2)_x \leq 0 \quad \text{and} \quad (b_1)_y \leq 0.$$

Then for all $(x, y) \in \Omega$, one has

$$|u_x(x, y)| \leq C(1 + \varepsilon^{-1} \exp(-\beta_1(1 - x)/\varepsilon)),$$
$$|u_y(x, y)| \leq C(1 + \varepsilon^{-1} \exp(-\beta_2(1 - y)/\varepsilon)),$$

where $b_1 > \beta_1 > 0$ and $b_2 > \beta_2 > 0$. This result generalizes older estimates [Lis83, OS89] that make the separability assumption $(b_2)_x = (b_1)_y = 0$. The proof uses a maximum principle for the elliptic system of equations satisfied by (u_x, u_y). ♣

Example 1.24. (Bounds for first-order derivatives: parabolic layers)
Let us consider the problem

$$-\varepsilon \Delta u + u_y + cu = f \quad \text{in } \Omega = (0, 1) \times (0, 1),$$
$$u = 0 \quad \text{on } \Gamma,$$

with $c \geq \gamma > 0$ and under conditions such that a differentiable classical solution exists in $C^{2,\alpha}(\bar{\Omega})$. As in Example 1.16, there are parabolic layers at $x = 0$ and $x = 1$ and an exponential layer at $y = 1$. Vulanović [Vul91] proves the estimates

$$|u_y(x, y)| \leq C(1 + \varepsilon^{-1} \exp(-\beta_2(1 - y)/\varepsilon)),$$
$$|u_x(x, y)| \leq C \left[1 + \varepsilon^{-1/2} \left(\exp(-\gamma^{1/2} x/\varepsilon^{1/2}) + \exp(-\gamma^{1/2}(1 - x)/\varepsilon^{1/2}) \right) \right].$$

where $0 < \beta_2 < 1$. The last bound clearly demonstrates the influence of the parabolic layer and shows that $|u_x|$ can be estimated using standard exponentials instead of the complicated function v_0 of (1.11).

It seems difficult to extend this analysis to higher-order derivatives because differentiation of the differential equation leads to difficulties. ♣

Is there a suitable *decomposition of the solution of our elliptic boundary value problem,* as in Parts I and II, into smooth and layer components?

In [Shi92c] Shishkin derived a solution decomposition for elliptic problems. We first describe his approach for the problem with exponential layers, i.e., for the problem (1.13):

$$Lu := -\varepsilon \Delta u + b(x, y) \cdot \nabla u + c(x, y)u = f(x, y) \quad \text{in } \Omega = (0, 1)^2,$$
$$u = 0 \quad \text{on } \Gamma,$$

with $b_1 > \beta_1 > 0$, $b_2 > \beta_2 > 0$. Assume that b_1, b_2, c and f are smooth. Then the solution u can be decomposed as

$$u = S + E \quad \text{with } S = u_0 + \varepsilon u_1 + u_2^*, \quad E = E_0 + E_1^*. \qquad (1.20)$$

The components of the smooth part S and the layer part E are defined by

$$
\begin{aligned}
L_0 u_0 &= f, \quad u_0|_{\Gamma_-} = 0, \\
L_0 u_1 &= -L_2 u_0, \quad u_0|_{\Gamma_-} = 0, \quad \text{(writing } L = \varepsilon L_2 + L_0\text{)} \\
L u_2^* &= -\varepsilon^2 L_2 u_1, \quad u_2^*|_\Gamma \text{ given,}
\end{aligned}
$$

and

$$
\begin{aligned}
L E_0 &= 0, \quad E_0|_\Gamma = -(u_0 + \varepsilon u_1)|_\Gamma, \\
L E_1^* &= 0, \quad E_1^*|_\Gamma = -u_2^*|_\Gamma.
\end{aligned}
$$

But the details of this analysis in [Shi92c, Chapter III, p.131] contain several misprints and the functions u_2^*, E_1^* (which Shishkin calls v_1, v_2) are not defined there but in [Shi92c, Appendix C, Section 2]. While u_0 and u_1 are standard terms in an asymptotic expansion, the other terms are defined are nonstandard. The layer component E_0 can be decomposed further into two exponential layers and a corner layer:

$$E_0 = E_1 + E_2 + E_{12}. \qquad (1.21)$$

Let $r((x, y), \Gamma)$ denote the distance from a point (x, y) to the boundary.

Theorem 1.25. *[Shi92c] In addition to the hypotheses above, assume that $u_0, u_1 \in C^{3,\alpha}(\bar{\Omega})$. Then the solution of (1.13) can be decomposed as (1.20)–(1.21) and its components satisfy the following estimates for $0 \le k \le 3$:*

$$\left| \frac{\partial^k S(x, y)}{\partial x^{k_1} \partial y^{k_2}} \right| \le C \left[1 + \varepsilon^{2-k} + \varepsilon^2 \, r^{-k}((x, y), \Gamma) \right], \qquad (1.22a)$$

$$\left| \frac{\partial^k E_1(x, y)}{\partial x^{k_1} \partial y^{k_2}} \right| \le C \left[\varepsilon^{-k_1} + \varepsilon^{1-k} + r^{-k}((x, y), \Gamma) \right] e^{-\beta_1(1-x)/\varepsilon}, \qquad (1.22b)$$

$$\left| \frac{\partial^k E_2(x, y)}{\partial x^{k_1} \partial y^{k_2}} \right| \le C \left[\varepsilon^{-k_2} + \varepsilon^{1-k} + r^{-k}((x, y), \Gamma) \right] e^{-\beta_2(1-y)/\varepsilon}, \qquad (1.22c)$$

$$\left| \frac{\partial^k E_{12}(x, y)}{\partial x^{k_1} \partial y^{k_2}} \right| \le C \, T(\varepsilon, r) \, e^{-[\beta_1(1-x)+\beta_2(1-y)]/\varepsilon} \qquad (1.22d)$$

where $k = k_1 + k_2$ and

$$T := \varepsilon^{-k} + r^{-k}((x, y), \Gamma).$$

If moreover one has $u_2^*, E_1^* \in C^{3,\alpha}(\bar{\Omega})$, then the terms containing $r((x, y), \Gamma)$ can be omitted.

The proof of Theorem 1.25 is sketched in [Shi92c, Appendix C], but the arguments are presented in a very concise way and it is difficult to understand the precise steps used. Furthermore, it is not easy to check if $u_2^*, E_1^* \in C^{3,\alpha}(\bar{\Omega})$.

Extending an argument first outlined in [DR97], Linß and Stynes [LS01a] presented the following precise sufficient conditions for the validity of a decomposition in the sense of the above theorem. Define the notation

$$\mathcal{L}_i v := \frac{\partial v}{\partial y} \frac{\partial^i}{\partial x^i}\left(\frac{b_2}{b_1}\right) + v \frac{\partial^i}{\partial x^i}\left(\frac{c}{b_1}\right) \quad \text{for } i = 0, 1, \dots,$$

and let $\|\cdot\|_{\nu,\alpha}$ denote the $C^{\nu,\alpha}(0,1)$ norm with respect to $(\cdot)$.

Theorem 1.26. *(S-type decomposition) [LS01a] Consider the boundary value problem (1.13) in the unit square with $b_1 > \beta_1 > 0$, $b_2 > \beta_2 > 0$ (so the solution has only exponential boundary layers) and $f \in C^{4,\alpha}(\bar{\Omega})$. Suppose that f satisfies the compatibility conditions*

$$f(0,0) = f(0,1) = f(1,0) = f(1,1) = 0$$

and

$$\left(\frac{f}{b_1}\right)_y (0,0) = \left(\frac{f}{b_2}\right)_x (0,0),$$

$$\left(\left(\frac{f}{b_1}\right)_x - \mathcal{L}_0\left(\frac{f}{b_1}\right)\right)_y (0,0) = \left(\frac{f}{b_2}\right)_{xx} (0,0),$$

$$\left[\left(\frac{f}{b_1}\right)_{xx} - \mathcal{L}_0\left(\left(\frac{f}{b_1}\right)_x - \mathcal{L}_0\left(\frac{f}{b_1}\right)\right) - 2\mathcal{L}_1\left(\frac{f}{b_1}\right)\right]_y (0,0) = \left(\frac{f}{b_2}\right)_{xxx} (0,0),$$

$$\left(b_1\left(\frac{f}{b_1}\right)_{yy}\right) (0,0) = \left(b_2\left(\frac{f}{b_2}\right)_{xx}\right) (0,0).$$

Let $n \geq 2$ be an integer. If $n \geq 4$, assume in addition that

$$b_{2,x}(1,1) = b_{1,y}(1,1).$$

Then the given boundary value problem has a classical solution $u \in C^{3,\alpha}(\bar{\Omega})$ that can be decomposed as

$$u = S + E_1 + E_2 + E_{12} \tag{1.23}$$

where

$$\|S\|_{C^2} + \varepsilon^\alpha \|S\|_{C^{2,\alpha}} \leq C, \tag{1.24}$$

and for all $x, y \in [0,1]$ one has

$$\left\|\frac{\partial^{k_1}}{\partial x^{k_1}} E_1(x, \cdot)\right\|_{\nu,\alpha} \leq C\varepsilon^{-k_1} e^{-\beta_1(1-x)/\varepsilon}, \tag{1.25a}$$

$$\left\|\frac{\partial^{k_2}}{\partial y^{k_2}} E_2(\cdot, y)\right\|_{\mu,\alpha} \leq C\varepsilon^{-k_2} e^{-\beta_2(1-y)/\varepsilon}, \tag{1.25b}$$

and

$$\left| \frac{\partial^k}{\partial x^{k_1} \partial y^{k_2}} E_{12}(x,y) \right| \leq C\,\varepsilon^{-k}\,e^{-(\beta_1(1-x)+\beta_2(1-y))/\varepsilon} \tag{1.26}$$

for $0 \leq \nu, \mu \leq 2$ and $0 \leq k_1, k_2 \leq n$. Moreover, for all $(x,y) \in \Omega$ one has

$$|LE_1(x,y)| \leq C\varepsilon e^{-\beta_1(1-x)/\varepsilon}, \tag{1.27a}$$

$$|LE_2(x,y)| \leq C\varepsilon e^{-\beta_2(1-y)/\varepsilon}, \tag{1.27b}$$

and

$$|LE_{12}(x,y)| \leq C\,\varepsilon\,e^{-(\beta_1(1-x)+\beta_2(1-y))/\varepsilon}. \tag{1.27c}$$

A similar decomposition in [Kop03] requires less compatibility at the corner $(0,0)$.

For problems with parabolic boundary layers the situation is different. Let us study the typical problem

$$Lu := -\varepsilon \Delta u + u_x = f \quad \text{in } \Omega = (0,1) \times (0,1) \tag{1.28a}$$

$$u = 0 \quad \text{on } \Gamma, \tag{1.28b}$$

whose solution u has in general parabolic boundary layers at $y = 0$ and $y = 1$.

The decomposition

$$u = u_0 + u_1^* + v_0 + v_1^* + w_0 + w_1^* \tag{1.29}$$

from [Shi92c] contains the leading terms u_0, v_0, w_0 of an asymptotic expansion, representing the solution of the reduced problem, the exponential and the parabolic boundary layer correction, and also the nonstandard terms u_1^*, v_1^*, w_1^* that guarantee that the layer terms are in the null space of L. But [Shi92c, Appendix D, Section 1, Theorem 1.1] gives only estimates for the derivatives of u_0, v_0, w_0 and, under some compatibility assumptions, the bounds,

$$\|u - u_0\|_\infty \leq C\varepsilon, \quad \|v - v_0\|_\infty \leq C\varepsilon^{1/3}, \quad \|w - w_0\|_\infty \leq C\varepsilon^{1/5}. \tag{1.30}$$

These estimates are too weak to give optimal-order error estimates for numerical methods.

The use of standard asymptotic expansions in the construction of decompositions has the disadvantage that they entail the use of hyperbolic or parabolic operators whose regularity properties differ from those of elliptic operators. It is therefore attractive to construct decompositions whose terms are solutions of elliptic problems – a so-called *elliptic decomposition*. This idea is related to the use of *extended domains*; see [OS07b, OS07a, Roo02].

The solution of the boundary value problem (1.28) is now decomposed as

$$u = S + E_1 + E_2 + E_3 \tag{1.31}$$

where the smooth component S is the solution of, for instance (other extended domains are possible),

$$LS = f^* \text{ in the half plane } x > 0,$$
$$u = 0 \text{ on } x = 0.$$

Here and in what follows g^* denotes a smooth extension of g with compact support. The component E_1 representing the exponential layer at $x = 1$ is defined by

$$LE_1 = 0 \text{ in the strip } 0 < x < 1,$$
$$E_1 = 0 \text{ on } x = 0, \quad E_1 = -S^* \text{ on } x = 1.$$

The parabolic boundary layers are contained in E_2, which satisfies

$$LE_2 = 0 \text{ in } x > 0, \ 0 < y < 1,$$
$$E_2 = 0 \text{ on } x = 0, \quad E_2 = -S^* \text{ on } y = 0, \ y = 1.$$

Finally E_3, the corner layer component:

$$LE_3 = 0 \text{ in } \Omega.$$
$$E_3 = -E_2 \text{ on } x = 1, \quad E_3 = 0 \text{ on } x = 0,$$
$$E_3 = -E_1 \text{ on } y = 0 \text{ and } y = 1.$$

Then one expects under certain compatibility assumptions the following estimates to hold for certain i, j and $0 < \alpha < 1$:

$$\left| \frac{\partial^{i+j}}{\partial x^i \partial y^j} S(x, y) \right| \leq C, \tag{1.32a}$$

$$\left| \frac{\partial^{i+j}}{\partial x^i \partial y^j} E_1(x, y) \right| \leq C\varepsilon^{-i} e^{-\alpha(1-x)/\varepsilon}, \tag{1.32b}$$

$$\left| \frac{\partial^{i+j}}{\partial x^i \partial y^j} E_2(x, y) \right| \leq C\varepsilon^{-j/2} B(y), \tag{1.32c}$$

$$\left| \frac{\partial^{i+j}}{\partial x^i \partial y^j} E_3(x, y) \right| \leq C \varepsilon^{-(i+j/2)} e^{-\alpha(1-x)/\varepsilon} B(y) \tag{1.32d}$$

with

$$B(y) := \exp(-\gamma^* y/\sqrt{\varepsilon}) + \exp(-\gamma^*(1-y)/\sqrt{\varepsilon})$$

and some $\gamma^* > 0$. Under sufficient compatibility assumptions, the above ideas are used in [KS05, KS07a] to prove that the estimates (1.32) are valid, but the construction and estimation of the decomposition there is complicated (it includes the cases of little or no corner compatibility) and minimal sufficient conditions for these estimates are unknown. A related problem is examined in [KS06], where a jump discontinuity in an inflow boundary condition generates

an interior characteristic layer. This is a test case that is often mentioned in the literature but is difficult analytically since the solution of such a problem does not lie in $H^1(\Omega)$.

For the important practical case of a Neumann outflow boundary condition, see [NKS08].

Remark 1.27. (Solution decomposition for reaction-diffusion problems) Let us consider the two-dimensional reaction-diffusion problem

$$Lu := -\varepsilon \Delta u + c(x,y)u = f(x,y) \quad \text{in } \Omega \subset \mathbb{R}^2,$$
$$u = g \quad \text{on } \Gamma, \tag{1.33}$$

where $c > \gamma > 0$ and g is continuous. Under suitable regularity hypotheses on c, f and g one has $u \in C^\alpha(\bar{\Omega})$ or even $u \in C^{1,\alpha}(\bar{\Omega})$, but in general one does not obtain $u \in C^{2,\alpha}(\bar{\Omega})$ for domains with non-smooth boundary; see Example 1.3 and [Gri85b].

The classical theory of matched asymptotic expansions gives

$$u = S + E^{BL} + E^{CL} + R, \tag{1.34}$$

where E^{BL} contains several boundary layer terms and E^{CL} the corner layer terms while R is a remainder. For a rectangular domain, say $\Omega = (0,1)^2$, Butuzov [But75] proved that for each $n \geq 0$ there exists an asymptotic expansion (1.34) such that

$$\|R\|_\infty \leq C_n \, \varepsilon^{n+1}.$$

Han and Kellogg [HK90] extended this analysis by showing that S, E^{BL} and R are smooth and consequently any corner singularities are contained in the corner layer terms. With the notation

$$d_\nu^2(x,y) := \min \left\{ x^2 + y^2, (1-x)^2 + y^2, x^2 + (1-y)^2, (1-x)^2 + (1-y)^2 \right\}$$
$$d_s(x,y) := \min\{x, y, 1-x, 1-y\}$$

they derived, for the case of constant c, the following estimates for derivatives of the solution of the given reaction-diffusion problem:

$$|D_{xy}^m u(x,y)| \leq C + C\varepsilon^{-m/2} E_{bl} + \begin{cases} C \, E_{cl} & \text{if } m = 1, \\ C[1 + |\ln(d_\nu/(\varepsilon^{1/2})|] \, E_{cl} & \text{if } m = 2, \\ C(d_\nu/(\varepsilon^{1/2}))^{-(m-2)} \, E_{cl} & \text{if } m \geq 3, \end{cases}$$

where

$$E_{bl} = e^{-\gamma d_s(x,y)/\varepsilon^{1/2}}, \qquad E_{cl} = e^{-\gamma d_\nu/\varepsilon^{1/2}}.$$

As Andreev [And06] pointed out, singularities in the fourth-order derivatives appear only in the mixed derivatives: the pure derivatives satisfy

$$\left| \frac{\partial^4 u}{\partial x^4} \right| \leq C\varepsilon^{-2}, \qquad \left| \frac{\partial^4 u}{\partial y^4} \right| \leq C\varepsilon^{-2}.$$

The paper [And06] also contains a critical discussion of [HK90, Theorem 3.3] (in the case of variable c) and of the decomposition of [CGO05, Theorem 2.2].

More general problems can be found in [Shi92c] and in the book [Mel02] where Melenk allows the domain to be a curvilinear polygon whose boundary Γ is assumed to consist of finitely many curves. While the proof of a decomposition in [Shi92c, Section I.3] is incomplete, Melenk [Mel02, Theorem 2.3.4] gives full information about the behaviour of S, E^{BL}, E^{CL} and R.   ♣

2

Finite Difference Methods

2.1 Finite Difference Methods on Standard Meshes

2.1.1 Exponential Boundary Layers

Consider the convection-diffusion problem

$$Lu := -\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega := (0,1) \times (0,1), \qquad (2.1a)$$
$$u = 0 \quad \text{on } \partial \Omega, \qquad (2.1b)$$

with $0 < \varepsilon \leq 1$, $b = (b_1(x,y), b_2(x,y)) > (\beta_1, \beta_2) > (0,0)$ on $\bar{\Omega}$ and $c \geq 0$ on $\bar{\Omega}$. Assume that the data of the problem are smooth.

Example 1.15 shows that in general the solution u has exponential boundary layers at the sides $x = 1$ and $y = 1$ of $\bar{\Omega}$. Just like problems in one dimension, these layers cause serious instabilities in standard difference schemes.

The solution u_0 of the *reduced problem* is defined by:

$$L^0 u_0 := b \cdot \nabla u_0 + cu_0 = f \quad \text{in } \Omega,$$
$$u_0(x,0) = 0 \quad \text{on } \{(x,0) : 0 \leq x \leq 1\},$$
$$u_0(0,y) = 0 \quad \text{on } \{(0,y) : 0 < y \leq 1\}.$$

This is the same type of problem as (II.3.1), and much of Section II.3.1 still applies here. We again refer to the characteristic curves of the reduced problem as the *subcharacteristics* of (2.1). Any satisfactory scheme for (2.1) should, on setting $\varepsilon = 0$, become a satisfactory scheme for the reduced problem. This implies that one should use upwinded schemes, i.e., schemes that are equivalent to approximating the convection terms $(b \cdot \nabla u)(x,y)$ by means of asymmetrical finite differences that are centered on some point upstream of (x,y).

Let M and N be positive integers. Let $0 = x_0 < x_1 < \ldots < x_M = 1$ and $0 = y_0 < y_1 < \ldots < y_N = 1$. Throughout Chapter 2, we consider rectangular tensor-product grids on Ω, whose nodes are (x_i, y_j) for $i = 0, \ldots, M$ and

$j = 0, \ldots, N$. Set $h_i = x_i - x_{i-1}$ for each i and $k_j = y_j - y_{j-1}$ for each j. Let $h = \max\{h_i\}$ and $k = \max\{y_j\}$. Given any function v that is defined on the grid, v_{ij} denotes $v(x_i, y_j)$. In each scheme considered, u_h denotes the computed solution and the computed nodal values are written as u_{ij}.

We shall follow standard practice by working within a discrete L_∞-norm framework in Section 2.1.

A scheme $L_h v = \tilde{f}$ is said to be *consistent of $\mathcal{O}(\alpha)$* with (2.1) in the maximum norm if

$$\|L_h u - \tilde{f}\|_{\infty,d} \le K(h^\alpha + k^\alpha) \quad \text{as } h, k \to 0, \tag{2.2}$$

where K and α are fixed constants. (For convenience the restriction operator R_h that restricts u to the mesh is ignored.) Here $\| \cdot \|_{\infty,d} = \max_{ij} |(\cdot)_{ij}|$ is the discrete $L_\infty(\Omega)$ norm. If (2.2) is proved only under the assumption that ε is constant, then the scheme is said to be *formally consistent of $\mathcal{O}(\alpha)$*.

The scheme $L_h v = \tilde{f}$, $v|_{\partial\Omega} = 0$, is L_∞ *stable* if its solution v satisfies

$$\|v\|_{\infty,d} \le K\|\tilde{f}\|_{\infty,d}$$

for all mesh functions $\tilde{f}$ and some constant K.

Stability is often proved by the following argument described already in Part I. Suppose that L^h is *inverse-monotone*, i.e., $(L^h)^{-1} \ge 0$. Suppose also that we can find a mesh function ϕ and constants K_1 and K_2 such that $0 \le \phi_{ij} \le K_1$ and $0 < K_2 \le (L_h\phi)_{ij}$ for all i and j. (The choice $\phi(x,y) = 1+y$ often works for (2.1a), since b_2 is positive.) Then

$$L_h(\phi\|\tilde{f}\|_{\infty,d}/K_2)_{ij} \ge \|\tilde{f}\|_{\infty,d} \ge \tilde{f}_{ij},$$

so inverse-monotonicity implies that

$$\|v\|_{\infty,d} \le \|\tilde{f}\|_{\infty,d}\|\phi\|_{\infty,d}/K_2 \le (K_1/K_2)\|\tilde{f}\|_{\infty,d}.$$

Here ϕ acts as a *barrier function*.

Construction of Difference Schemes and Upwinding

A common way of generating schemes for (2.1) is to take some stable scheme for the reduced problem (see Section II.3.1) and to it add a standard difference approximation of $-\varepsilon\Delta u$. Our first example is in this vein.

Example 2.1. (Simple upwind scheme) Example II.3.4 discussed the simple upwind scheme for (II.3.1). To approximate (2.1), we generalize this scheme to

$$-\frac{2\varepsilon}{h_i + h_{i+1}}\left(\frac{u_{i+1,j} - u_{ij}}{h_{i+1}} - \frac{u_{ij} - u_{i-1,j}}{h_i}\right)$$
$$-\frac{2\varepsilon}{k_j + k_{j+1}}\left(\frac{u_{i,j+1} - u_{ij}}{k_{j+1}} - \frac{u_{ij} - u_{i,j-1}}{k_j}\right)$$
$$+ (b_1)_{ij}\frac{u_{ij} - u_{i-1,j}}{h_i} + (b_2)_{ij}\frac{u_{ij} - u_{i,j-1}}{k_j} + c_{ij} = f_{ij} \tag{2.3a}$$

for $i = 1, \ldots, M - 1$ and $j = 1, \ldots, N - 1$, with

$$u_{ij} = 0 \quad \text{when } (x_i, y_j) \in \partial\Omega. \tag{2.3b}$$

This is *simple upwinding*. The matrix of the scheme is an M-matrix and irreducibly diagonally dominant [OR70], so the discrete problem (2.3) is inverse-monotone and has a unique solution. One can find a barrier function (consider $1 + x_i$) and prove L_∞ stability uniformly with respect to the parameter ε.

It is straightforward to verify that the scheme is formally only first-order consistent. It is stable, but will smear layers as it did already in one dimension. In practice it is usually first-order accurate away from layers. We do not prove this rigorously (see [Tob83] for a closely related result), but merely observe that on equidistant meshes with $\varepsilon \ll \min\{h, k\}$, the scheme is very similar to the scheme of [GS93], whose local behaviour was described in Example II.4.6.

When the condition $b > (0, 0)$ is violated, the continuous problem may be unstable for $c \equiv 0$. Then simple upwinding is also unstable and may give a seriously inaccurate solution. Brandt and Yavneh [BY91] discuss such an example of linearized recirculating flow with $c \equiv 0$, where Ω is an annulus, the subcharacteristics are circles and, except near $\partial\Omega$, the solution of an analogue of (2.3a) is $\mathcal{O}(1)$ distant from the solution of the differential equation. We saw in Example I.2.35 that this effect is already possible for one-dimensional problems with turning points. ♣

An alternative approach to the construction of schemes for (2.1) is to take a scheme for the one-dimensional analogue of (2.1), then form the "tensor product" of that scheme in two dimensions. Simple upwinding fits into this category also; it is the tensor product of the upwind scheme (I.2.12). Solutions of tensor-product schemes generally suffer from excessive smearing of layers. To see why this happens, consider Example 2.1. When ε is much smaller than the local mesh size, u_{ij} is essentially computed from $u_{i-1,j}$ and $u_{i,j-1}$, while information from the natural candidate $u_{i-1,j-1}$ is not used – even in the case when the subcharacteristic through (x_i, y_j) also passes through (x_{i-1}, y_{j-1}). This means that information is moved *across* the subcharacteristics rather than *along* them. Consequently the method will not compute sharp internal layers.

A third technique used in devising difference schemes for (2.1) is to artificially increase the diffusion coefficient ε, then to apply some standard non-upwinded scheme (cf. Section I.2.1.2). For example, in the case of a square mesh, replacing ε in (2.1a) by $\varepsilon + h$ then applying a central difference scheme yields again (2.3a). Here we have added isotropic diffusion to the problem; that is, when modifying the differential equation, all directions were treated in exactly the same manner.

Numerical experience shows that isotropic artificial diffusion can make internal layers over-diffuse and that it is sufficient to add diffusion only in the direction of the subcharacteristics [KNZ80]. Hegarty [Heg82] adds anisotropic

diffusion of this type in the following systematic way, drawing inspiration from the familiar one-dimensional case. (See also [Lay93].)

Assume that b_1 and b_2 are constants and that, without loss of generality, $b_1 \geq b_2$. Transform coordinates from (x, y) to (ζ, η) by rotating the axes so that η is constant along each subcharacteristic while ζ is constant along lines perpendicular to subcharacteristics. That is,

$$\zeta := \frac{b_1 x + b_2 y}{\sqrt{b_1^2 + b_2^2}} \quad \text{and} \quad \eta := \frac{b_2 x - b_1 y}{\sqrt{b_1^2 + b_2^2}}.$$

The given equation can be written in terms of these variables as

$$-\varepsilon \hat{u}_{\zeta\zeta} - \varepsilon \hat{u}_{\eta\eta} + \sqrt{b_1^2 + b_2^2}\,\hat{u}_\zeta + c\hat{u} = \hat{f}, \tag{2.4}$$

where $\hat{u}(\zeta, \eta) := u(x, y)$. The idea now is to treat

$$\hat{u} \mapsto -\varepsilon \hat{u}_{\zeta\zeta} + \sqrt{b_1^2 + b_2^2}\,\hat{u}_\zeta$$

as a one-dimensional differential operator.

We work on a square mesh. Hegarty's first scheme imitates the one-dimensional Il'in-Allen-Southwell scheme (see Section I.2.1.3) by altering $-\varepsilon \hat{u}_{\zeta\zeta}$ to $-(H/2)\coth[H/(2\varepsilon)]\,\hat{u}_{\zeta\zeta}$, where $H := h\sqrt{b_1^2 + b_2^2}/b_1$ is the effective mesh width in the subcharacteristic direction ζ. The term $-\varepsilon \hat{u}_{\eta\eta}$ becomes $-\hat{\varepsilon}\hat{u}_{\eta\eta}$, where $\hat{\varepsilon}$ is yet to be determined. He then transforms back to the (x, y) variables and introduces a central difference approximation. An examination of the truncation error motivates the choice of $\hat{\varepsilon}$ (when $b_1 = b_2$, it is $b_1 h/\sinh(b_1 h/\varepsilon)$). Numerical results in [Heg82] show that this scheme gives sharp internal boundary layers, but its solution may exhibit oscillations.

To exclude oscillations, Hegarty develops an alternative scheme, which is of positive type (see Remark 2.3). Again transform (2.1a) to (2.4), but now after adding the same artificial diffusion as before, apply central differencing in the ζ and η variables. In general, this introduces points not on the original mesh; values at these points are approximated by linear interpolation. Set $\chi = b_2/b_1$ and $\theta = H/(2\varepsilon)$. At each $(x_i, y_j) \in \Omega$, the scheme is

$$\frac{b_1 h}{2}(\coth\theta)[(1-\chi)\delta_{xx}u_{ij} + \chi\delta_{\xi\xi}u_{ij}]$$
$$+ \frac{b_1 h}{2}(\sinh\theta)^{-1}[(1-\chi)\delta_{yy}u_{ij} + \chi\delta_{\eta\eta}u_{ij}]$$
$$+ b_1[(1-\chi)\delta_x u_{ij} + \chi\delta_\xi u_{ij}] + c_{ij}u_{ij} = f_{ij}.$$

Here $\delta_{xx}u$ and $\delta_{yy}u$ are the standard second-order difference approximations to u_{xx} and u_{yy} respectively, while

$$\delta_\xi u_{ij} := (u_{i+1,j+1} - u_{i-1,j-1})/2h,$$
$$\delta_{\xi\xi}u_{ij} := (u_{i+1,j+1} - 2u_{ij} + u_{i-1,j-1})/h^2$$

and

$$\delta_{\nu\nu} u_{ij} := (u_{i+1,j-1} - 2u_{ij} + u_{i-1,j+1})/h^2$$

approximate derivatives in the north-east and south-east directions. This scheme yields more diffuse internal layers than the first scheme [Heg82], but its solutions are oscillation-free.

Remark 2.2. When the diffusion coefficient is altered *a priori*, the extra diffusion added is called *artificial viscosity* (AVIS). This AVIS can easily be quantified from an inspection of the consistency error of the scheme. Bank et al. [BBF90] work with a variant of AVIS when they compare various upwind methods with the standard Galerkin method by inspection of the 3×3 element stiffness matrix associated with the diffusion term.

We carefully distinguish between AVIS – which appears in the difference scheme – and *numerical viscosity* (NVIS), which is said to be present when the numerical solution of the scheme has excessively diffuse layers. There is no standard definition of NVIS. In [To93b] it is specified as that increase in the diffusion parameter such that the exact solution of the resulting differential equation is closest (in some norm) to the computed numerical solution of the original problem. It is in general difficult to compute NVIS precisely, unlike AVIS – but NVIS is much more important, because a numerical solution is accurate if and only if its NVIS is small.

The two quantities are often confused in the literature. In particular, the amount of AVIS is sometimes taken as a measure of the amount of NVIS. This is fallacious as we know from Section I.2.1.3: the Il'in-Allen-Southwell scheme, applied to a two-point boundary value problem with constant coefficients has non-zero AVIS, but it yields exact nodal solutions (i.e., it has zero NVIS!). ♣

Remark 2.3. On an equidistant mesh, consider a consistent difference scheme $L^h v = \tilde{f}$ whose matrix (a_{ij}) is of *positive type*. Such a matrix satisfies the following conditions:

- $a_{ii} > 0$; $a_{ij} \leq 0$ when $i \neq j$; $a_{ii} \geq -\sum_{j\neq i} a_{ij}$ for all i with strict inequality for at least one i;
- the matrix is irreducible.

For example, the matrix of the simple upwind scheme has these properties. Schemes of positive type are at best first-order consistent when ε is small compared to $\min\{h, k\}$. This follows from Godunov's analogous result for the reduced problem (II.3.1), which can be proved by applying the technique of [Yse83] (see also [Lax61, Wid71]. ♣

If the scheme $L^h v = \tilde{f}$ is of positive type, then L^h is an M-matrix [OR70] and hence inverse-monotone. Inverse-monotone schemes are usually well behaved and can often be analysed using barrier functions. The easiest way to obtain such schemes is to use matrices of positive type or M-matrices – but then Remark 2.3 imposes a restriction of moderate accuracy on the scheme. It is a challenging problem to construct inverse-monotone schemes that are second-order accurate when ε is small compared with the local mesh diameter.

One such scheme is given by Kratsch and Roos [KR92] for the case when b_1 and b_2 are constant. Their difference scheme matrix is not of positive type but it can be expressed as a product of two M-matrices and consequently is inverse-monotone. It is formally second-order consistent.

A similar idea appears in the LECUSSO scheme of Günther [Gün92], but he writes the scheme as a product of M-matrices only in the one-dimensional case. This paper also provides an interesting comparison of numerical results for various schemes applied to a demanding heat flow problem in two space dimensions.

One can use local properties of the problem to determine the set of points used in the stencil of the difference scheme. Our next example is of this type.

Example 2.4. Roe and Sidilkover [RS92] use a variable stencil approach to develop a scheme of positive type on a uniform square mesh for a time-dependent constant coefficient first-order hyperbolic problem in two space dimensions. Their scheme minimizes truncation error within the class of four-point schemes of positive type. (They show that simple upwinding *maximizes* the truncation error in the same class.)

We generalize their scheme to (2.1a) by adding suitable approximations of cu and f and a standard difference approximation of $-\varepsilon \Delta u$. Then at each $(x_i, y_j) \in \Omega$, one gets the scheme

$$
\begin{aligned}
&- \varepsilon \frac{u_{i+1,j} - 2u_{ij} - u_{i-1,j}}{h^2} - \varepsilon \frac{u_{i,j+1} - 2u_{ij} - u_{i,j-1}}{h^2} \\
&+ (b_1)_{ij} \frac{u_{i,j} + u_{i,j-1} - u_{i-1,j} - u_{i-1,j-1}}{2h} \\
&+ (b_2)_{ij} \frac{u_{i,j} + u_{i-1,j} - u_{i,j-1} - u_{i-1,j-1}}{2h} \\
&+ |(b_1)_{ij} - (b_2)_{ij}| \frac{u_{i,j} - u_{i-1,j} - u_{i,j-1} + u_{i-1,j-1}}{2h} + c_{ij} u_{ij} = f_{ij}.
\end{aligned}
$$

The convective terms in this scheme always simplify to a two-point or three-point difference approximation of $b \cdot \nabla u$ and the scheme is of positive type. It is formally first-order consistent. ♣

Dalík [Dal95] also uses a variable stencil on an arbitrary quasi-uniform triangulation of Ω. His scheme is more complicated than that of Example 2.4. Assuming that ε is less than the minimum mesh diameter, he shows that the matrix of the scheme is inverse-monotone and that, away from layers,

$$\|u - u_h\|_\infty \le C(h^2 + \varepsilon),$$

where u_h is the computed solution and h is the mesh diameter.

Uniformly Convergent Methods

We now examine finite difference methods for (2.1) whose computed solutions u_h satisfy

$$\|u - u_h\|_{\infty,d} \le C(h^\alpha + k^\beta),$$

where α and β are positive constants that are independent of ε and of the mesh. That is, these methods are *uniformly convergent* in the discrete maximum norm $\|\cdot\|_{\infty,d}$.

We begin with a necessary condition from [Roo85] that generalizes Theorem I.2.17 to two dimensions.

Theorem 2.5. *In (2.1) assume that b_1 and b_2 are positive constants and that $c \equiv 0$. Let the mesh be square (i.e., $h_i = k_j = h$ for all i and j). Consider a nine-point difference scheme for (2.1) and assume that at each mesh point (x_i, y_j) the scheme can be written as*

$$\sum_{\nu,\mu=-1}^{1} a_{\nu\mu} u_{i+\nu,j+\mu} = h\tilde{f}_{ij},$$

where each $a_{\nu\mu}$ depends only on the ratio h/ε.

Then to achieve uniform convergence of any positive order in the discrete maximum norm $\|\cdot\|_{\infty,d}$, the coefficients of the scheme must satisfy the following three conditions:

$$e^{-b_1 h/\varepsilon} \sum_{\mu=-1}^{1} a_{-1,\mu} + \sum_{\mu=-1}^{1} a_{0,\mu} + e^{b_1 h/\varepsilon} \sum_{\mu=-1}^{1} a_{1,\mu} = 0, \qquad (2.5\text{a})$$

$$e^{-b_2 h/\varepsilon} \sum_{\nu=-1}^{1} a_{\nu,-1} + \sum_{\nu=-1}^{1} a_{\nu,0} + e^{b_2 h/\varepsilon} \sum_{\nu=-1}^{1} a_{\nu,1} = 0, \qquad (2.5\text{b})$$

$$\sum_{\nu,\mu=-1}^{1} a_{\nu\mu} e^{(\nu b_1 + \mu b_2)h/\varepsilon} = 0. \qquad (2.5\text{c})$$

Proof. The argument resembles the proof of Theorem I.2.17. □

If one weakens the hypothesis of uniform convergence in the discrete L_∞ norm to uniform convergence of order greater than $1/2$ in the discrete L_2 norm, then [ST95] the scheme must still satisfy (2.5a) and (2.5b).

A more general approach towards the assessment of numerical methods, based on a comparison of the fundamental systems of the continuous and discrete operators, is presented in [AD01].

Example 2.6. On a square mesh, consider a nine-point scheme that upwinds in each coordinate direction by an arbitrary amount. After multiplication by h, one can write the stencil of the scheme as

$$-\frac{\varepsilon}{h} \begin{bmatrix} \cdot & 1 & \cdot \\ 1 & -4 & 1 \\ \cdot & 1 & \cdot \end{bmatrix} + \frac{b_1}{2} \begin{bmatrix} \cdot & \cdot & \cdot \\ -1-p & 2p & 1-p \\ \cdot & \cdot & \cdot \end{bmatrix} + \frac{b_2}{2} \begin{bmatrix} \cdot & 1-q & \cdot \\ \cdot & 2q & \cdot \\ \cdot & -1-q & \cdot \end{bmatrix},$$

where p and q are upwinding parameters. The conditions (2.5) of Theorem 2.5 are satisfied if and only if

$$p = \coth\left(\frac{b_1 h}{2\varepsilon}\right) - \frac{2\varepsilon}{b_1 h} \quad \text{and} \quad q = \coth\left(\frac{b_2 h}{2\varepsilon}\right) - \frac{2\varepsilon}{b_2 h}. \tag{2.6}$$

This is the two-dimensional version of of the Il'in-Allen-Southwell scheme of Section I.2.1.3; it was proposed by Allen and Southwell [AS55] and independently by Il'in [Il'69].   ♣

Surprisingly, optimal error estimates for the Il'in-Allen-Southwell scheme in two dimensions are not known. Emel'janov [Eme73] proves

$$\|u - u_h\|_{\infty,d} \le Ch^{4/(2+\lambda)} \tag{2.7}$$

under the strong hypothesis that $u \in C^{4,\lambda}(\bar{\Omega})$. He uses the estimates

$$\|u\|_{k,\lambda} \le C\varepsilon^{-(k+\lambda)} \quad \text{for } k = 2, 3, 4,$$

and the bound

$$\|u - u_{as}\|_\infty \le C\varepsilon$$

which requires smoothness of the reduced solution (compare Example 1.15). An inspection of the proof shows that if we assume the existence of the decomposition (1.23), then one gets

$$\|u - u_h\|_{\infty,d} \le C(\varepsilon + h).$$

Numerical results in [HOS93] indicate that the error bound in (2.7) is fairly sharp and that with less compatibility of the data one obtains a lower rate of convergence.

2.1.2 Parabolic Boundary Layers

Consider now (2.1) with $b_1 \equiv 0$. That is, our problem becomes

$$-\varepsilon \Delta u + b_2(x, y)u_y + c(x, y)u = f \quad \text{in } \Omega := (0, 1) \times (0, 1), \tag{2.8a}$$
$$u = 0 \quad \text{on } \partial\Omega. \tag{2.8b}$$

Once again take $b_2(x, y) > \beta_2 > 0$ on $\bar{\Omega}$ and $c \ge 0$ on $\bar{\Omega}$, and assume that the data of the problem are smooth.

As in Example 1.16, the solution of this problem has typically an exponential boundary layer at $y = 1$ and parabolic boundary layers at $x = 0$ and $x = 1$. The exponential boundary layer can be treated as in the previous subsection. The parabolic layers raise interesting numerical issues. They cause numerical instabilities that are far less severe than those engendered by exponential layers, yet it is difficult in practice to approximate them accurately.

A common strategy for solving (2.8) is to use some form of upwinding to stabilize the exponential boundary layer, combined with mesh refinement near $x = 0$ and $x = 1$. Whether or not the mesh refinement yields an accurate solution in the parabolic layers, the solution elsewhere is often satisfactory.

Emel'janov [Eme70] uses a uniform mesh and applies the obvious variant of the Il'in-Allen-Southwell scheme to problem (2.8) (i.e., he sets $p = 0$ and defines q by (2.6) in Example 2.6). He proves that, on the subdomain of Ω obtained by excluding the parabolic layers, this method is first-order uniformly convergent.

Su [Su87] solves (2.8) by means of a more detailed asymptotic expansion than in Example 1.16. He uses the expansion $\tilde{u}$ of Butuzov [But75], which contains terms for the exponential layer, for each parabolic layer and for two corner layers. Assuming that $b_2 \equiv c \equiv 1$ and $f(0,0) = f(1,0) = 0$, Butuzov shows that

$$\|u - \tilde{u}\|_\infty \le C\varepsilon \quad \text{on } \bar{\Omega}.$$

All terms in the expansion can be computed explicitly, except the parabolic and corner layer terms. Su approximates each of these layer terms separately by a change of variable followed by the application of standard difference schemes. Finally, he adds all terms to obtain an approximation u_h that satisfies

$$|u(x_i, y_j) - u_{ij}| \le C(h^2 + \varepsilon)$$

at all mesh points (the position of the mesh points depends on the changes of variable, but they are clustered more densely in the parabolic layers than elsewhere). A numerical method of this type relies strongly, however, on the ability to construct an accurate asymptotic expansion of the solution u, which may be difficult in practice.

We turn now to methods for (2.8) that are *uniformly convergent* in the discrete maximum norm; that is, methods whose solutions satisfy

$$\|u - u_h\|_{\infty,d} \le C(h^\alpha + k^\beta),$$

where α and β are positive constants and u is the solution of (2.8). Before attempting to construct such a method, recall from Remark II.3.22 Shishkin's surprising obstacle result [Shi89] for parabolic reaction-diffusion problems: if a difference scheme whose stencil comprises a fixed number of points on an equidistant mesh has coefficients that are independent of the boundary data and satisfies a discrete maximum principle, then it cannot be uniformly convergent in the discrete maximum norm. The difficulty is caused by the *parabolic boundary layer* present in the solution of such problems. As these layers also occur in the solution of (2.8), we expect a similar impediment to exist. For some five-point schemes this negative result is proved in [RS96].

The following heuristic argument explains the difficulty faced by uniformly convergent schemes: necessary conditions for uniform convergence are induced by the layer terms that characterize the exact solution – see the proof of Theorem I.2.17. But the parabolic boundary layer problem (cf. Example 1.16)

$$-\frac{\partial^2 v_0}{\partial \xi^2} + \frac{\partial v_0}{\partial y} = 0, \quad v_0(0, y) = g(y), \quad v_0(\xi, 0) = 0$$

has, on taking $g(y) := y^i$ for $i = 0, 1, \ldots$, infinitely many linearly independent solutions v_0. Each of these solutions generates its own necessary condition, but it should not be possible to satisfy infinitely many conditions with the finite number of parameters available in a scheme with a fixed number of nodes (see [AD01, Shi97b] for further details).

2.2 Layer-Adapted Meshes

2.2.1 Exponential Boundary Layers

This subsection examines a model problem with exponential layers at $x = 0$ and $y = 0$, namely

$$Lu := -\varepsilon \Delta u - b \cdot \nabla u + cu = f \quad \text{on } \Omega := (0, 1) \times (0, 1), \tag{2.9a}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{2.9b}$$

where $b = (b_1(x, y), b_2(x, y)) > (\beta_1, \beta_2) > (0, 0)$ on $\bar{\Omega}$ and b, c, f are smooth. If the differential equation were $-\varepsilon \Delta u + b \cdot \nabla u + cu = f$, whose solution has layers at $x = 1$ and $y = 1$, then the change of variables $x \mapsto 1-x$ and $y \mapsto 1-y$ converts this problem to (2.9).

First consider the simple upwind scheme

$$L^{up} u_h = f_h,$$

which, written out in full, is

$$-\frac{2\varepsilon}{h_i + h_{i+1}} \left(\frac{u_{i+1,j} - u_{ij}}{h_{i+1}} - \frac{u_{ij} - u_{i-1,j}}{h_i} \right)$$

$$-\frac{2\varepsilon}{k_j + k_{j+1}} \left(\frac{u_{i,j+1} - u_{ij}}{k_{j+1}} - \frac{u_{ij} - u_{i,j-1}}{k_j} \right)$$

$$-(b_1)_{ij} \frac{u_{i+1,j} - u_{i,j}}{h_i} - (b_2)_{i,j} \frac{u_{i,j+1} - u_{i,j}}{k_j} + c_{ij} = f_{ij} \tag{2.10a}$$

for $i = 1, \ldots, M-1$ and $j = 1, \ldots, N-1$, with

$$u_{ij} = 0 \quad \text{when } (x_i, y_j) \in \partial\Omega. \tag{2.10b}$$

We shall study this scheme on a tensor-product mesh $\omega_x \times \omega_y$, where ω_x and ω_y are one-dimensional Shishkin-type meshes (see Section I.2.4) having the same number of mesh points. Thus ω_x is obtained from the continuous mesh-generating function λ, where

$$\lambda(\xi) = \frac{\sigma\varepsilon}{\beta_1}\tilde{\lambda}(\xi) \quad \text{for } \xi \in [0, 1/2].$$

The function $\tilde{\lambda}$ is monotone with $\tilde{\lambda}(0) = 0$ and $\tilde{\lambda}(1/2) = \ln N$; on $[1/2, 1]$ the function λ is linear with $\lambda(1) = 1$. Recall that the mesh-characterizing function ψ is defined by $\psi = \exp(-\lambda)$. Figure 2.1 shows the typical structure of a tensor-product Shishkin-mesh for a problem with two exponential layers at $x = 1$ and $y = 1$; in this diagram (see Section I.2.4) $\lambda_x = C_x \varepsilon \ln N$ and $\lambda_y = C_y \varepsilon \ln N$, where N mesh intervals are used in each coordinate direction.
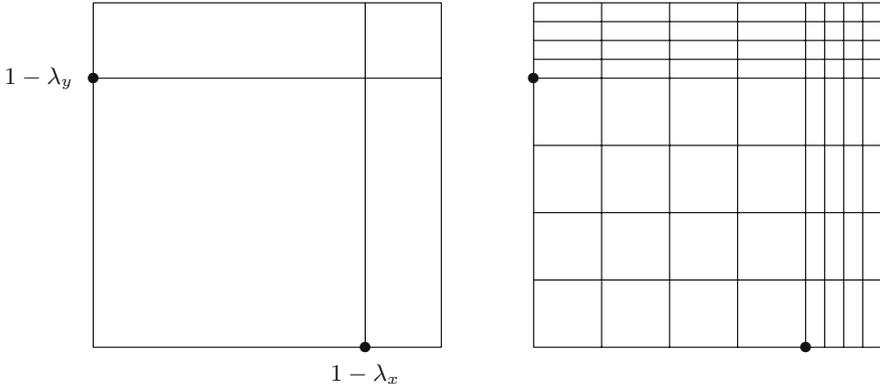
Fig. 2.1. Shishkin mesh for convection-diffusion with two outflow exponential layers

Theorem 2.7. *Assume that the solution of* (2.9) *can be decomposed similarly to Theorem 1.26 with $\alpha = 1$ and $n = 3$. Let the mesh-generating function be piecewise differentiable and satisfy*

$$\max \tilde{\lambda}'(\xi) \le CN, \qquad \int_0^{1/2} \tilde{\lambda}'(\xi)^2 d\xi \le CN.$$

Then for $\sigma \ge 2$, the error of the simple upwind scheme satisfies

$$|u(x_i, y_j) - u_{ij}| \le \begin{cases} C N^{-1} & \text{for } i, j = N/2, \ldots, N, \\ C N^{-1} \max |\psi'| & \text{otherwise.} \end{cases}$$

For a piecewise-equidistant Shishkin mesh the mesh-characterizing function ψ introduces a factor $\ln N$ into the error estimate. On the other hand, the Bakhvalov-Shishkin mesh, for which $\psi(\xi) = 1-(1-1/N)2\xi$, yields the optimal error estimate CN^{-1} because $|\psi'|$ is uniformly bounded.

Proof. Recalling the decomposition of Theorem 1.26, split the numerical solution in a similar manner: define the mesh function S^N as the solution of

$$\left[L^{up}S^N\right]_{ij} = [LS]_{ij} \text{ for all } i \text{ and } j, \quad S^N = S \text{ on } \partial\Omega,$$

and define E_1^N, E_2^N and E_{12}^N analogously. For the smooth component S, standard classical arguments give

$$|S(x_i, y_j) - S_{ij}^N| \leq C N^{-1} \text{ for all } i \text{ and } j.$$

For the layer term at $x = 0$ we have, like the proof of Lemma I.2.94,

$$0 \leq E_{1,ij}^N \leq W_{1,i} := C \prod_{\nu=1}^{i} \left(1 + \frac{\beta_1 h_\nu}{2\varepsilon}\right)^{-1} \quad \text{for } i, j = 0, \dots, N.$$

The smallness of E_1 on the coarse mesh leads to

$$\left| E_1(x_i, y_j) - E_{1,ij}^N \right| \leq C N^{-1} \quad \text{for } i = N/2, \dots, N, \; j = 0, \dots, N.$$

A Taylor expansion gives

$$\left| L^{up}(E_1 - E_1^N) \right| \leq C(N^{-1} + \varepsilon^{-1} W_{1,i} \max |\psi'|).$$

Appealing to a discrete comparison principle and using the barrier function $C(N^{-1} + W_{1,i} N^{-1} \max |\psi'|)$ yields

$$\left| E_1(x_i, y_j) - E_{1,ij}^N \right| \leq C N^{-1} \max |\psi'| \quad \text{for } i = 0, \dots, N/2 - 1, \; j = 0, \dots, N.$$

Similar arguments are used for the terms E_2 and E_{12} corresponding to the boundary layer at $y = 0$ and the corner layer.   □

   When the solution is less regular, one can nevertheless prove some positive rate of convergence; see [Shi00].

   In [LS99] a modified hybrid scheme on a tensor-product Shishkin mesh is considered. It is based on simple upwinding, but employs central differencing whenever the mesh allows one to do this without losing inverse-monotonicity. For this scheme the above proof avoids the factor $\ln N$ and gives the optimal error bound

$$\|u - u_h\|_{\infty, d} \leq C N^{-1}.$$

   Liseikin [Lis83] uses a tensor product of one-dimensional Bakhvalov-type meshes (see Section I.2.4). He assumes the validity of the estimates

$$\left| \frac{\partial^k u}{\partial x^k}(x, y) \right| \leq C \left[1 + \varepsilon^{-k} e^{-\beta_1 (1-x)/\varepsilon} \right]$$

and

$$\left| \frac{\partial^k u}{\partial y^k}(x, y) \right| \leq C \left[1 + \varepsilon^{-k} e^{-\beta_2 (1-y)/\varepsilon} \right]$$

on Ω for $0 \leq k \leq 3$. Such an assumption implies, as was seen in Chapter 1, that the data of the problem are smooth and satisfy strong compatibility conditions at the corners of $\bar{\Omega}$. The logarithmically graded mesh then controls the local truncation error of the simple upwind scheme. The computed solution satisfies

$$\|u - u_h\|_{\infty,d} \le CN^{-1}.$$

The proof of Theorem 2.7 used the discrete comparison principle and carefully chosen barrier functions. Alternatively, as we saw in Section I.2.4.2, one can use an improved stability result for the upwind scheme. Thus [And01] one has the following discrete stability analogue of the continuous stability bound of Theorem 1.22. Set $\bar{h}_i = (h_i + h_{i+1})/2$ and $\bar{k}_i = (k_i + k_{i+1})/2$ for each i. Define the discrete Green's function G_d by

$$L^{up}G_d(x_i, y_j; \xi_m, \eta_n) = \delta^d(x_i, \xi_m)\,\delta^d(y_j, \eta_n), \qquad G_d = 0 \text{ on } \partial\Omega,$$

with

$$\delta^d(x_i, \xi_m) = \begin{cases} (\bar{h}_i)^{-1} & \text{for } i = m, \\ 0 & \text{otherwise;} \end{cases}$$

the mesh function $\delta^d(y_j, \eta_n)$ is defined analogously.

Lemma 2.8. *The discrete Green's function G_d is nonnegative. One has the estimates*

$$\max_{x_i, y_j, \eta_n} \|G_d(x_i, y_j; \cdot, \eta_n)\|_{L_1,d} \le \frac{1}{\beta_2}$$

and

$$\max_{x_i, y_j, \xi_m} \|G_d(x_i, y_j; \xi_m, \cdot)\|_{L_1,d} \le \frac{1}{\beta_1},$$

where $\|\cdot\|_{L_1,d}$ is the one-dimensional discrete L_1 norm.

Proof. It suffices to prove the statement for $c = 0$ because its Green's function $\tilde{G}_d$ satisfies $0 \le G_d \le \tilde{G}_d$ owing to the inverse-monotonicity of the discrete problem. Thus assume that $c = 0$ in (2.9). Fix (x_i, y_j). Define the function of one variable $G^\Sigma(x_i, y_j; \cdot)$ by

$$G^\Sigma(x_i, y_j; \xi_m) := \sum_n G_d(x_i, y_j; \xi_m, \eta_n)\bar{k}_n.$$

As $G_d \ge 0$, this sum is simply the discrete L_1 norm of $G_d(x_i, y_j; \xi_m, \cdot)$. Analogously to the continuous Green's function, the function G_d satisfies the adjoint problem associated with (ξ_m, η_n) – for the simple upwind scheme, the term $-b_i D^+ u_i$ has adjoint $D^-(b_i u_i)$. Multiplying the adjoint equation by $\bar{k}_n$ then summing over n, we obtain a difference equation for G^Σ in its third argument:

$$-\varepsilon\delta^2(G^\Sigma)_m + D^-(b_1^* G^\Sigma)_m = \delta^d(x_i, \xi_m) - F(x_i, y_j, \xi_m). \qquad (2.11)$$

Here $b_1^* \ge \beta_1$ is derived from b_1 via a mean value theorem while

$$F(x_i, y_j, \xi_m) := b_{2,m,N-1}G_d(x_i, y_j; \xi_m, \eta_{N-1}) + \frac{\varepsilon}{k_{N-1}} G_d(x_i, y_j; \xi_m, \eta_{N-1})$$

$$+ \frac{\varepsilon}{k_1} G_d(x_i, y_j; \xi_m, \eta_1).$$

Denoting by G_d^* the Green's function associated with the one-dimensional operator on the left-hand side of (2.11), one then has

$$0 \le G_d^* \le \frac{1}{\beta_1}.$$

Now the solution representation

$$G^\Sigma(x_i, y_j, \xi_m) = G_d^*(x_i, y_j, \xi_m, \xi_m) - \sum_n G_d^*(x_i, y_j, \xi_m, z_n) F(x_i, y_j, z_n) \bar{h}_n$$

gives us immediately the second estimate of the lemma because $F \ge 0$. The other inequality is proved similarly.   □

Consider now the discrete boundary value problem

$$L^{up} u_h = f_h \text{ in } \Omega, \qquad u_h = 0 \text{ on } \partial\Omega.$$

The solution representation

$$u_h(x_i, y_j) = \sum_{m,n} \bar{h}_m \bar{k}_n G_d(x_i, y_j, \xi_m, \eta_n) f_h(\xi_m, \eta_n)$$

yields

Theorem 2.9. *Assume that $b_1 > \beta_1 > 0$. Then the simple upwind operator enjoys the anisotropic stability estimate*

$$\|v_h\|_{\infty,d} \le C \|L^{up} v_h\|_{1 \otimes \infty, d}.$$

The notation here is analogous to Theorem 1.22: first apply the maximum norm in the y-direction and then the discrete L_1 norm with respect to x.

When $b_1 > 0$ and $b_2 > 0$ on $\bar{\Omega}$, Theorem 2.9 gives an alternative proof of

$$\|u - u_h\|_{\infty,d} \le C \begin{cases} N^{-1} & \text{for a Bakhvalov mesh,} \\ N^{-1} \ln N & \text{for a Shishkin mesh.} \end{cases}$$

Remark 2.10. (Reaction-diffusion problem) For the reaction-diffusion problem discussed in Remark 1.27 with $\Omega = (0,1)^2$ one expects uniform convergence on a Shishkin mesh for the standard finite difference method obtained by setting $b_1 \equiv b_2 \equiv 0$ in (2.10). With sufficient compatibility to ensure that $u \in C^{4,\alpha}(\bar{\Omega})$ and that one has a suitable decomposition of u, it is straightforward to prove

$$\|u - u_h\|_{\infty,d} \le C(N^{-1} \ln N)^2. \tag{2.12}$$

This was shown in [CGO05] using a barrier function technique; alternatively, one could use an improved stability estimate based on the Green's function as in the one-dimensional problem considered by Savin [Sav95].

Remarkably, Andreev [And06] was able to avoid the use of compatibility conditions (see the discussion on solution decomposition in Remark 1.27) when proving

$$\|u - u_h\| \leq C\, N^{-2}(\ln N)^4$$

for this scheme. Subsequently Andreev [And] and Andreev and Kopteva [AK08] extended these results to problems with stronger corner singularities. In the first of these papers, Dirichlet and Neumann boundary conditions meet at a corner of the unit square, while in the second an L-shaped domain with Dirichlet boundary conditions is treated. In both cases the solution lies only in $C^{0,\alpha}(\bar{\Omega})$ with $0 < \alpha < 1$. The analysis combines layer-adapted meshes with geometrically graded meshes near the corner singularity; for related work, see [Mel02].

Kopteva [Kop07a, Kop07b] studies semilinear problems of the type

$$Lu := -\varepsilon \triangle u + b(x,u) = 0 \quad \text{in } \Omega,$$
$$u = g \quad \text{on } \Gamma,$$

in a domain with a smooth boundary while assuming some standard stability property of the reduced solution. The discretization combines features of finite differences and finite elements. In a strip "parallel" to the boundary, whose thickness corresponds to the construction of a Bakhvalov or a Shishkin mesh, a finite difference method is used, and in the interior of Ω the difference scheme is generated via linear finite elements on a Delaunay triangulation. Optimal error bounds in the maximum norm are derived.

Systems of reaction-diffusion problems in two-dimensional domains are solved in [KLS, KMS08, Shi07b] using standard schemes on Bakhvalov and Shishkin meshes and error bounds like (2.12) are proved. ♣

Are there second-order schemes for convection-diffusion problems? In [Kop03] Kopteva derives an error expansion for the simple upwind scheme on a piecewise equidistant mesh. This expansion is used to show that Richardson extrapolation generates a robust almost (i.e., up to a logarithmic factor) second-order method.

Comparing numerical results for simple upwinding, a hybrid scheme, central differencing and defect correction on a Shishkin mesh, it is concluded in [LS01b] that defect correction is the most efficient of these because it combines the accuracy of central differencing with the good stability properties of upwinding. But up to now, no complete analysis of defect correction for two-dimensional convection-diffusion problems has been given.

We do not know of any theoretical result for central differencing on layer-adapted meshes for (2.9), but there are some results for related schemes generated by finite element methods; see Chapter 3.

2.2.2 Parabolic Layers

Compared with exponential layers, satisfactory convergence results for difference schemes for parabolic boundary layers are thin on the ground. Let us consider the model problem

$$Lu := -\varepsilon \Delta u + u_x + cu = f \quad \text{in } \Omega := (0,1) \times (0,1), \qquad (2.13a)$$
$$u = 0 \quad \text{on } \partial\Omega. \qquad (2.13b)$$

The problem has an exponential layer of width $\mathcal{O}(\varepsilon |\ln \varepsilon|)$ at the outflow boundary $x = 1$ and parabolic layers of width $\mathcal{O}(\sqrt{\varepsilon} |\ln \varepsilon|)$ at the characteristic boundaries $y = 0$ and $y = 1$.

We know already that on standard meshes it is impossible to construct difference schemes that are uniformly convergent pointwise for problems with parabolic boundary layers. Thus for (2.13) one could combine fitted schemes in the x-direction with a layer-adapted scheme in the y-direction, or use layer-adapted meshes in both directions to try to achieve uniform convergence. In a finite difference framework, we know of only one paper that avails itself of the former strategy: Shishkin [Shi86] uses the one-dimensional Il'in-Allen-Southwell scheme to approximate $-\varepsilon u_{xx} - u_x$ and central differencing to approximate $-\varepsilon u_{yy}$. He proves the following result, where the mesh uses N points in each coordinate direction – equidistant in x but layer-adapted in y.

Theorem 2.11. (*Il'in-Allen-Southwell scheme and a Bahkvalov mesh*) *Assume that $c, f \in C^3(\bar{\Omega})$, that $u \in C^4(\bar{\Omega})$, and $f(1,0) = f(1,1) = 0$. Then*

$$\|u - u_h\|_{\infty,d} \leq CN^{-1/4}.$$

If more smoothness and compatibility of the data are assumed, then the conclusion of the Theorem becomes $\|u - u_h\|_{\infty,d} \leq CN^{-1/2}$.

As tensor products of layer-adapted meshes were quite successful for exponential layers, we now introduce a mesh of this type for (2.13), using the mesh transition parameters (see Section I.2.4)

$$\lambda_x = \min\left\{1/2, \ \sigma_x \varepsilon \ln N\right\}, \qquad \lambda_y = \min\left\{1/4, \ \sigma_y \sqrt{\varepsilon} \ln N\right\}$$

where N mesh intervals are used in each coordinate direction and the mesh is fine at $x = 0$ and at $y = 0$, $y = 1$. Note that the ε of the exponential layer transition point becomes $\sqrt{\varepsilon}$ for the parabolic layer; this is due to the different asymptotic structure of these layers. Figure 2.2 shows the typical structure of such a mesh for (2.13); in each coordinate direction, half the mesh intervals are in the coarse mesh and half in the fine mesh. For simplicity, only the standard Shishkin mesh is discussed here but Shishkin-type and Bakhvalov-type meshes are also possible.

Numerical results in several papers [CMOS01, FHS96a, FHS96b, FHS96c, HMOS95] and the monograph [FHM+00] demonstrate numerically the almost
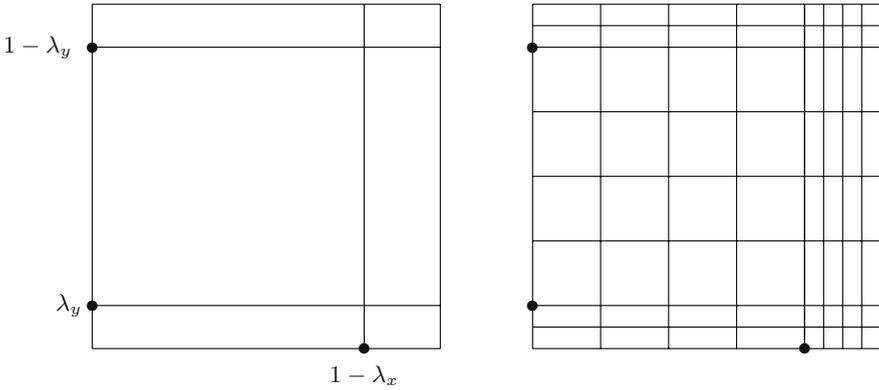
Fig. 2.2. Shishkin mesh for one exponential and two parabolic layers

first-order uniform convergence of the simple upwind scheme, but a rigorous proof of this convergence under minimal regularity assumptions is not easy. Using the decomposition (1.29) and the estimates (1.30), Shishkin [Shi90b] proved that

$$\|u - u_h\|_{\infty,d} \leq C(N^{-1} \ln N)^p$$

with $p = 1/18$ or $p = 1/14$ depending on the precise assumptions on the problem data. If we assume the validity of the decomposition

$$u = S + E_1 + E_2 + E_{12}$$

and of bounds like (1.32) for the third-order derivatives needed in the analysis of the simple upwind scheme, then it should be possible to prove that

$$\|u - u_h\|_{\infty,d} \leq CN^{-1} \ln N.$$

In fact [OS07a] derives the error estimate $\|u - u_h\|_{\infty,d} \leq CN^{-1}(\ln N)^2$.

Remark 2.12. (The A-mesh) In [Wes96] Wesseling implicitly assumes the existence of a decomposition of the solution (by ignoring higher-order terms in an asymptotic expansion) for a problem with parabolic boundary layers and a weak exponential layer. He proves first-order uniform convergence for an upwind scheme with a refined piecewise equidistant mesh near the characteristic boundaries but his choice of transition point is

$$\lambda_y = \min\{1/4, \sigma_y \sqrt{\varepsilon}\,|\ln \varepsilon|\},$$

i.e., the factor $\ln N$ of the Shishkin mesh is replaced by $|\ln \varepsilon|$. This mesh is sometimes called the A-mesh. Numerical experiments in [HMOS97] demonstrate however that this choice of transition point for a piecewise equidistant mesh is not as good as Shishkin's if one wants also to approximate scaled derivatives of the solution. ♣

Remark 2.13. The authors of [CGLS02] study a problem with Robin boundary conditions on the characteristic boundary, so the parabolic boundary layer is weak. Assuming the existence of a decomposition of the type (1.31) with estimates similar to (1.32) for the fourth-order derivatives, it is shown that

$$\|u - u_h\|_{\infty,d} \leq C \left[(N^{-1} \ln N)^2 + \sqrt{\varepsilon} N^{-1} \ln N \right]$$

for a scheme which in the x-direction is related to the midpoint upwind scheme. ♣

Remark 2.14. (Interior parabolic layers) In [HS94], Hemker and Shishkin study a singularly perturbed parabolic equation with a discontinuous initial condition that generates an *interior parabolic layer* and construct a uniformly convergent (fitted) scheme on an equidistant mesh. Unlike the situation with parabolic boundary layers, the equation determining the layer correction now has only one solution (the classical error function).

One would expect a similar result for an elliptic problem of type (2.13) with constant coefficients, if a discontinuous boundary condition generates an interior parabolic layer at the subcharacteristic through the point of discontinuity. For nonconstant coefficients (curved subcharacteristics) the situation is more complicated and is unclear. ♣

Remark 2.15. (Hemker's problem) In [Hem97] Hemker proposes the following benchmark problem: solve

$$-\varepsilon \triangle u + u_x = 0$$

in the plane region exterior to the unit circle with the boundary conditions

$$u(x, y) = 1 \text{ for } x^2 + y^2 = 1, \qquad u(x, y) \to 0 \text{ as } x^2 + y^2 \to \infty.$$

This is a complicated problem: the solution has an exponential layer and two interior parabolic layers – in particular the asymptotic situation is quite complicated at the points $(0, \pm 1)$ where the parabolic layers are "born" from the exponential layer. (The unboundedness of the domain is unimportant.)

Numerical results for this problem can be found in [HHH00] (here the so-called over-set grid technique is used) and [NH00], where an adaptive sparse-grid technique is developed. ♣

3

Finite Element Methods

Throughout most of Chapter 3, we consider the problem

$$Lu := -\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in} \quad \Omega, \tag{3.1a}$$

$$u = 0 \quad \text{on} \quad \Gamma := \partial\Omega, \tag{3.1b}$$

where Ω is a bounded polygonal domain in $\mathbb{R}^2$. (In Section 3.3, $\mathbb{R}^2$ is replaced by $\mathbb{R}^d$ with $d \geq 2$.) Nevertheless many of the ideas below can be transferred to problems posed in more than two space dimensions.

The weak formulation of problem (3.1) is:

Find $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$ one has $a(u, v) = (f, v)$,

where the bilinear form $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is defined by

$$a(v, w) := \varepsilon(\nabla v, \nabla w) + (b \cdot \nabla v + cv, w)$$

with $(\cdot, \cdot)$ the $L_2(\Omega)$ inner product. As in the one-dimensional case (Section I.2.2.1), choose a finite-dimensional space $V_h \subset H_0^1(\Omega)$ that comprises continuous piecewise polynomial splines. Then the standard Galerkin finite element method is

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has $a(u_h, v_h) = (f, v_h)$. (3.2)

We know already from Parts I and II that standard Galerkin finite element methods usually yield inaccurate approximate solutions to convection-diffusion problems. This disappointing behaviour occurs because such methods lose stability and cannot adequately approximate solutions inside layers. Thus modified finite element methods whose performance is more satisfactory will now be discussed.

To stabilize (3.2), the simplest technique is to add artificial viscosity (AVIS) of magnitude $\mathcal{O}(h)$ to the diffusion coefficient ε. In the one-dimensional case, this modification is equivalent to an upwind discretization of the convective term – recall (I.2.14). But in the multi-dimensional case, the term

"upwinding" encompasses various stabilization techniques, not all of which are equivalent to the addition of AVIS. Upwinding may produce excessive numerical viscosity (NVIS) in the computed solution; Remark 2.2 discussed the distinction between AVIS and NVIS.

The finite element spaces considered in Chapter 3 are based on an associated family of triangulations $\mathcal{T}_h$ of Ω that has to satisfy some conditions. Let T denote any triangle (or convex quadrilateral) of $\mathcal{T}_h$. Then the diameter of T is denoted by h_T and the diameter of its largest inscribed circle by ρ_T. Set $h = \max_{T \in \mathcal{T}_h} h_T$. A family of triangulations $\mathcal{T}_h$ is called *shape-regular* if there is a positive constant C, independent of h and ε, such that $h_T/\rho_T \leq C$ for all $T \in \mathcal{T}_h$. *It is always assumed here that the family of triangulations is shape-regular* unless we say otherwise. A shape-regular family of triangulations $\mathcal{T}_h$ is said to be *quasi-uniform* if there is a positive constant C', independent of h and ε, such that $C'h \leq h_T$ for all $T \in \mathcal{T}_h$. The quasi-uniform property is stronger than the shape-regular property; in particular, arbitrary locally-refined meshes are not quasi-uniform. Regrettably, the precise meaning of terms such as these varies in the research literature, so when reading any work one must be careful to check the definitions of the terminology used.

Chapter 3 is organized as follows. Section 3.1 studies schemes that preserve the inverse-monotonicity of the continuous problem. This is a desirable attribute in many applications, but such schemes are restricted to first-order accuracy. Then, in Section 3.2 we switch to higher-order methods, viz., the streamline diffusion, Galerkin least squares and residual-free bubble finite element methods. Here the addition of weighted residuals to the standard Galerkin finite element method yields improved stability properties, despite the absence of a discrete maximum principle. Stabilization methods based on adding symmetric terms are studied in Section 3.3. The underlying concept of stabilization by discontinuous Galerkin finite element methods will be investigated in Section 3.4. In general, the methods presented in Sections 3.1–3.4 are not ε-uniformly convergent – methods of that type are considered in Section 3.5. Finally, adaptive finite element methods are examined in Section 3.6.

3.1 Inverse-Monotonicity-Preserving Methods Based on Finite Volume Ideas

Consider first the case where the coefficients b and c of the operator L are continuous with $c(x) \geq c_0 \geq 0$. These hypotheses ensure L has the following two properties, which are closely related to the classical comparison principle of Theorem 1.4.

Let $w \in C(\bar{\Omega}) \cap C^2(\Omega)$. The operator L is said to be *inverse-monotone* if the inequalities

$$\left. \begin{array}{l} Lw(x) \geq 0 \quad \text{for all} \quad x \in \Omega \\ w(x) \geq 0 \quad \text{for all} \quad x \in \Gamma \end{array} \right\} \quad \text{imply} \quad w(x) \geq 0 \quad \text{for all } x \in \bar{\Omega} \, . \tag{3.3}$$

We say that L satisfies a *maximum principle* if

$Lu(x) = 0$ for all $x \in \Omega$ implies that
$$\min_{x \in \Gamma}\{u(x), 0\} \leq u(x) \leq \max_{x \in \Gamma}\{u(x), 0\} \quad \text{for all } x \in \bar{\Omega}. \qquad (3.4)$$

Note that (3.4) is a direct consequence of (3.3). It has the following physical interpretation: suppose that $c = 0$ and u denotes the density of a substance, where no source of the substance is present – then the maximum principle states that the greatest density occurs on the boundary Γ and the density never takes negative values.

Finite element methods for which (3.4) fails may produce solutions that take negative values. Consequently inverse-monotonicity-preserving methods are particularly desirable in some applications.

When searching for a weak solution of the boundary value problem (3.1), i.e., for $u \in H_0^1(\Omega)$, then the operator L must be interpreted in the weak sense that $L : H_0^1(\Omega) \to H^{-1}(\Omega)$. In this case, sufficient conditions for the inverse-monotonicity of L follow from a weak maximum principle, see [GT83, Tro87].

Inverse-monotonicity of the discretized problem is often proved using the M-criterion of Theorem I.2.7, which is now presented in a slightly different form.

Theorem 3.1. *Let $A = (a_{ij})$ be an $N \times N$ matrix with $a_{ij} \leq 0$ for $i \neq j$. Then the following conditions are equivalent:*
(i) A is an M-matrix;
(ii) A is inverse-monotone, i.e., $Az \geq 0$ implies $z \geq 0$;
(iii) there is a vector $e \geq 0$ with $Ae \geq 0$, and for each $i \in \{1, \ldots, N\}$ with $(Ae)_i = 0$, there is a chain $i_0 = i, i_1, \ldots, i_k \in \{1, \ldots, N\}$ such that $a_{i_{\nu-1} i_\nu} < 0$ for $\nu = 1, \ldots, k$ and $(Ae)_{i_k} > 0$.

We begin with the case of pure diffusion (i.e., b and c identically zero), for which the operator $L := -\Delta$ is inverse-monotone. Consider a finite element space V_h of piecewise linear functions on triangles where V_h satisfies Dirichlet homogeneous boundary conditions. Then at the inner meshpoints $p_k \in \Omega$, where $k = 1, \ldots, N$, the matrix $A = (a_{ij})$ has the form

$$a_{ij} = (\nabla \varphi_j, \nabla \varphi_i) \quad \text{for } i, j = 1, \ldots, N,$$

where φ_i denotes the usual basis function satisfying $\varphi_i(p_j) = \delta_{ij}$. Now $a_{ij} \neq 0$ only when p_i and p_j belong to the same triangle T. Since $\nabla \varphi_i$ is constant on T and has direction opposite to the outer normal on the side not containing p_i, it follows that the sign of a_{ij}, for $i \neq j$, depends only on the angle between the two sides that do not simultaneously contain p_i and p_j. More precisely, for $i \neq j$, one has $a_{ij} < 0$ if this angle is less than $\pi/2$, while $a_{ij} = 0$ if the angle equals $\pi/2$. Moreover, we conclude that

$$a_{ij} = 0 \text{ for } i \neq j \quad \text{implies that} \quad a_{ik} < 0 \text{ and } a_{kj} < 0, \text{ for } i \neq k, k \neq j.$$

Triangulations of the polygonal domain Ω are said to be *weakly acute* if all angles of their triangles are less than or equal to $\pi/2$. We now examine properties of the matrix $A = (a_{ij})$, with $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$, for conforming finite element methods that use piecewise linears on weakly acute meshes.

Let $e = (1, \ldots, 1)$ be an N-dimensional vector and p_i an inner vertex of a triangle that has no boundary node. Then

$$(Ae)_i = \sum_{j=1}^{N} a_{ij} = \left(\nabla \sum_{j=1}^{N} \varphi_j, \nabla\varphi_i \right) = 0,$$

because $\sum_{j=1}^{N} \varphi_j = 1$ on the support of φ_i. If instead p_i is a vertex of a boundary triangle, let p_j, for $j = N + 1, \ldots, N + M$, be the vertices at the boundary that belong to the support of φ_i. Then

$$(Ae)_i = \left(\nabla \sum_{j=1}^{N} \varphi_j, \nabla\varphi_i \right) = -\left(\nabla \sum_{j=N+1}^{N+M} \varphi_j, \nabla\varphi_i \right) = -\sum_{j=N+1}^{N+M} a_{ij} \geq 0.$$

The final case is where T is a triangle with one inner vertex p_i and two boundary vertices. As T has at most one angle equal to $\pi/2$, we have $(Ae)_i > 0$. Thus the hypotheses of Theorem 3.1(iii) have been verified and we infer that A is an M-matrix. This proves

Lemma 3.2. *On meshes of weakly acute type, the discretization of $-\Delta$ by piecewise linear functions preserves the inverse-monotonicity of the continuous problem.*

The above argument ensured the desired sign pattern in A by considering the contribution of each triangle $T \in \mathcal{T}_h$. But for piecewise linear functions, all elements that share the edge $p_i p_j$ contribute to $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$. This observation permits the replacement of the hypothesis of weakly acute meshes in Lemma 3.2 by the following less stringent condition from [XZ99] for simplicial meshes in $\mathbb{R}^d$ with $d \geq 2$:

for each $T \in \mathcal{T}_h$ and all $E \in \mathcal{E}_h$, one has the inequality

$$\frac{1}{d(d-1)} \sum_{E \subset T} |\kappa_{E,T}| \cot \Theta_{E,T} \geq 0. \tag{3.5}$$

Here $\mathcal{E}_h$ is the set of all edges in the mesh; when the edge E joins the vertices p_i and p_j, then $F_{i,T}$ and $F_{j,T}$ are the faces of T opposite the vertices p_i and p_j respectively, $|\kappa_{E,T}|$ is the $(d-2)$-dimensional measure of the simplex $F_{i,T} \cap F_{j,T}$ opposite the edge E of T, and $\Theta_{E,T}$ is the angle between the faces $F_{i,T}$ and $F_{j,T}$. In the two-dimensional case $d = 2$, this assumption means that the sum of the two angles facing any edge in the mesh is less than π, which implies that we have a *Delaunay triangulation* – these are characterized by the condition that the circumcircle of each triangle $T \in \mathcal{T}_h$ contains no vertex

of any triangle other than T. It has been studied intensively in computational geometry and fast algorithms for constructing a Delaunay triangulation using a given set of vertices are known [Pac93].

The condition (3.5) on the mesh in $\mathbb{R}^d$ is necessary and sufficient [XZ99, Lemma 2.1] for the M-matrix property in the discretization of $-\Delta$ by linear finite elements. For the three-dimensional case see also [KKN00].

It is important to note that in the case of a discretization of $-\Delta$ by piecewise quadratic elements, inverse-monotonicity can be proved only in special geometric situations [HM81]. Consequently it is natural in Section 3.1 to restrict our analysis to piecewise linear elements and to shun the construction of inverse-monotonicity preserving methods for higher-order elements.

Consider now the more general case where $c(x) \geq 0$. If $c(x) = 1$, then when φ_i and φ_j have overlapping supports, clearly

$$(c\,\varphi_j, \varphi_i) > 0 \quad \text{for } i \neq j.$$

This produces positive off-diagonal matrix entries in the corresponding matrix, which therefore is not an M-matrix. Nevertheless, there is a simple remedy: if one calculates each term $(c\,\varphi_j, \varphi_i)$ approximately by means of a quadrature rule such as

$$\int_T \Phi(x)\,dx \sim \frac{|T|}{3}\big[\Phi(p_i) + \Phi(p_j) + \Phi(p_k)\big],$$

where $|T|$ denotes the measure of T; this will contribute positive entries only on the main diagonal of A, which will now be an M-matrix. This quadrature rule also maintains the order of convergence of the method.

The right-hand side f can be handled in the same way. In particular, the approximation $(f, \varphi_i) \sim (f(p_i)/3) \sum_{T \cap p_i \neq \emptyset} |T|$ is non-negative if $f \geq 0$. In some cases this quadrature rule resembles is like the *lumping* technique that is often applied to the mass matrix in parabolic problems.

The construction of inverse-monotonicity-preserving discretizations of the convective term in (3.1) is much more complicated; various techniques have been developed in the literature. We shall study certain modifications of the standard Galerkin finite element method whose corresponding discrete operators are inverse-monotone.

Consider a triangulation $\mathcal{T}_h$ of Ω of weakly acute type. (We should more properly deal with a family of such triangulations, indexed by a parameter h, but for simplicity here and below we work with a single such triangulation.) Triangle vertices that do not lie on the boundary Γ are denoted by p_i, for $i = 1, \ldots, N$. Approximate the solution space V by the standard piecewise linear space

$$V_h := \{ v \in C(\bar{\Omega}) : \quad v_h\big|_T \in P_1(T) \text{ for all } T \in \mathcal{T}_h, \quad v_h\big|_\Gamma = 0 \}.$$

Let $\varphi_i \in V_h$ be the canonical basis function that satisfies $\varphi_i(p_j) = \delta_{ij}$, so $V_h = \text{span}\{\varphi_1, \ldots, \varphi_N\}$.

The discrete problem now reads:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$\varepsilon(\nabla u_h, \nabla v_h) + b_h(u_h, v_h) + c_h(u_h, v_h) = f_h(v_h), \qquad (3.6)$$

where $b_h(\cdot, \cdot)$, $c_h(\cdot, \cdot)$ and $f_h(\cdot)$ are approximations of the continuous bilinear forms $(b \cdot \nabla \cdot, \cdot)$, $(c \cdot, \cdot)$ and the linear form $(f, \cdot)$. As was mentioned earlier, c_h and f_h are computed by applying the quadrature rule

$$\int_T \Phi(x)\, dx \sim \frac{|T|}{3} \left[\Phi(p_i) + \Phi(p_j) + \Phi(p_k)\right],$$

where p_i, p_j and p_k are the vertices of T. More precisely,

$$(c\, u_h, \varphi_i) = \sum_T \int_T c(x) u_h(x) \varphi_i(x)\, dx \sim \sum_{T \cap p_i \neq \emptyset} \frac{|T|}{3} c(p_i)\, u_h(p_i),$$

thus set

$$c_h(u_h, \varphi_i) := \frac{1}{3} c(p_i)\, u_h(p_i) \sum_{T \cap p_i \neq \emptyset} |T|, \qquad (3.7a)$$

$$f_h(\varphi_i) := \frac{1}{3} f(p_i) \sum_{T \cap p_i \neq \emptyset} |T|. \qquad (3.7b)$$

Now we come to the discretization of the convective term. Define the directional derivative $\partial(\cdot)/\partial b$ by

$$|b|\, \frac{\partial u_h}{\partial b} = b \cdot \nabla u_h,$$

then apply the same integration rule; this gives

$$(b \cdot \nabla u_h, \varphi_i) \sim \frac{1}{3} |b(p_i)| \frac{\partial u_h}{\partial b}(p_i) \sum_{T \cap p_i \neq \emptyset} |T|,$$

so now one need only approximate the directional derivative at p_i.

The first idea for this is due to Tabata [Tab77], who proposes to set

$$|b(p_i)| \frac{\partial u_h}{\partial b}(p_i) \sim b(p_i) \cdot \nabla u_h\big|_{T_i},$$

where T_i denotes an *upwind triangle*. For the vertex p_i, a triangle T_i is called upwind with respect to b (see Figure 3.1) if
(i) p_i is a vertex of T_i,
(ii) the vector $-b(p_i)$ points from p_i into T_i.
Observe here that if $b(p_i) = 0$ it is not necessary to define an upwind triangle and in all other cases at least one upwind triangle T_i exists. Thus the discretization of the convective term is given explicitly by

$$b_h(u_h, \varphi_i) := \frac{1}{3} b(p_i) \cdot \nabla u_h\big|_{T_i} \sum_{T \cap p_i \neq \emptyset} |T|. \qquad (3.8)$$
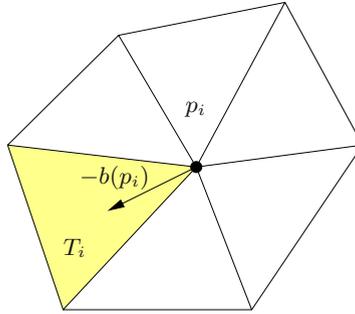
Fig. 3.1. Upwind triangle T_i associated with p_i

Example 3.3. Consider the upwind triangle method in the case $b(x) \equiv (1,0)$, $c(x) \equiv 0$ where we use a uniform square mesh of Friedrichs-Keller type as shown in Figure 3.2. Denote the distance between adjacent nodes by h. In this example each meshpoint p_{ij} has two upwind triangles T_{ij} and T_{ij}^* but

$$b \cdot \nabla u_h \big|_{T_{ij}} = b \cdot \nabla u_h \big|_{T_{ij}^*}.$$

Write $u_{ij} := u_h(p_{ij})$ for the discrete solution at the meshpoints. After scaling one gets the difference scheme

$$-\frac{\varepsilon}{h^2} \left(u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} \right) + \frac{1}{h}(u_{ij} - u_{i-1,j}) = f(p_{ij}),$$

for $i, j = 1, \ldots, N$. In this case the upwind triangle method coincides with the simple upwind scheme (2.3a) on setting $h_i = h_{i+1} = k_j = k_{j+1} = h$, $b_1 = 1$, and $b_2 = c = 0$. ♣
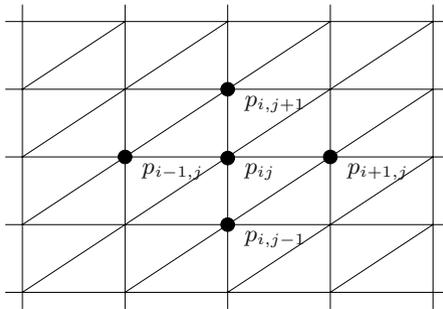
Fig. 3.2. Triangulation of Friedrichs-Keller type

To investigate the properties of the discrete problem (3.6)–(3.8), consider the matrix $L_h = (l_{ij})$ of the scheme, which is defined by

$$l_{ij} := \varepsilon(\nabla\varphi_j, \nabla\varphi_i) + b_h(\varphi_j, \varphi_i) + c_h(\varphi_j, \varphi_i) \quad \text{for } i, j = 1, \dots, N.$$

Theorem 3.4. *Assume that the coefficients b and c and the right-hand side f of (3.1) are sufficiently smooth, with $c(x) \geq 0$. Let $\mathcal{T}_h$ be a triangulation of weakly acute type. Then the discrete problem (3.6)–(3.8) preserves the inverse-monotonicity property of L, i.e., L_h is inverse-monotone.*

For the proof of Theorem 3.4 and further properties of the upwind triangle method see [Ike83, Tab77].

Remark 3.5. Bristeau et al. [BGP79] proposed a modification of the upwind triangle method that improves the approximation of the directional derivative and can be applied to arbitrary functions b for which $|b| \neq 0$ on weakly acute meshes. In this method the directional derivative is approximated using two points that are upwind of p_i with respect to the flow direction. A detailed study of the properties of the scheme including numerical experiments is given in [Kra87, KR92].

An extension of the Bristeau et al. scheme to a formally third-order scheme is proposed in [TF91], where an improved approximation of the directional derivative at p_i is sought by using two upwind points and two downwind points. ♣

We will now concentrate on methods that use some ideas from the finite volume method to discretize the convection term and combine them with the finite element approximation of the diffusion term. Thus these methods are sometimes called combined finite volume-finite element methods. This approach has been extended to nonstationary nonlinear convection-diffusion and compressible flow problems in [FFLM95, FFLM97, FSS99].

Example 3.3 showed that under certain circumstances the triangle up-wind method is equivalent to the simple upwind difference scheme, which in one dimension (see Section I.2.1.2) smears the boundary layer and so has excessive numerical viscosity. Can we construct methods that preserve the inverse-monotonicity of the continuous problem but have acceptable NVIS?

As above we use piecewise linear functions. For weakly acute meshes the discretization of $-\Delta$ is an M-matrix. Terms of the form $cu - f$ are handled either by a quadrature rule or by the lumping technique described below. The main emphasis in the following discussion is the discretization of the convective term $b \cdot \nabla u$, which will be carried out in a finite volume framework.

Start from a weakly acute triangulation of the polygonal bounded domain Ω. A *secondary grid* is another partition of Ω that is derived from the original triangulation. It takes two main forms: *barycentric* and *circumcentric*.

Let p_i, for $i = 1, \dots, N + M$, be the vertices of our triangulation, where $p_1, \dots, p_N$ are in the interior of Ω while the remaining p_i lie on Γ. A dual domain D_i of the secondary grid is associated with each p_i. This region D_i is defined in the barycentric case by
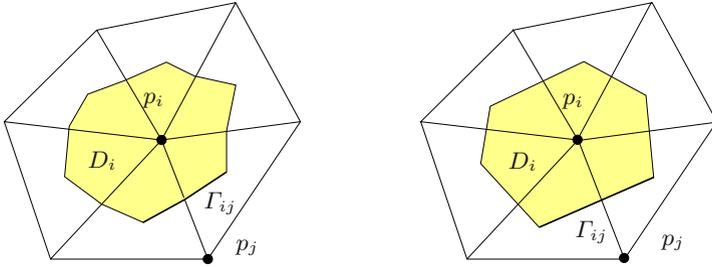
Fig. 3.3. Barycentric and circumcentric dual domains

$$D_i := \bigcup_{T \cap p_i \neq \emptyset} \{\, p \in T \,:\, \lambda_i^T(p) \geq \lambda_j^T(p) \text{ for all } j \text{ with } p_j \in T \,\}$$

and in the circumcentric case by

$$D_i := \bigcup_{T \cap p_i \neq \emptyset} \{\, p \in T \,:\, |p_i p| \leq |p_j p| \text{ for all vertices } p_j \in T \,\},$$

where $\lambda_i^T(p)$ is the barycentric coordinate of the point p with respect to T and $|\cdot|$ is the length of a line segment; see Figure 3.3. The boundary ∂D_i of D_i is polygonal in both cases. For $i = N+1, \dots, N+M$, some of ∂D_i lies on Γ. For $i = 1, \dots, N$, let Γ_{ij} denote the face(s) of ∂D_i that meet the line segment $p_i p_j$. In the barycentric case, Γ_{ij} comprises those parts of the medians of the two triangles that intersect the interior of $p_i p_j$. In the circumcentric case, Γ_{ij} lies on the perpendicular bisector of $p_i p_j$.

For later use, introduce the index set

$$\Lambda_i := \{\, j \neq i \,:\, \exists\, T \text{ with } p_i, p_j \in T \,\}$$

for $i = 1, \dots, N+M$. Let χ_i be the characteristic function of the dual domain D_i. Define the lumping operator l_h by

$$l_h w := \sum_{i=1}^{N+M} w(p_i) \chi_i.$$

We seek an approximation of the solution of (3.6) in the discrete space V_h of piecewise linears. To approximate the convective term, start from the identity

$$(b \cdot \nabla u, v) = (\mathrm{div}\,(bu), v) - (\mathrm{div}\, b, uv).$$

Applying the lumping operator and Green's formula, one obtains

$$(b \cdot \nabla u, v) \approx (\mathrm{div}\,(bu), l_h v) - (\mathrm{div}\, b, l_h(uv))$$

$$= \sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} v(p_i) \int_{\Gamma_{ij}} \big[u - u(p_i) \big]\, b \cdot n_{ij}\, d\gamma,$$

where n_{ij} denotes the unit normal to Γ_{ij} that points out of D_i. With upwinding in mind, replace u on Γ_{ij} by a linear combination of function values at the neighbouring nodes p_i and p_j:

$$u \approx \lambda_{ij}\, u(p_i) + (1 - \lambda_{ij})\, u(p_j) \quad \text{on } \Gamma_{ij}.$$

The parameter $\lambda_{ij} \in [0,1]$ controls the amount of upwinding; some guidance in its choice will be given below. The new discretization of the convective term is now

$$b_h(u_h, v_h) = \sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} \beta_{ij}\, (1 - \lambda_{ij}) \big[u_h(p_j) - u_h(p_i) \big]\, v_h(p_i),$$

for all $u_h, v_h \in V_h$, where β_{ij} is some approximation of the flux

$$\int_{\Gamma_{ij}} b \cdot n_{ij}\, d\gamma$$

across Γ_{ij}. Assume that for β_{ij} one has

(A1) $| \int_{\Gamma_{ij}} b \cdot n_{ij}\, d\gamma - \beta_{ij}| \le C\, h^3$,
(A2) $\beta_{ij} + \beta_{ji} = 0$ if $\Gamma_{ij} \cap \Gamma = \emptyset$.

Both (A1) and (A2) are satisfied if one uses the mid-point rule

$$\int_{\Gamma_{ij}} b \cdot n_{ij}\, d\gamma \sim |\Gamma_{ij}|\, b(q_{ij}) \cdot n_{ij} =: \beta_{ij},$$

where $q_{ij} := (p_i + p_j)/2$ is the mid-point of the line segment $p_i p_j$.

One simple choice of λ_{ij} is motivated by the sign of the approximated flux β_{ij} through Γ_{ij}: set $\lambda_{ij} = 1$ if $\beta_{ij} > 0$, and $\lambda_{ij} = 0$ if $\beta_{ij} < 0$. In practice the parameter λ_{ij} is often chosen as a function of the *mesh Peclet number* $\beta_{ij}/(2\varepsilon)$. We shall study the properties of the scheme obtained when λ_{ij} is determined by

$$\lambda_{ij} = \Phi\left(\frac{\beta_{ij}}{2\varepsilon} \right),$$

where $\Phi(\cdot)$ is a general *weighting function*. Let us assume that:

(B1) $\Phi(t) = 1 - \Phi(-t) \quad \forall\, t > 0 \quad \text{and} \quad 0 \le \Phi(t) \le 1 \quad \forall\, t \in \mathbb{R}$;
(B2) $t \left[\Phi(t) - \frac{1}{2} \right] \ge 0 \quad \forall\, t \in \mathbb{R}$;
(B3) $\Psi(t) := t\, \Phi(t)$ is Lipschitz continuous on $\mathbb{R}$.

Some candidates for $\Phi(\cdot)$ that satisfy the assumptions (B1)–(B3) and have been used in practical computations are

$$\Phi_1(t) = \begin{cases} 1 & \text{if } t \ge 0, \\ 0 & \text{if } t < 0, \end{cases} \quad \text{and} \quad \Phi_2(t) = \begin{cases} (1 + 2t)/(2 + 2t) & \text{if } t \ge 0, \\ 1/(2 - 2t) & \text{if } t < 0. \end{cases}$$

In fact this Φ_1 generates the choice of λ described at the beginning of the paragraph. In the present context, Φ_1 is called simple upwinding and Φ_2 Samarskiĭ upwinding (cf. Section I.2.1.2).

Example 3.6. Consider a uniform triangulation of Friedrichs-Keller type as in Figure 3.2, with a circumcentric secondary grid. Assume that $b(x) \equiv (1,0)$ and $c(x) \equiv 0$. Let h denote the distance between adjacent nodes. Writing $u_{ij} := u_h(p_{ij})$ for the discrete solution at the meshpoints, after scaling one gets the difference scheme

$$-\frac{\varepsilon}{h^2}\left(u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij}\right)$$

$$+ \alpha \, \frac{u_{ij} - u_{i-1,j}}{h} + (1-\alpha) \, \frac{u_{i+1,j} - u_{ij}}{h} = f(p_{ij}),$$

for $i, j = 1, \ldots, N$, where $\alpha := \Phi(h/(2\varepsilon))$. For simple upwinding one has $\alpha = 1$, which yields the simple upwind scheme (2.3a). Samarskiĭ upwinding, on the other hand, has

$$\alpha = \frac{1}{2}\left(1 + \frac{q}{1+q}\right) \quad \text{where } q = \frac{h}{2\varepsilon},$$

which approaches the central difference scheme as h/ε tends to zero and the simple upwind scheme as ε/h tends to zero. Compared with the upwind triangle method, it is apparent that this secondary grid method provides more flexibility in controlling the amount of upwinding and numerical viscosity. ♣

The reaction term cu and the right-hand side f of (3.1) are discretized via the following lumping procedure:

$$(c\,u_h, v_h) \approx (l_h(c\,u_h), l_h v_h) = \sum_{i=1}^{N+M} |D_i|\, c(p_i) u_h(p_i) v_h(p_i),$$

$$(f, v_h) \approx (l_h f, l_h v_h) = \sum_{i=1}^{N+M} |D_i|\, f(p_i) v_h(p_i).$$

In the case of a barycentric secondary grid, one has

$$|D_i| = \frac{1}{3} \sum_{T \cap p_i \neq \emptyset} |T|,$$

so lumping can also be viewed as a simple quadrature rule.

The discrete problem can now be formulated:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a_h(u_h, v_h) = f_h(v_h), \tag{3.9a}$$

where

$$a_h(u_h, v_h) := \varepsilon(\nabla u_h, \nabla v_h) + b_h(u_h, v_h) + c_h(u_h, v_h), \tag{3.9b}$$

$$b_h(u_h, v_h) := \sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} \beta_{ij}\left(1 - \Phi\left(\frac{\beta_{ij}}{2\varepsilon}\right)\right)[u_h(p_j) - u_h(p_i)] \, v_h(p_i), \tag{3.9c}$$

$$c_h(u_h, v_h) := \sum_{i=1}^{N+M} |D_i| \, c(p_i) u_h(p_i) v_h(p_i), \tag{3.9d}$$

$$f_h(v_h) := \sum_{i=1}^{N+M} |D_i| \, f(p_i) v_h(p_i). \tag{3.9e}$$

In the case of simple upwinding, the following result is valid:

Theorem 3.7. *Assume that the coefficients b and c and the right-hand side f of (3.1) are sufficiently smooth, and that $c(x) \geq 0$. Let $\mathcal{T}_h$ be a weakly acute triangulation. Let the weighting function be $\Phi(t) := 1/2 \, (1 + \operatorname{sgn} t)$. Furthermore, assume that the approximation of the flux satisfies (A1) and (A2). Then the discrete problem (3.9) preserves the inverse-monotonicity property, i.e., the matrix L_h of the associated difference scheme is inverse-monotone.*

Proof. Consider the convective term. For $i \neq j$ and $t_{ij} := \beta_{ij}/(2\varepsilon)$,

$$b_h(\varphi_j, \varphi_i) = \beta_{ij}(1 - \Phi(t_{ij})) = 2\varepsilon \, t_{ij}(1 - \Phi(t_{ij})) \leq 0.$$

Hence the off-diagonal entries of L_h are non-positive. Choosing $e = (1, \ldots, 1)$, one can verify condition (iii) of Theorem 3.1. □

Corollary 3.8. *Assume the hypotheses of Theorem 3.7, and that one has a majorizing element $e_h \in V_h$ that satisfies $L_h e_h \geq (e_0, e_0, \ldots, e_0) > 0$ and $\|e_h\|_{\infty,d} \leq e_{\max}$ with constants e_0 and $e_{\max}$ that are independent of h and ε. Then the discrete problem (3.9) is L_∞ stable uniformly with respect to ε, i.e.,*

$$\|v_h\|_{\infty,d} \leq \frac{e_{\max}}{e_0} \, \|L_h v_h\|_{\infty,d} \qquad \forall \, v_h \in V_h.$$

We now discuss error estimates for the method (3.9), using general weighting functions Φ, in the ε-weighted H^1 norm

$$\|v_h\|_\varepsilon := \left(\varepsilon |v_h|_1^2 + \|v_h\|_0^2\right)^{1/2},$$

a one-dimensional analogue of which was used in Section I.2.2.2. Our results are based on the V_h-ellipticity of the bilinear form $a_h(\cdot, \cdot)$, which preserves the ellipticity of the continuous problem.

Lemma 3.9. *Suppose that the coefficients b, c and f of (3.1) are sufficiently smooth and that $c - \frac{1}{2}\nabla \cdot b \geq c_0 > 0$. Let the assumptions (A1), (A2), (B1) and (B2) be satisfied. Then there is a positive constant h_0, independent of ε, such that for all $h < h_0$ one has*

$$a_h(v_h, v_h) \geq \varepsilon |v_h|_1^2 + \frac{c_0}{2} \|v_h\|_0^2 \quad \forall v_h \in V_h; \tag{3.10}$$

that is, $a_h(\cdot, \cdot)$ is V_h-elliptic with respect to $\|\cdot\|_\varepsilon$.

Proof. Now

$$a_h(v_h, v_h) = \varepsilon |v_h|_1^2 + b_h(v_h, v_h) + c_h(v_h, v_h).$$

In $b_h(v_h, v_h)$ the summation index i runs only from 1 to N since $v_h(p_i) = 0$ for $i = N+1, \ldots, N+M$. Writing the convective term in two ways and using (B2) and (B1) yields, with $\lambda_{ij} = \Phi(\beta_{ij}/(2\varepsilon))$,

$$b_h(v_h, v_h) = \frac{1}{2} \sum_{i=1}^N \sum_{j \in \Lambda_i} \beta_{ij}(1 - \lambda_{ij})[v_h(p_j) - v_h(p_i)]v_h(p_i)$$

$$+ \frac{1}{2} \sum_{i=1}^N \sum_{j \in \Lambda_i} \beta_{ji}(1 - \lambda_{ji})[v_h(p_i) - v_h(p_j)]v_h(p_j)$$

$$= \frac{1}{2} \sum_{i=1}^N \sum_{j \in \Lambda_i} \beta_{ij}[v_h(p_j) - v_h(p_i)][(1 - \lambda_{ij})v_h(p_i) + \lambda_{ij}v_h(p_j)]$$

$$= \frac{1}{2} \sum_{i=1}^N \sum_{j \in \Lambda_i} \beta_{ij}\left(\lambda_{ij} - \frac{1}{2}\right)[v_h(p_j) - v_h(p_i)]^2$$

$$- \frac{1}{2} \sum_{i=1}^N \sum_{j \in \Lambda_i} \beta_{ij}\, [v_h(p_i)]^2$$

$$\geq -\frac{1}{2} \sum_{i=1}^N \sum_{j \in \Lambda_i} \beta_{ij}\, [v_h(p_i)]^2\, .$$

Recall assumption (A1) and note that

$$\sum_{j \in \Lambda_i} \int_{\Gamma_{ij}} b \cdot n_{ij}\, ds = \int_{\partial D_i} b \cdot n_i\, ds = \int_{D_i} \operatorname{div} b\, dx.$$

We therefore obtain

$$b_h(v_h, v_h) + c_h(v_h, v_h) \geq \sum_{i=1}^{N} [v_h(p_i)]^2 \int_{D_i} \left(c - \frac{1}{2} \mathrm{div}\, b \right) dx$$

$$+ \sum_{i=1}^{N} [v_h(p_i)]^2 \int_{D_i} (c(p_i) - c)\, dx$$

$$+ \frac{1}{2} \sum_{i=1}^{N} [v_h(p_i)]^2 \sum_{j \in \Lambda_i} \left(\int_{\Gamma_{ij}} b \cdot n_{ij}\, ds - \beta_{ij} \right)$$

$$\geq c_0 \sum_{i=1}^{N} |D_i|\, [v_h(p_i)]^2 - \mathcal{O}(h^3) \sum_{i=1}^{N} [v_h(p_i)]^2$$

$$\geq \frac{c_0}{2}\, \|v_h\|_0^2,$$

since the discrete norm $\left(\sum_{i=1}^{N} |D_i|\, [v_h(p_i)]^2 \right)^{1/2}$ and the continuous L^2 norm $\|v_h\|_0$ are equivalent on V_h. $\square$

Lemma 3.9 implies existence of a unique solution to (3.9). An error estimate in $\|\cdot\|_\varepsilon$ for this solution can now be proved. While doing this we consider also triangulations of Ω by three-directional meshes. Such meshes have the following properties:

(i) each element side is parallel to one of three fixed directions;
(ii) each inner node is surrounded by six triangles;
(iii) each inner node has six neighbouring nodes – two in each of the three fixed directions.

An example of a three-directional mesh is a triangulation of Friedrichs-Keller type (see Figure 3.2).

Theorem 3.10. *Suppose that b, c and f in (3.1) are sufficiently smooth and that $c - \frac{1}{2} \nabla \cdot b \geq c_0 > 0$. Assume that (A1), (A2), (B1), (B2) and (B3) hold true and that $u \in H^2(\Omega)$, where u is the solution of (3.1). Let u_h be the solution of (3.9). Then for all $h < h_0$, with h_0 independent of ε,*

$$\|u - u_h\|_\varepsilon \leq C\, \frac{h}{\sqrt{\varepsilon}}\, (\|u\|_2 + \|f\|_{1,p}) \tag{3.11a}$$

for $p > 2$. This error estimate can be strengthened to

$$\|u - u_h\|_\varepsilon \leq C\, h\, (\|u\|_2 + \|f\|_{1,p}) \tag{3.11b}$$

if Ω is triangulated by a three-directional mesh.

Proof. The proof, many of whose details we omit, is based on the following modification of Strang's first lemma (see [Cia02]):

$$\|u - u_h\|_\varepsilon \leq C \left\{ \inf_{v_h \in V_h} \left[\|u - v_h\|_1 + \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_\varepsilon} \right] \right.$$
$$\left. + \sup_{w_h \in V_h} \frac{|(f, w_h) - f_h(w_h)|}{\|w_h\|_\varepsilon} \right\}.$$

Let us explain how the presence of a three-directional mesh improves the result. The assumption $f \in W^{1,p}(\Omega)$, $p > 2$, guarantees that f is continuous; hence $l_h f$ is well-defined. Furthermore, the approximation property

$$\|f - l_h f\|_{0,p} \leq Ch\|f\|_{1,p} \quad \forall f \in W^{1,p}(\Omega)$$

holds true for the lumping operator l_h. In the case of a general mesh, defining q by $1/p + 1/q = 1$, one can invoke the standard argument

$$\begin{aligned} |(f, w_h) - f_h(w_h)| &= |(f, w_h - l_h w_h) + (f - l_h f, l_h w_h)| \\ &\leq C \, h \, (\|f\|_0 \, |w_h|_1 + \|f\|_{1,p} \, \|l_h w_h\|_{0,q}) \\ &\leq C \, \frac{h}{\sqrt{\varepsilon}} \, \|f\|_{1,p} \, \|w_h\|_\varepsilon \,, \end{aligned}$$

where one uses $\|f\|_{0,2} \leq C\|f\|_{0,p}$, the stability result $\|l_h w_h\|_{0,q} \leq C\|w_h\|_{0,q}$, and $\|w_h\|_{0,q} \leq \|w_h\|_{0,2}$. This last inequality follows from the equivalence of the norms $\hat{w} \mapsto \|\hat{w}\|_{0,q,\hat{T}}$ and $\hat{w} \mapsto \|\hat{w}\|_{0,2,\hat{T}}$ in the finite-dimensional space on the reference cell $\hat{T}$, scaling properties and summing over all cells $T \in \mathcal{T}_h$. But for a three-directional mesh, it can be shown that

$$(v_h, l_h w_h) = (l_h v_h, w_h) \quad \forall v_h, \, w_h \in V_h.$$

Consequently, writing I_h for the piecewise linear nodal interpolation operator, one has

$$\begin{aligned} |(f, w_h) - f_h(w_h)| &= |(f - I_h f, w_h - l_h w_h) + (I_h f, w_h - l_h w_h) \\ &\quad + (f - l_h f, l_h w_h)| \\ &= |(f - I_h f, w_h - l_h w_h) + (I_h f - l_h I_h f, w_h) \\ &\quad + (f - l_h f, l_h w_h)| \\ &\leq Ch \, (\|f\|_{1,p} \, \|w_h\|_{0,q} + \|I_h f\|_{1,p} \, \|w_h\|_{0,q} \\ &\quad + \|f\|_{1,p} \, \|l_h w_h\|_{0,q}) \\ &\leq Ch\|f\|_{1,p} \, \|w_h\|_\varepsilon. \end{aligned}$$

The consistency error of the bilinear form clearly satisfies

$$\begin{aligned} |a(v_h, w_h) &- a_h(v_h, w_h)| \\ &\leq |(b \cdot \nabla v_h, w_h) - b_h(v_h, w_h)| + |(c \, v_h, w_h) - c_h(v_h, w_h)|. \end{aligned}$$

The last term can be handled in the same way as f above. The estimate of the convective term is very technical and is not presented here. For the

simple upwind method, i.e., for the weighting function $\Phi(t) = (1 + \operatorname{sgn} t)/2$, Risch [Ris86] proves

$$|(b \cdot \nabla I_h u, w_h) - b_h(I_h u, w_h)| \leq C\, h\, \|u\|_2 \, \|w_h\|_1$$

on general meshes, and in the case of a three-directional mesh the bound

$$|(b \cdot \nabla I_h u, w_h) - b_h(I_h u, w_h)| \leq C\, h\, \|u\|_2 \, \|w_h\|_0$$

is derived.   □

Remark 3.11. For the standard Galerkin finite element method with piecewise linears, one can easily prove the energy norm estimate

$$\|u - u_h\|_\varepsilon \leq C\, h\, \|u\|_2$$

on shape-regular meshes, but this method is only L^2 stable as $\varepsilon \to 0$. The method of Corollary 3.8 is L_∞ stable on fairly general meshes under reasonable assumptions, but (3.11a) is inferior to the bound just stated for the standard Galerkin method. Is the standard Galerkin method more accurate? No; numerical computations show that for $\varepsilon \ll h$ it has a wildly oscillatory solution that is clearly unsatisfactory.   ♣

Remark 3.12. The estimates (3.11) seem useless because in general $\|u\|_2$ is large when ε is small, but Risch [Ris86] has obtained analogous estimates on subdomains $\Omega' \subset \Omega$ by means of special cut-off functions. (This technique was first applied to convection-diffusion problems in [Näv82], to analyse the streamline diffusion method.) More precisely, Risch proves the local error estimates

$$\|u - u_h\|_{\varepsilon, \Omega'} \leq C\, \frac{h}{\sqrt{\varepsilon}}\, \left(\|u\|_{2, \Omega''} + \|f\|_{1, p, \Omega''} \right) \qquad (3.12a)$$

with $p > 2$ on a general mesh, and

$$\|u - u_h\|_{\varepsilon, \Omega'} \leq C\, h\, \left(\|u\|_{2, \Omega''} + \|f\|_{1, p, \Omega''} \right) \qquad (3.12b)$$

if Ω'' is triangulated by a three-directional mesh. The subdomains Ω'' and Ω', with $\Omega' \subset \Omega'' \subset \Omega$, should be chosen so that Ω'' intersects no layer of the solution u. In this way one gets a ε-uniform local convergence result (cf. Theorem 3.41, which is a similar result for the streamline diffusion method).   ♣

Next we examine local estimates in the L_∞ norm that can be applied on subdomains $\Omega' \subset \Omega'' \subset \Omega$, where one should choose Ω'' to avoid layers in the solution u.

Theorem 3.13. *Suppose that b, c and f in (3.1) are sufficiently smooth and that $c - \frac{1}{2}\nabla \cdot b \geq c_0 > 0$. Assume that the solution u of (3.1) belongs to*

$H_0^1(\Omega) \cap W^{2,p}(\Omega)$ *for some* $p > 2$. *Let the triangulation of* Ω *be weakly acute. Assume that (A1) and (A2) hold true and that the weighting function is* $\Phi(t) = (1 + \operatorname{sgn} t)/2$. *Let* u_h *be the solution of (3.9). Then for* $h < h_0$, *with* h_0 *independent of* ε, *one has the local error estimate*

$$\|u - u_h\|_{0,\infty,\Omega'} \le C \, \frac{h^{\kappa(p)} |\ln h|}{\sqrt{\varepsilon}} \, \|u\|_{2,p,\Omega''} , \tag{3.13a}$$

where $1/2 < \kappa(p) < 1$ *and* $\lim_{p \to \infty} \kappa(p) = 1$. *This estimate can be strengthened to*

$$\|u - u_h\|_{0,\infty,\Omega'} \le C \, h^{\kappa(p)} |\ln h| \, \|u\|_{2,p,\Omega''} \tag{3.13b}$$

if Ω'' *is triangulated by a three-directional mesh.*

The proof of this theorem in [Ris86, Ris90] employs cut-off functions Ψ, a variant of V_h-ellipticity in a Ψ-weighted energy norm, and the M-matrix property of the discrete problem. The technique of [Näv82] cannot be applied directly to secondary grid methods because they lack Galerkin orthogonality. Risch's proof exploits the inverse-monotonicity of the discrete problem and is suited only to monotonicity-preserving discretizations.

Remark 3.14. The local error estimates (3.12) and (3.13) are very useful if the domain Ω'' does not include interior or boundary layers. As we know from Section 1.3, the position of layers in the solution u of (3.1) depends on the data of the problem. A typical situation is the case when $b = (b_1, 0)$ with $b_1 > 0$, f is smooth and Ω is the unit square: one then has boundary layers along the sides $x = 1$, $y = 0$ and $y = 1$ of Ω (see Section 1.2). In the local error estimates (3.12), (3.13), when ε and h are small, it is therefore sensible to choose $\Omega' = (0, 1 - \sigma') \times (\delta', 1 - \delta')$ and $\Omega'' = (0, 1 - \sigma'') \times (\delta'', 1 - \delta'')$ with $0 < \sigma'' < \sigma' << 1$ and $0 < \delta'' < \delta' << 1$. More general situations are studied in [Ris86, Ris90]. ♣

In deriving the secondary grid method (3.9), the convective term was discretized using essentially a finite volume technique, while the diffusion term was handled in a purely finite element manner. Also, if $\varepsilon = 0$ and $c = 0$, then the non-positivity of the off-diagonal entries of the corresponding matrix is ensured only if the weighting function is $\Phi(t) = 1/2 \, (1 + \operatorname{sgn} t)$. Furthermore, for $\varepsilon > 0$ the diffusion term can yield negative off-diagonal entries. Taking these facts into consideration, it seems worthwhile to look for discretizations that treat

$$-\varepsilon \Delta u + b \cdot \nabla u$$

as a single term, unlike our previous approach.

With this aim in mind, for convenience rewrite (3.1) in the form

$$\operatorname{div}\left(-\varepsilon \nabla u + bu\right) + (c - \operatorname{div} b)\, u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma. \tag{3.14}$$

Let the triangulation of Ω be weakly acute. Construct a secondary grid of circumcentric type as before. We retain our old notation: the inner nodes are p_i for $i = 1, \ldots, N$, the dual domain around each inner node p_i is D_i, the index set of neighbour nodes of p_i is Λ_i, while Γ_{ij} denotes the face of ∂D_i that meets the line segment $p_i p_j$ and n_{ij} is the unit normal to Γ_{ij} that points out of D_i. Let m_i denote the area of D_i, d_{ij} the length of the line segment $p_i p_j$, and m_{ij} the length of Γ_{ij}.

Integrating (3.14) over D_i and applying Green's formula yields

$$-\sum_{j \in \Lambda_i} \int_{\Gamma_{ij}} n_{ij} \cdot (\varepsilon \nabla u - bu)\, ds + \int_{D_i} (c - \nabla \cdot b) u\, dx = \int_{D_i} f\, dx. \qquad (3.15)$$

To simplify the notation, set

$$N_{ij} = \frac{1}{m_{ij}} \int_{\Gamma_{ij}} (n_{ij} \cdot b)\, ds, \quad c_i = \frac{1}{m_i} \int_{D_i} c\, dx, \quad f_i = \frac{1}{m_i} \int_{D_i} f\, dx.$$

Apply the lumping operator l_h to the second term in (3.15):

$$\int_{D_i} (c - \nabla \cdot b) u\, dx \approx \int_{D_i} (c - \nabla \cdot b)\,(l_h u)\, dx$$

$$= c_i m_i u(p_i) - \sum_{j \in \Lambda_i} N_{ij} m_{ij} u(p_i).$$

The term

$$-\sum_{j \in \Lambda_i} \int_{\Gamma_{ij}} n_{ij} \cdot (\varepsilon \nabla u - bu)\, ds. \qquad (3.16)$$

must still be approximated. This can be done by replacing the integrand by a constant, i.e.,

$$-n_{ij} \cdot (\varepsilon \nabla u - bu) \approx S_{ij}.$$

To define this S_{ij}, we consider the following boundary value problem on the line segment between p_i and p_j:

$$-\frac{\varepsilon}{d_{ij}} \frac{dw}{d\xi} + N_{ij} w = S_{ij}, \quad w(0) = u(p_i),\ w(1) = u(p_j),$$

where $w = w(\xi)$ for $0 \le \xi \le 1$. This yields

$$S_{ij} = N_{ij} \frac{u(p_i) \exp(N_{ij} d_{ij} / \varepsilon) - u(p_j)}{\exp(N_{ij} d_{ij} / \varepsilon) - 1},$$

and the integral in (3.16) is replaced by $S_{ij} m_{ij}$.

Assembling these approximations gives finally the system of discrete equations

$$\sum_{j \in \Lambda_i} \frac{\varepsilon m_{ij}}{d_{ij}} B\left(\frac{N_{ij}d_{ij}}{\varepsilon}\right)(u_i - u_j) + c_i m_i u_i = f_i m_i \tag{3.17}$$

for $i = 1, \ldots, N$, where $B(\cdot)$ denotes the Bernoulli function defined by

$$B(t) = \begin{cases} \dfrac{t}{\exp t - 1} & \text{for } t \neq 0, \\ 1 & \text{for } t = 0. \end{cases}$$

and $u_k = u(p_k)$ for all k. This scheme was studied by Angermann in several papers[Ang91b, Ang93, Ang95b], with a more general control function

$$\widetilde{B}(t) = 1 - t[1 - r(t)], \tag{3.18}$$

under the assumptions that

(C1) $r(t)$ is monotone for all $t \in \mathbb{R}$,
(C2) $\lim\limits_{t \to -\infty} r(t) = 0, \quad \lim\limits_{t \to +\infty} r(t) = 1$,
(C3) $1 + r(t)t \geq 0 \quad$ for all $t \in \mathbb{R}$,
(C4) $[1 - r(t) - r(-t)]t = 0 \quad$ for all $t \in \mathbb{R}$,
(C5) $[r(t) - 1/2]t \geq 0 \quad$ for all $t \in \mathbb{R}$,
(C6) the mapping $t \mapsto t\,r(t)$ is Lipschitz continuous on $\mathbb{R}$.

In particular the choice

$$r(t) = 1 - \frac{1}{t}\left(1 - \frac{t}{\exp t - 1}\right)$$

satisfies assumptions (C1)–(C6) and $\widetilde{B}$ is then the Bernoulli function B.

Remark 3.15. The scheme (3.17)–(3.18) resembles the scheme (3.9), but the control functions $\Phi(\cdot)$ and $r(\cdot)$ depend on different mesh Peclet numbers, namely

$$\frac{N_{ij}m_{ij}}{2\varepsilon} \quad \text{and} \quad \frac{N_{ij}d_{ij}}{\varepsilon}$$

respectively. In the case of a uniform mesh of Friedrichs-Keller type with meshsize h, one has $m_{ij} = d_{ij} = h$, and if $\Phi(t) = r(2t)$ then the schemes are identical. The averaged finite element scheme developed by Xu and Zikatanov [XZ99] for any space dimension d is also closely related to (3.17)–(3.18); these schemes are identical for $d = 2$ and constant functions b. ♣

We state the analogue of Theorem 3.10 for the present scheme.

Theorem 3.16. *Suppose that b, c and f in (3.14) are sufficiently smooth and that $c - \frac{1}{2}\nabla \cdot b \geq c_0 > 0$. Let $\mathcal{T}_h$ be a weakly acute triangulation. Let u_h be the solution of (3.17)–(3.18), and assume that the control function $r(\cdot)$ satisfies (C1)–(C6). Then for $h < h_0$, with h_0 independent of ε, the discrete problem*

(3.17)–(3.18) *is inverse-monotone. If $u \in H^2(\Omega)$, where u is the solution of* (3.14), *then*

$$\|u - u_h\|_\varepsilon \le C \, \frac{h}{\sqrt{\varepsilon}} \, (\|u\|_2 + \|f\|_{1,p}) \tag{3.19a}$$

for $p > 2$. This error estimate can be strengthened to

$$\|u - u_h\|_\varepsilon \le C \, h \, (\|u\|_2 + \|f\|_{1,p}) \tag{3.19b}$$

if Ω is triangulated by a three-directional mesh.

Proof. See [Ang95b]. □

Remark 3.17. Angermann [Ang95b] asks the interesting question: can one choose the control function $r(\cdot)$ for an arbitrary mesh in such a way that the factor $1/\sqrt{\varepsilon}$ can be removed from the error bound (3.19a)? While this would not give an ε-uniformly convergent method, it would nevertheless be helpful in the context of adaptive methods (see Section 3.6). Unfortunately, on general meshes the answer to this question is negative.   ♣

Remark 3.18. Angermann [Ang95c] and Johannsen [Joh96] consider finite volume schemes where the control volumes (dual domains) are aligned with the piecewise constant approximation of the convection field b, i.e., the edges of the control volumes are parallel or perpendicular to the convection field. Though no complete theoretical analysis is available, numerical experiments show the effectiveness of this finite volume variant.   ♣

Finally, we discuss two extensions to nonconforming methods, i.e., methods where $V_h \not\subset V := H_0^1(\Omega)$. Nonconforming finite element methods are very useful when solving the incompressible Navier-Stokes equations (see Chapter IV.2). Here only piecewise linear basis functions are examined.

First, consider a method closely related to (3.9). Start from a weakly acute triangulation. Denote by B_i, for $i = 1, \ldots, N$, all mid-points of inner edges of triangles T of the triangulation and by B_i, for $i = N + 1, \ldots, N + M$, the mid-points of edges lying on the boundary Γ of Ω. We introduce the nonconforming finite element space

$$V_h := \{\, v_h \,:\, v_h\big|_T \in P_1(T) \,\forall T, \quad v_h \text{ continuous at } B_i \text{ for } i = 1, \ldots, N,$$
$$v_h(B_i) = 0, \text{ for } i = N + 1, \ldots, N + M \,\}$$

In general $V_h \not\subset C(\Omega)$, so our discretization will be nonconforming. Consequently $(\nabla u_h, \nabla v_h)$ is not defined and is replaced by $(\nabla u_h, \nabla v_h)_h$, where the inner products are computed element by element. With each inner node B_i, for $i = 1, \ldots, N$, is associated a dual domain D_i of the secondary grid, as shown in Figure 3.4. We complete our secondary grid by analogously defining dual domains D_i that correspond to the boundary nodes B_i for $i = N+1, \ldots, N+M$.

Let Γ_{ij} denote the segment of ∂D_i that intersects the line segment $B_i B_j$. For each i, set

$$\Lambda_i := \{j \neq i \ : \ \exists T \text{ with } B_i, B_j \in T\}.$$

Unlike the earlier conforming case, Λ_i contains at most four indices irrespective of the triangulation.
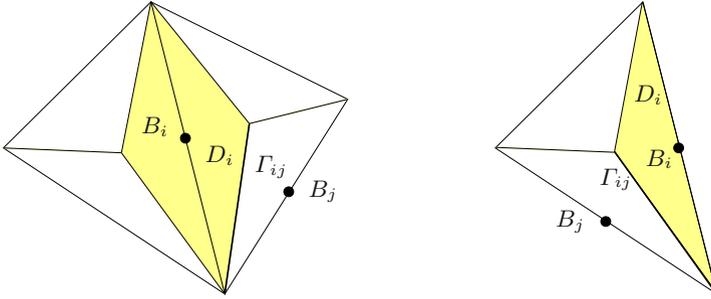

Fig. 3.4. Dual domains D_i associated with an inner (left) and boundary (right) node B_i for the nonconforming element

Define a lumping operator l_h by

$$l_h w := \sum_{i=1}^{N+M} w(B_i)\chi_i,$$

where χ_i is the characteristic function of the dual domain D_i. Then, retracing our steps in the derivation of the scheme (3.9), we arrive at the discrete problem:

Find $u_h \in V_h$ such that, for all $v_h \in V_h$,

$$a_h(u_h, v_h) = f_h(v_h), \tag{3.20a}$$

where

$$a_h(u_h, v_h) := \varepsilon(\nabla u_h, \nabla v_h)_h + b_h(u_h, v_h) + c_h(u_h, v_h), \tag{3.20b}$$

$$b_h(u_h, v_h) := \sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} \beta_{ij}\left(1 - \Phi\left(\frac{\beta_{ij}}{2\varepsilon}\right)\right)[u_h(B_j) - u_h(B_i)]\, v_h(B_i), \tag{3.20c}$$

$$c_h(u_h, v_h) := \sum_{i=1}^{N+M} |D_i|\, c(B_i)u_h(B_i)v_h(B_i), \tag{3.20d}$$

$$f_h(v_h) := \sum_{i=1}^{N+M} |D_i|\, f(B_i)v_h(B_i). \tag{3.20e}$$

Theorem 3.19. *Assume that the coefficients b and c and the right-hand side f of (3.1) are sufficiently smooth, and that $c(x) \geq 0$. Let $\mathcal{T}_h$ be a weakly acute triangulation. Let the weighting function be $\Phi(t) := (1+\mathrm{sgn}\, t)/2$. Furthermore, assume that the approximation of the flux satisfies (A1) and (A2). Then the discrete problem (3.20) preserves the inverse-monotonicity property, i.e., the matrix L_h of the associated difference scheme is inverse-monotone.*

Proof. The proof is similar to that of Theorem 3.10. See also [OU84].   □

Remark 3.20. Error estimates similar to (3.11) are also true in the noncon-forming case, as one might expect. In particular on three-directional meshes the factor $1/\sqrt{\varepsilon}$ can be removed by using the property

$$(v_h, l_h w_h) = (l_h v_h, w_h)$$

as described in the proof of Theorem 3.10.   ♣

The last method that we describe in this section is the nonconforming finite element method of [MW94]. The problem (3.1) is rewritten as

$$-\mathrm{div}\,(\varepsilon \nabla u - bu) + du = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \Gamma. \tag{3.21}$$

Assume that the functions b, d and f are sufficiently smooth and that

$$b_1^2 + b_2^2 \neq 0, \quad d + \frac{1}{2}\nabla \cdot b \geq 0 \quad \text{and } d \geq 0 \text{ on } \bar{\Omega}. \tag{3.22}$$

Begin again with a triangulation of Ω such that, for every triangle T, the circumcircle of T contains no vertices other than vertices of T itself. As mentioned earlier such a triangulation is called a Delaunay triangulation and ensures that the discretization of $-\Delta$ by piecewise linear elements yields an M-matrix.

Our usual notation is used. With each inner vertex p_i, for $i = 1, \ldots, N$, associate the circumcentric dual domain

$$D_i := \bigcup_{T \cap p_i \neq \emptyset} \{ p \in T : |p_i p| \leq |p_j p| \text{ for all vertices } p_j \in T \}$$

of the secondary grid. Let Γ_{ij} denote the face of ∂D_i that crosses the line segment $p_i p_j$. Let γ_{ij} be a unit vector along Γ_{ij} that is oriented in an anti-clockwise manner around p_i and let d_{ij} denote the length of the segment $p_i p_j$; see Figure 3.5.

Complete the secondary grid by analogously defining dual domains D_i that correspond to the boundary nodes P_i for $i = N + 1, \ldots, N + M$. Some faces of their boundaries ∂D_i will lie on the boundary Γ. Define the index set

$$\Lambda_i := \{ j \neq i : \exists T \text{ with } p_i, p_j \in T \}.$$
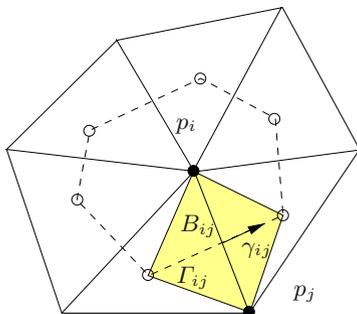
Fig. 3.5. Dual domain D_i bounded by the set of Γ_{ij} and the box B_{ij}

Two finite-dimensional finite element spaces will be used – one defined on each of our two grids. Using the characteristic function χ_i of the dual domain D_i, define the lumping operator l_h by

$$l_h w := \sum_{i=1}^{N+M} w(p_i)\chi_i,$$

which maps into the piecewise constant space

$$Q_h := \{q_h \ : \ q_h|_T \in P_0(T)\} = \text{span}\ \{\chi_1, \ldots, \chi_{N+M}\}.$$

To construct a nonconforming piecewise exponential space V_h, proceed as follows. Write e_{ij} for the line segment $p_i p_j$. Define the exponential function Φ_{ij} on e_{ij} to be the solution of the two-point boundary value problem

$$\frac{d}{dn_{ij}}\left(\varepsilon\frac{d\Phi_{ij}}{dn_{ij}} - \bar{b}_{ij}\Phi_{ij}\right) = 0, \quad \Phi_{ij}(p_i) = 1, \quad \Phi_{ij}(p_j) = 0, \tag{3.23}$$

where n_{ij} denotes the unit vector from p_i to p_j and $\bar{b}_{ij}$ is a piecewise constant approximation of $b \cdot n_{ij}$ that preserves constants – for instance, we can set $\bar{b}_{ij} := [(b(p_i) + b(p_j)) \cdot n_{ij}]/2$. Let B_{ij} be the quadrilateral whose diagonals are Γ_{ij} and e_{ij}; see Figure 3.5. Then extend the domain of definition of Φ_{ij} to B_{ij} by defining Φ_{ij} to be constant along each perpendicular to e_{ij}. We can now define a canonical basis function for V_h as follows:

$$\Phi_i = \begin{cases} \Phi_{ij} & \text{on } B_{ij} \text{ if } j \in \Lambda_i, \\ 0 & \text{otherwise.} \end{cases}$$

The finite element space $V_h := \text{span}\ \{\Phi_1, \ldots, \Phi_N\}$, which satisfies the homogeneous boundary conditions, is not a subset of $H_0^1(\Omega)$ since in general it lacks the property of continuity across the boundaries of boxes B_{ij}. Finally, on each box B_{ij} define the vector-valued function $\hat{b}_{ij}$ by

$$\hat{b}_{ij} = \bar{b}_{ij}n_{ij} + (b \cdot \gamma_{ij})\gamma_{ij}. \tag{3.24}$$

Now we are ready to define our discrete Petrov-Galerkin problem:

Find $u_h \in V_h$ such that for all $q_h \in Q_h$ one has

$$a_h(u_h, q_h) + (l_h(du_h), q_h) = (f, q_h), \tag{3.25a}$$

where the bilinear form $a_h : V_h \times Q_h \to \mathbb{R}$ is given by

$$a_h(u_h, q_h) := -\sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} \int_{\Gamma_{ij}} (\varepsilon \nabla u_h - \hat{b}_{ij} u_h) \cdot n_{ij} q_h \; ds. \tag{3.25b}$$

One can write (3.25) as the system of discrete equations

$$\sum_{j \in \Lambda_i} \frac{\varepsilon m_{ij}}{d_{ij}} \left[B\left(-\frac{\bar{b}_{ij}d_{ij}}{\varepsilon} \right) u_i - B\left(\frac{\bar{b}_{ij}d_{ij}}{\varepsilon} \right) u_j \right] + d_i m_i u_i = \int_{D_i} f \; dx, \tag{3.26}$$

where $B(t) := t/(\exp(t) - 1)$ is the Bernoulli function, m_{ij} denotes the length of Γ_{ij}, m_i is the area of D_i, $u_i := u_h(p_i)$ and $d_i = d(p_i)$.

Theorem 3.21. *Assume that the coefficients b and d of (3.21) satisfy (3.22). Let $\mathcal{T}_h$ be a Delaunay triangulation. Then the discrete problem (3.26) preserves the inverse-monotonicity of the continuous problem (3.21).*

Proof. Since the Bernoulli function $B(\cdot)$ is always positive, the coefficient matrix of (3.22) has non-positive off-diagonal entries and positive diagonal entries. Moreover, the chain property of Theorem 3.1 (iii) can be proved using $e = (1, \dots, 1)$. The coefficient matrix is therefore an M-matrix. $\square$

Remark 3.22. The scheme (3.26) in [MW94] is similar in construction to the scheme (3.17) of Angermann. The principal difference is the use of piecewise linears by Angermann whereas Miller & Wang use exponentials. In fact (3.26) is a type of *exponential box scheme*; such schemes are widely used when solving the drift-diffusion equations of semiconductor device modelling [PHSM87]. ♣

In [MW94], the error of the solution of (3.26) is studied in the mesh-dependent norm

$$|||v_h|||_{MW} := \left[\sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} h \; \text{area}(B_{ij}) \left(\frac{v_h(p_i) - v_h(p_j)}{d_{ij}} \right)^2 \right.$$

$$\left. + [v_h(p_i)]^2 \left(d_i m_i + \frac{1}{2} \int_{\Gamma_{ij}} \hat{b}_{ij} \cdot n_{ij} \; ds \right) \right]^{1/2}, \tag{3.27}$$

under the hypothesis that

(A) there is a positive constant α_0 such that, for $i = 1, \ldots, N+M$, the function $\hat{b}$ defined in (3.24) satisfies the relations

$$\min_{j \in \Lambda_i} |\hat{b}_{ij} \cdot n_{ij}| \geq \alpha_0 > 0, \tag{3.28a}$$

$$d_i m_i + \frac{1}{2} \int_{\Gamma_{ij}} \hat{b}_{ij} \cdot n_{ij} \, ds \geq 0. \tag{3.28b}$$

The first part of (3.27) corresponds to a discrete H^1 norm weighted by $h^{1/2}$ and the second part to a discrete L^2 norm. The assumption (3.28b) is closely related to (3.22). For, integrating (3.22) over the dual domain D_i and using the lumping operator, one obtains

$$d_i m_i + \frac{1}{2} \int_{\Gamma_{ij}} \hat{b}_{ij} \cdot n_{ij} \, ds \approx \int_{D_i} d \, dx + \frac{1}{2} \int_{\Gamma_{ij}} b_{ij} \cdot n_{ij} \, ds \geq 0.$$

To state the convergence result, we need another mesh-dependent seminorm on $(W^{1,\infty})^2$, namely

$$|g|_{1,\infty,h} := \left[\sum_{i=1}^{N+M} \sum_{j \in \Lambda_i} \mathrm{area}(B_{ij}) \, |g|^2_{1,\infty,B_{ij}} \right]^{1/2}.$$

Theorem 3.23. *Let $I_h u$ interpolate to the exact solution u of (3.21) in the nonconforming exponential fitted space V_h. Suppose that Assumption (A) holds true. Then for the discrete solution u_h of (3.25), one has the error estimate*

$$|||u_h - I_h u|||_{MW} \leq C h^{1/2} (|du|_1 + |b|_{1,\infty,h} \|u\|_\infty + |\varepsilon \nabla u - bu|_{1,\infty,h}). \tag{3.29}$$

Proof. See [MW94]. $\square$

Remark 3.24. The global error bound (3.29) does not imply that the method converges uniformly with respect to ε: in general $|du|_1$ and $|\varepsilon \nabla u - bu|_{1,\infty,h}$ are not bounded uniformly in ε. ♣

3.2 Residual-Based Stabilizations

This section discusses two stabilization techniques that add weighted residuals to the standard Galerkin finite element method: the *streamline diffusion* and *Galerkin least squares* finite element methods. These methods combine good global stability properties with high accuracy in subdomains that exclude boundary layers but they do not always preserve monotonicity. Since the residual of the exact solution is zero, the methods are automatically consistent in the finite element sense – i.e., the solution of the original boundary value problem also satisfies the discrete system of equations as in (3.33) below – unlike the monotonicity-preserving methods considered in the previous section. Furthermore, in Section 3.2.3 we shall use a multiscale framework to reveal that the streamline diffusion finite element method can be recovered from the standard Galerkin approach by taking into account the effect of the fine scales on the coarse scales.

3.2.1 Streamline Diffusion Finite Element Method (SDFEM)

The streamline diffusion finite element method (SDFEM) was introduced by Hughes and Brooks [HB79] for the numerical solution of convection-dominated convection-diffusion problems. Its manifestation in the simpler case of one space dimension was already studied in Section I.2.2.3. As we shall see, the SDFEM can be interpreted as a Petrov-Galerkin method, so it is also known as the *streamline upwind Petrov-Galerkin method* (SUPG method). For first-order hyperbolic problems (when $\varepsilon = 0$) similar ideas appeared in [Den74, Wah74]; these older methods can be considered precursors of the SDFEM.

We start with a weak formulation of the convection-diffusion problem (3.1):

Find $u \in V := H_0^1(\Omega)$ such that for all $v \in V$ one has

$$\varepsilon(\nabla u, \nabla v) + (b \cdot \nabla u, v) + (c\,u, v) = (f, v), \tag{3.30}$$

where $(\cdot, \cdot)$ is the $L_2(\Omega)$ inner product. Assume that b, c and f are sufficiently smooth with $c - \frac{1}{2}\nabla \cdot b > 0$.

Let $V_h \subset V$ be a conforming finite element space that consists of piecewise polynomials of degree k, i.e.,

$$V_h := \{v_h \in V \ : \ v_h\big|_T \in P_k(T) \quad \text{for all} \quad T \in \mathcal{T}_h\,\},$$

and assume that the triangulation $\mathcal{T}_h$ of Ω is shape-regular. Let $u \in H^{k+1}(T)$ with $k \geq 1$ so that $u \in C(\bar{\Omega})$. Then one can define its interpolant u^I in V_h and [Cia02, GRS07] this enjoys the approximation property

$$|u - u^I|_{m,T} \leq C\,h_T^{k+1-m}\,|u|_{k+1,T} \qquad \text{for } m = 0, 1, 2 \tag{3.31}$$

on each $T \in \mathcal{T}_h$. Moreover, using a scaling argument and the equivalence of norms in finite-dimensional spaces, one can prove the local inverse inequality

$$\|\Delta v_h\|_{0,T} \le \mu_{\mathrm{inv}} h_T^{-1} |v_h|_{1,T} \quad \forall v_h \in V_h, \tag{3.32}$$

where the constant μ_{inv} is independent of T and h.

The SDFEM adds weighted residuals to the usual Galerkin finite element method. Thus, assuming that the solution u of (3.30) is regular, in the sense that

$$-\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in} \quad L_2(T) \quad \forall T \in \mathcal{T}_h,$$

it follows that u satisfies

$$a_h(u, v_h) = f_h(v_h) \quad \forall v_h \in V_h, \tag{3.33}$$

where

$$a_h(w, v) := \varepsilon(\nabla w, \nabla v) + (b \cdot \nabla w, v) + (cw, v)$$
$$+ \sum_{T \in \mathcal{T}_h} \delta_T(-\varepsilon \Delta w + b \cdot \nabla w + cw, b \cdot \nabla v)_T, \tag{3.34}$$

$$f_h(v) := (f, v) + \sum_{T \in \mathcal{T}_h} \delta_T(f, b \cdot \nabla v)_T . \tag{3.35}$$

Here $(\cdot, \cdot)_T$ denotes the inner product in $L_2(T)$. The user-chosen constant δ_T is called the SD parameter. Since in general $\Delta u_h \notin L_2(\Omega)$ but $\Delta u_h \in L_2(T)$ for each T, we calculate Δu_h element by element.

Now the SDFEM is defined as follows:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a_h(u_h, v_h) = f_h(v_h). \tag{3.36}$$

The SDFEM satisfies (3.33) and so is consistent ; combining this with (3.30) yields the *projection property*

$$a_h(u - u_h, v_h) = 0 \quad \forall v_h \in V_h$$

for solutions $u \in H^2(\Omega)$ of (3.30). This identity is also known as *Galerkin orthogonality*.

It is natural to measure stability and errors in the following norm that is related to the discrete bilinear form a_h:

$$|||v|||_{SD} := \left(\varepsilon |v|_1^2 + \sum_{T \in \mathcal{T}_h} \delta_T \|b \cdot \nabla v\|_{0,T}^2 + \omega \|v\|_{0,T}^2 \right)^{1/2} .$$

Set $c_T = \max_{x \in T} |c(x)|$ for each $T \in \mathcal{T}_h$. For stability of the SDFEM, let the constant ω satisfy

$$c - \frac{1}{2}\nabla \cdot b \ge \omega > 0 \quad \text{on } \Omega.$$

The stability properties of the SDFEM are a consequence of

Lemma 3.25. *Let the SD parameter δ_T satisfy*

$$0 < \delta_T \leq \frac{1}{2} \min\left\{ \frac{\omega}{c_T^2}, \frac{h_T^2}{\varepsilon \mu_{\mathrm{inv}}^2} \right\}$$

for each $T \in \mathcal{T}_h$. Then the discrete bilinear form is coercive, i.e.,

$$a_h(v_h, v_h) \geq \frac{1}{2} |||v_h|||_{SD}^2 \qquad \forall v_h \in V_h.$$

Proof. For each $v_h \in V_h$, we have

$$a_h(v_h, v_h) \geq \varepsilon |v_h|_1^2 + \omega \|v_h\|_0^2 + \sum_{T \in \mathcal{T}_h} \delta_T \|b \cdot \nabla v_h\|_{0,T}^2$$

$$+ \sum_{T \in \mathcal{T}_h} \delta_T(-\varepsilon \Delta v_h + c v_h, b \cdot \nabla v_h)_T.$$

The local inverse inequality (3.32) and the hypothesis on δ_T give

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T(-\varepsilon \Delta v_h + c v_h, b \cdot \nabla v_h)_T \right|$$

$$\leq \sum_{T \in \mathcal{T}_h} \varepsilon^2 \delta_T \|\Delta v_h\|_{0,T}^2 + \sum_{T \in \mathcal{T}_h} c_T^2 \delta_T \|v_h\|_{0,T}^2 + \frac{1}{2} \sum_{T \in \mathcal{T}_h} \delta_T \|b \cdot \nabla v_h\|_{0,T}^2$$

$$\leq \frac{\varepsilon}{2} |v_h|_1^2 + \frac{\omega}{2} \|v_h\|_0^2 + \frac{1}{2} \sum_{T \in \mathcal{T}_h} \delta_T \|b \cdot \nabla v_h\|_{0,T}^2,$$

which yields the desired result.   □

Remark 3.26. Lemma 3.25 implies the *a priori* estimate

$$|||u_h|||_{SD} \leq C \left(\|f\|_0^2 + \sum_{T \in \mathcal{T}_h} \delta_T \|f\|_{0,T}^2 \right)^{1/2}$$

and the stability inequality

$$|||u_h|||_{SD} \leq 2 |||A_h u_h|||_*$$

for the discrete operator $A_h : V_h \to V_h^*$ defined by

$$< A_h v_h, w_h >:= a_h(v_h, w_h) \qquad \forall v_h, w_h \in V_h,$$

where the norm on the dual space V_h^* is

$$|||g_h|||_* := \sup_{w_h \in V_h} \frac{< g_h, w_h >}{|||w_h|||_{SD}} \quad \text{for } g_h \in V_h^*.$$

Compared with the standard Galerkin finite element method, where $\delta_T = 0$, the SDFEM provides additional control over the convective derivative in the streamline direction because of the definition of the norm $||| \cdot |||_{SD}$. This additional bound prevents the discrete solution from oscillating over a large part of Ω, as Schieweck [Sch08] explains.   ♣

To get an error estimate for the SDFEM, consider first the error between u_h and the interpolant $u^I \in V_h$ of the exact solution u. Lemma 3.25 and the projection property (3.40) give

$$\frac{1}{2}|||u^I - u_h|||_{SD}^2 \le a_h(u^I - u_h, u^I - u_h) = a_h(u^I - u, u^I - u_h).$$

We estimate the right-hand side term by term, invoking the interpolation properties (3.31) for smooth functions u in $V \cap H^{k+1}(\Omega)$:

$$\varepsilon(\nabla(u^I - u), \nabla(u^I - u_h)) \le \varepsilon^{1/2}|u^I - u|_1 \, |||u^I - u_h|||_{SD}$$
$$\le C\varepsilon^{1/2}h^k|u|_{k+1} \, |||u^I - u_h|||_{SD} ,$$

$$(b \cdot \nabla(u^I - u) + c(u^I - u), u^I - u_h)$$
$$= ((c - \nabla \cdot b)(u^I - u), u^I - u_h) - (u^I - u, b \cdot \nabla(u^I - u_h))$$
$$\le \left[C\left(\sum_{T \in \mathcal{T}_h} \|u^I - u\|_{0,T}^2 \right)^{1/2} + \left(\sum_{T \in \mathcal{T}_h} \delta_T^{-1}\|u^I - u\|_{0,T}^2 \right)^{1/2} \right] |||u^I - u_h|||_{SD}$$
$$\le C\,h^k \left[\sum_{T \in \mathcal{T}_h} h_T^2(1 + \delta_T^{-1})|u|_{k+1,T}^2 \right]^{1/2} |||u^I - u_h|||_{SD}$$

and

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T \big(-\varepsilon\Delta(u^I - u) + b \cdot \nabla(u^I - u) + c(u^I - u), b \cdot \nabla(u^I - u_h) \big)_T \right|$$
$$\le C \sum_{T \in \mathcal{T}_h} \delta_T^{1/2}(\varepsilon h_T^{k-1} + h_T^k + h_T^{k+1})|u|_{k+1,T}\delta_T^{1/2}\|b \cdot \nabla(u^I - u_h)\|_{0,T}$$
$$\le C \left[\sum_{T \in \mathcal{T}_h} (\varepsilon + \delta_T)h_T^{2k}|u|_{k+1,T}^2 \right]^{1/2} |||u^I - u_h|||_{SD}.$$

The last inequality appealed to the bound $\varepsilon\delta_T \le Ch_T^2$ that was assumed in Lemma 3.25. Combining all of these estimates, one gets

$$|||u^I - u_h|||_{SD} \le C \left[\sum_T (\varepsilon + \delta_T + \delta_T^{-1}h_T^2 + h_T^2)\, h_T^{2k} \, |u|_{k+1,T}^2 \right]^{1/2} . \qquad (3.37)$$

In order to extract the best possible convergence rate from (3.37), one must balance the terms ε, δ_T, and $\delta_T^{-1}h_T^2$ while respecting the constraints on δ_T in the hypotheses of Lemma 3.25. This balance is achieved by setting

$$\delta_T = \begin{cases} \delta_0 \, h_T & \text{if } Pe_T > 1 \text{ (convection-dominated case),} \\ \delta_1 \, h_T^2/\varepsilon & \text{if } Pe_T \le 1 \text{ (diffusion-dominated case),} \end{cases} \qquad (3.38)$$

with appropriate positive constants δ_0 and δ_1. Here and in what follows the *local mesh Péclet number* is defined by

$$Pe_T := \frac{\|b\|_{0,\infty,T}\, h_T}{2\varepsilon}\,.$$

Theorem 3.27. *Let the hypotheses of Lemma 3.25 be satisfied. Choose δ_T according to (3.38). Then the solution u_h of the SDFEM satisfies the global error estimate*

$$|||u - u_h|||_{SD} \le C\,(\varepsilon^{1/2} + h^{1/2})\, h^k |u|_{k+1}\,.$$

Proof. It follows from (3.37) and (3.38) that

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

Applying the interpolation properties (3.31), we get

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

A triangle inequality completes the proof. □

Remark 3.28. In the convection-dominated case one has $\varepsilon < \|b\|_{0,\infty,T}\, h_T/2$ and $\delta_T = \delta_0 h_T$. Hence on meshes satisfying $h \le C h_T$ one obtains the global estimate

$$\|u - u_h\|_0 + h^{1/2}\Big(\sum_T \|b \cdot \nabla(u - u_h)\|_{0,T}^2\Big)^{1/2} \le C\, h^{k+1/2} |u|_{k+1}.$$

As
$$\|u - u^I\|_0 \le C\, h^{k+1}|u|_{k+1} \qquad \text{and} \qquad |u - u^I|_1 \le C\, h^k |u|_{k+1},$$

we see that the L_2 error of the derivative in the streamline direction is optimal, but the bound on $\|u - u_h\|_0$ is order $1/2$ less than optimal.   ♣

In the special case where $b = (b_1, b_2)$ is constant, $c \equiv 0$ and V_h comprises piecewise linear elements, the bilinear form associated with the SDFEM can be written as

$$a_h(u_h, v_h) = \varepsilon(\nabla u_h, \nabla v_h) + \sum_{T \in \mathcal{T}_h} \delta_T |b|^2 \left(\frac{\partial u_h}{\partial b}, \frac{\partial v_h}{\partial b}\right)_T + (b \cdot \nabla u_h, v_h).$$

Thus the SDFEM adds artificial diffusion of $\mathcal{O}(\delta_T |b|^2)$ in the direction of the streamline – this explains why it is called a streamline diffusion method.

The SDFEM can also be regarded as a Petrov-Galerkin method, for choosing the test functions on each triangle T to be $v + \delta_T b \cdot \nabla v$ gives (3.34)–(3.36).

Remark 3.29. In the special case of piecewise linear elements, the expression $\sum_T \varepsilon \delta_T (\Delta u_h, b \cdot \nabla v_h)_T$ contributes nothing to $a_h(u_h, v_h)$. The upper bound for the SD parameter δ_T in Lemma 3.25 can then be relaxed to

$$0 < \delta_T \leq \frac{\omega}{c_T^2} .$$

This upper bound is independent of the discretization constant μ_{inv} of (3.32). ♣

Remark 3.30. When choosing δ_T for second-order or higher-order elements, one has to take into consideration the constant μ_{inv} from the local inverse inequality (3.32). The value of μ_{inv} depends on the mesh and on the finite element space V_h used; estimates of this constant are given in [HH92] and its numerical computation is investigated in [Fra94]. ♣

Remark 3.31. Optimal-order error bounds with respect to the mesh size h and the polynomial order $p = k$ of the finite element space V_h have been given in [HS01b]. In this hp-type analysis the inverse inequality (3.32), the range of δ_T in Lemma 3.25, and its choice in (3.38) have to be replaced by

$$\|\Delta v_h\|_{0,T} \leq \mu_{\mathrm{inv}} p^2 h_T^{-1} |v_h|_{1,T} \quad \forall v_h \in V_h,$$

$$0 < \delta_T \leq \frac{1}{2} \min \left\{ \frac{\omega}{c_T^2}, \frac{h_T^2}{\varepsilon \mu_{\mathrm{inv}}^2 p^4} \right\},$$

$$\delta_T = \begin{cases} \dfrac{\delta_0\, h_T}{p} & \text{if } Pe_T > 1, \\[2ex] \dfrac{1}{2} \min \left\{ \dfrac{\omega}{c_T^2}, \dfrac{h_T^2}{\varepsilon \mu_{\mathrm{inv}}^2 p^4} \right\} & \text{if } Pe_T \leq 1. \end{cases}$$

♣

Remark 3.32. Can one make an optimal choice of the SD parameter δ_T inside the range assumed in Theorem 3.27? In general the answer is unknown. If piecewise linear finite elements are used for a one-dimensional problem with constant coefficients, then the optimal value of δ_T can be computed and produces the Il'in-Allen-Southwell scheme (see Section I.2.1.3), which yields the exact solution at the nodes. For two or more dimensions some fresh criterion is needed to fix δ_T. One possibility is to try to fulfil some necessary conditions for convergence, uniformly with respect to ε, in a certain norm [Roo85, SE99, ST95]. An alternative is the reduction of the numerical viscosity of the scheme [Tob95]. The reduced solution outside layers has been taken into account in [Kno, MS96] when trying to optimize δ_T. A symbiotic relationship between "best" solution approximation and fast convergence of smoothers based on the standard GMRES iteration has been reported in [FRSW99]. ♣

Numerical results comparing the SDFEM and its modifications with other methods can be found in [JK07b]; see also [LS01b] and Remark 3.119 for a related discussion on layer-adapted meshes.

Example 3.33. We write down the difference scheme generated by the SDFEM for the model problem

$$-\varepsilon \Delta u + 2u_x + u_y = f \quad \text{in } \Omega = (0,1)^2, \quad u = 0 \text{ on } \partial\Omega,$$

using piecewise linear elements on a uniform square mesh of Friedrichs-Keller type. The distance between adjacent nodes is denoted by h. After scaling, one gets the difference stencil

$$\frac{\varepsilon}{h^2}
\begin{bmatrix} \cdot & -1 & \cdot \\ -1 & 4 & -1 \\ \cdot & -1 & \cdot \end{bmatrix}
+ \frac{1}{2h}
\begin{bmatrix} \cdot & \cdot & 1 \\ -1 & \cdot & 1 \\ -1 & \cdot & \cdot \end{bmatrix}
+ \frac{\delta_0}{h}
\begin{bmatrix} \cdot & 1 & -2 \\ -2 & 6 & -2 \\ -2 & 1 & \cdot \end{bmatrix},$$

where in accordance with the convection-dominated case $\varepsilon \ll h$ we have set $\delta_T = \delta_0 h$ with a user-chosen constant δ_0 as in (3.38).

One cannot apply Theorem 3.1 to this scheme because in general positive off-diagonal terms appear, so the sufficient conditions for a discrete maximum principle are not satisfied. In fact, numerical calculations show that the discrete maximum principle does indeed fail for the SDFEM: oscillations can be observed near sharp layers – see for example [JK07b].   ♣

A *nonlinear* modification of the SDFEM that satisfies the discrete maximum principle for meshes of weakly acute type has been proposed by Mizukami and Hughes [MH85]. Suppose that $b = (b_1, b_2)$ is piecewise constant and $c \equiv 0$. We use piecewise linear elements and denote by φ_i the usual basis function associated with the node p_i. Assume that p_1, p_2, and p_3 are the vertices of the triangle $T \in \mathcal{T}_h$. The contributions from T to the convective part of the matrix $a_h(\varphi_j, \varphi_i)$ are given by

$$e_{ij} := (b \cdot \nabla\varphi_j, \varphi_i + \delta_T b \cdot \nabla\varphi_i)_T \quad \text{for } i, j = 1, 2, 3.$$

For each $i \in \{1, 2, 3\}$, Mizukami and Hughes replace b and $\delta_T b \cdot \nabla\varphi_i$ by $\tilde{b}_i$ and M_i respectively; these constants are such that the modified contributions

$$\tilde{e}_{ij} := (\tilde{b}_i \cdot \nabla\varphi_j, \varphi_i + M_i)_T \quad \text{for } i, j = 1, 2, 3$$

satisfy $\tilde{e}_{ij} \leq 0$ for $i \neq j$. (The element entries of the SDFEM can be recovered by setting $\tilde{b}_i = b$ and $M_i = \delta_T b \cdot \nabla\varphi_i$ for $i = 1, 2, 3$.)

We now examine the choice of $\tilde{b}_i$ and M_i. Consider first M_i. From the representation

$$\tilde{e}_{ij} = b \cdot \nabla\varphi_j\big|_T \int_T (\varphi_i + M_i)\, dx = b \cdot \nabla\varphi_j\big|_T \left(\frac{1}{3} + M_i \right) |T| \qquad (3.39)$$

it follows that that the sign of $\tilde{e}_{ij}$ is determined by $b \cdot \nabla\varphi_j|_T$ if

$$M_i \geq -1/3 \quad \text{for } i = 1, 2, 3. \tag{3.40}$$

A second requirement when selecting M_i is the fulfillment of a discrete conservation law [MH85]. Consequently we look for constants M_i (for $i = 1, 2, 3$) such that (3.40) and

$$M_1 + M_2 + M_3 = 0 \tag{3.41}$$

are satisfied. The different cases that can arise will now be discussed.

If the orientation of the triangle T with respect to the flow direction b is such that the term $(b \cdot \nabla\varphi_j)$ is positive for a single value of j, say $j = 1$, and non-negative for $j = 2$ and $j = 3$, then it suffices to choose $M_2 = M_3 = -1/3$ and no replacement of b is needed. This guarantees that $\tilde{e}_{i1} = 0$ for $i = 2, 3$. To fulfil (3.41) one must choose $M_1 = 2/3$.

If instead two of the terms $(b \cdot \nabla\varphi_j)$, where $j = 1, 2, 3$, are positive, then for definiteness let $(b \cdot \nabla\varphi_j) > 0$ for $j = 2, 3$. In this case one cannot choose the M_i so that $\tilde{e}_{ij} \leq 0$ for $i \neq j$. Nevertheless we can profit from the observation that

$$\tilde{b}_i \cdot \nabla u_h = b \cdot \nabla u_h$$

whenever $b - \tilde{b}$ is perpendicular to ∇u_h. Thus for $i = 1, 2, 3$ replace b by $\tilde{b}_i$ in (3.39), where

$$\tilde{b}_1 = b, \quad \tilde{b}_2 = b + w_2 \quad \text{and} \quad \tilde{b}_3 = b + w_3$$

for some as yet unspecified $w_2, w_3 \perp \nabla u_h$. A careful analysis shows that it is possible to find $w_2, w_3 \perp \nabla u_h$ such that at least one set of inequalities

$$\tilde{b}_2 \cdot \nabla\varphi_1 < 0, \quad \tilde{b}_2 \cdot \nabla\varphi_2 > 0, \quad \tilde{b}_2 \cdot \nabla\varphi_3 < 0 \tag{3.42a}$$

and

$$\tilde{b}_3 \cdot \nabla\varphi_1 < 0, \quad \tilde{b}_3 \cdot \nabla\varphi_2 < 0, \quad \tilde{b}_2 \cdot \nabla\varphi_3 > 0 \tag{3.42b}$$

holds true. If (3.42a) is true, set $M_1 = M_3 = -1/3$ and $M_2 = 2/3$ to get non-negative off-diagonals. If (3.42b) holds true, choose $M_1 = M_2 = -1/3$ and $M_3 = 2/3$. If one can find w_2 and w_3 satisfying both (3.42a) and (3.42b), then set $M_1 = -1/3$ and choose $M_2 > -1/3$ and $M_3 > -1/3$ with $M_2 + M_3 = 1/3$; only in this case, by taking $\delta_T = (3|b \cdot \nabla\varphi_1|)^{-1}$, $M_2 = \delta_T b \cdot \nabla\varphi_2$ and $M_3 = \delta_T b \cdot \nabla\varphi_3$, does one recover the original SDFEM. Note that the choice of $\tilde{b}$ depends on ∇u_h which is *a priori* unknown, so the choice of the constants M_i, for $i = 1, 2, 3$, depends also on ∇u_h, thereby generating a nonlinear system of equations.

Each contribution $\tilde{e}_{ij}$ from each triangle T gives a matrix with non-positive off-diagonal entries, so the global matrix also has this property. As the diffusion matrix for piecewise linear elements is an M-matrix, it follows that the coefficient matrix for the discrete problem will be an M-matrix.

Remark 3.34. Numerical experiments in [MH85] show that the method gives accurate solutions with little crosswind diffusion, but in some cases the method does not give satisfactory results [Kno06] since a small change of b can change drastically the constants M_i for $i = 1, 2, 3$. Improvements of the method in layer regions are suggested in [Kno06] where extensions to convection-diffusion-reaction equations and three-dimensional problems are also considered. But the method remains nonlinear, even in the constant coefficient case, since the matrix of the difference scheme generated depends on ∇u_h.   ♣

Lemma 3.25 shows that the SDFEM (3.34)–(3.36) has improved stability properties when compared with the standard Galerkin method. Its theoretical convergence rate for the global L_2 error is order $1/2$ less than optimal for smooth solutions; see Remark 3.28. We now give an optimal L_2-convergence result for a special mesh and sufficiently regular solutions.

To concentrate on the essential features of the argument, consider as a model problem the case $b \equiv (1, 0)$, $c \equiv 1$, viz., the boundary value problem

$$-\varepsilon \Delta u + u_x + u = f \quad \text{in } \Omega = (0,1)^2, \qquad u\big|_{\partial\Omega} = 0, \tag{3.43}$$

and restrict the discussion to the case of piecewise linear approximations, i.e.,

$$V_h = \{v_h \in H_0^1(\Omega) : v_h\big|_T \in P_1(T) \quad \text{for all} \quad T \in \mathcal{T}_h, \ v_h\big|_{\partial\Omega} = 0\}.$$

The interesting case $0 < \varepsilon \leq h$ will be examined here, so interior and boundary layers are not resolved; we aim for a numerical method in which these layers do not pollute regions where the solution is smooth. With these assumptions, the SDFEM of (3.34)–(3.36) is:

Find $u^h \in V_h$ such that for all $v^h \in V_h$ one has

$$\varepsilon(\nabla u^h, \nabla v^h) + (u_x^h, v^h) + (u^h, v^h) + \sum_{T \in \mathcal{T}_h} \delta_T(u_x^h + u^h, v_x^h)_T$$

$$= (f, v^h) + \sum_{T \in \mathcal{T}_h} \delta_T(f, v_x^h)_T. \tag{3.44}$$

Consider now a more general method of streamline-diffusion type that adds *artificial crosswind diffusion* [JSW87, Nii90, Zho95, ZR96]. That is, consider the following numerical method for solving (3.43):

Find $u^h \in V_h$ such that for all $v^h \in V_h$ one has

$$(\varepsilon + \delta)(u_x^h, v_x^h) + \varepsilon_m(u_y^h, v_y^h + (1 - \delta)(u_x^h, v^h) + (u^h, v^h))$$

$$= (f, v^h + \delta v_x^h), \tag{3.45}$$

where the artificial crosswind diffusion ε_m is as yet unspecified. Clearly (3.45) can be derived from (3.44) by setting $\delta_T = \delta$ for all $T \in \mathcal{T}_h$, changing the crosswind diffusion from ε to ε_m, and integrating by parts the term $\delta(u^h, v_x^h)$.

In what follows we assume that δ and ε_m are positive. To analyse (3.45), introduce the bilinear form

$$b_h(w, v) := (\varepsilon + \delta)(w_x, v_x) + \varepsilon_m(w_y, v_y) + (1 - \delta)(w_x, v) + (w, v),$$

the linear form

$$l_h(v) := (f, v + \delta v_x),$$

and the mesh-dependent norm

$$|||v|||_{ACD} := \left[(\varepsilon + \delta)\|v_x\|_0^2 + \varepsilon_m\|v_y\|_0^2 + \|v\|_0^2\right]^{1/2}. \qquad (3.46)$$

The discrete problem (3.45) can be rewritten as

Find $u^h \in V_h$ such that for all $v^h \in V_h$ one has

$$b_h(u^h, v^h) = l_h(v^h).$$

Since $b_h(v^h, v^h) = |||v^h|||_{ACD}^2$, the Lax-Milgram lemma ensures that (3.45) has a unique solution. If the exact solution of (3.43) lies in $H_0^1(\Omega) \cap H^2(\Omega)$, one obtains the quasi-orthogonality relation

$$b_h(u - u^h, v^h) = \text{Per}(u, v^h) \quad \forall v^h \in V_h \qquad (3.47a)$$

with the perturbation term

$$\text{Per}(u, v^h) := (\varepsilon \Delta u, \delta v_x^h) + (\varepsilon_m - \varepsilon)(u_y, v_y^h). \qquad (3.47b)$$

First we derive a global error estimate for the artificial crosswind diffusion method. Start from the triangle inequality

$$|||u - u^h|||_{ACD} \le |||u - u^I|||_{ACD} + |||u^I - u^h|||_{ACD},$$

where u^I denotes the interpolant from V_h to the exact solution u. Here

$$|||u - u^I|||_{ACD} \le C\left[(\varepsilon + \delta)^{1/2}h + \varepsilon_m^{1/2}h + h^2\right]|u|_2.$$

For the second term, the coercivity of b_h and the quasi-orthogonality relation (3.47a) yield

$$
\begin{aligned}
|||u^I &- u^h|||_{ACD}^2 \\
&= b_h(u^I - u, u^I - u^h) + \text{Per}(u, u^I - u^h) \\
&\le C\left[(\varepsilon + \delta)^{1/2}h + \varepsilon_m^{1/2}h + \delta^{-1/2}h^2 + h^2\right]|u|_2 \, |||u^I - u^h|||_{ACD} \\
&\quad + C\left(\varepsilon\delta^{1/2}\|\Delta u\|_0 + |\varepsilon_m - \varepsilon| \, \|u_{yy}\|_0\right) |||u^I - u^h|||_{ACD}.
\end{aligned}
$$

Combining these bounds and minimizing the resulting right-hand side with respect to δ yields $\delta \sim h$; using $0 < \varepsilon \le h$ one then gets

$$|||u - u^h|||_{ACD} \le C\big(h^{3/2} + \varepsilon_m^{1/2}h + h^2 + |\varepsilon_m - \varepsilon|\big)|u|_2.$$

Hence

$$|||u - u_h|||_{ACD} \le Ch^{3/2}|u|_2,$$

provided that the added crosswind diffusion $|\varepsilon_m - \varepsilon|$ is of $\mathcal{O}(h^{3/2})$.

We see that the L_2-norm convergence rate is $\mathcal{O}(h^{3/2})$ for the modified SDFEM (3.45), as for the standard SDFEM (3.34)–(3.36). In general this result cannot be sharpened (Remark 3.44), but on certain structured meshes the modified SDFEM (3.45) will yield second-order convergence in L_2. We now prove this for the model problem (3.43) on a three-directional mesh. A triangle inequality and (3.47a) yield

$$\|u - u^h\|_0 \le \|u - u^I\|_0 + \|u^I - u^h\|_0$$

$$\le Ch^2 + \frac{b_h(u^I - u, u^I - u^h)}{|||u^I - u^h|||_{ACD}} + \frac{\mathrm{Per}(u, u^I - u^h)}{|||u^I - u^h|||_{ACD}}.$$

For an improved error estimate, a sharper bound on the approximation error term $b_h(\cdot, \cdot)$ is needed. The technique that was applied earlier does not benefit from any possible interaction of error terms from adjacent triangles, but it will be shown below that a useful cancellation of low-order error terms occurs.

In order to give a detailed formula for the local interpolation error, we introduce some notation for an arbitrary fixed triangle T. For $i = 1, 2, 3$, let p_i denote the vertices of T in anti-clockwise ordering, let the side opposite p_i be S_i, let $h_i = \lambda_i h$ be the length of S_i, write $n^i = (n_x^i, n_y^i)$ for the outer normal unit vector along S_i, denote the directional derivative along S_i (in the anti-clockwise direction) by D_i and let $|T|$ be the area of T.

Lemma 3.35. *Let u^I be the linear nodal interpolant to the function u on a triangular element T and let $w \in V_h$. Then the following error expansions hold true:*

$$\int_T (u - u^I)_\mu w_\nu \, dxdy = \mathcal{O}\big(h^2(\|u\|_{4,T} + h\|u\|_{5,T})\,\|w\|_{0,T}\big)$$

$$+ \frac{h^4}{24|T|} \sum_{i=1}^3 \int_{S_i} \big(\lambda_i^3 \lambda_{i+1} n_\mu^i n_\nu^i D_{i+1} D_i^2 u - \lambda_{i+2}^4 n_\mu^{i+2} n_\nu^{i+2} D_i D_{i+2}^2 u\big) w \, ds$$

$$+ \frac{h^4}{24|T|} \sum_{i=1}^3 \int_{S_i} \big(\lambda_i^3 \lambda_{i+2} n_\mu^i n_\nu^{i+2} D_i^2 u - \lambda_{i+1}^4 n_\mu^{i+1} n_\nu^{i+1} D_{i+1}^2 u\big) D_i w \, ds, \quad (3.48)$$

$$\int_T (u - u^I) w_x \, dxdy = -\frac{h^2}{24} \sum_{i=1}^3 \int_{S_i} \Big(\sum_{j=1}^3 \lambda_j^2 D_j^2 u\Big) n_x^i w \, ds$$

$$+ \mathcal{O}(h^2 \|u\|_{3,T} \, \|w\|_{0,T}), \quad (3.49)$$

where μ and ν may be x or y and the indices $i + 1$, $i + 2$ are used modulo 3.

Proof. This uses techniques that are expounded in [BLR86, ZL94, Zho97]. □

We resume our examination of the approximation error for a three-directional mesh. Let $S_1 = S_1'$ be a common side of the two triangles T and T'. On a three-directional mesh one has $\lambda_i = \lambda_i'$, $D_i = -D_i'$, $|T| = |T'|$ and $n^i = -(n')^i$. Consequently all line integrals over interior sides cancel when summing (3.48) and (3.49) over all triangles T. Moreover, all line integrals over the boundary vanish also, because $w \equiv D_i w \equiv 0$ on $S_i \subset \partial\Omega$. In this way, for each $w_h \in V_h$ one has

$$b_h(u^I - u, w^h) = (\varepsilon + \delta)((u^I - u)_x, w_x^h) + \varepsilon_m((u^I - u)_y, w_y^h)$$
$$- (1 - \delta)((u^I - u), w_x^h) + ((u^I - u), w^h)$$
$$\leq C(\varepsilon + \delta + \varepsilon_m + 1)\, h^2 \, \|u\|_5 \|w^h\|_0.$$

For the perturbation term,

$$|\mathrm{Per}(u, w^h)| \leq (\varepsilon\delta\|\Delta u_x\|_0 + |\varepsilon_m - \varepsilon|\, \|u_{yy}\|_0)\|w^h\|_0.$$

Taking $w^h = u^I - u^h$ above, one now gets

$$\|u - u^h\|_0 \leq \|u - u^I\|_0 + |||u^I - u^h|||_{ACD}$$
$$\leq Ch^2|u|_2 + \frac{b_h(u^I - u, u^I - u^h)}{|||u^I - u^h|||} + \frac{\mathrm{Per}(u, u^I - u^h)}{|||u^I - u^h|||_{ACD}}$$
$$\leq Ch^2(\varepsilon + \delta + \varepsilon_m + 1)\|u\|_5 + C\,(\varepsilon\delta + |\varepsilon_m - \varepsilon|)\|u\|_5.$$

We have proved

Theorem 3.36. *Let the solution u of (3.43) belong to $H_0^1(\Omega) \cap H^5(\Omega)$. Suppose that $0 < \varepsilon < \varepsilon_m \leq Ch^2$ for some positive (generic) constant C. Let u^h be the solution of the modified SDFEM (3.45). Then on a three-directional mesh one has the error estimate*

$$\|u - u^h\|_0 \leq Ch^2\|u\|_5. \tag{3.50}$$

Remark 3.37. No lower bounds on δ and ε_m are needed for the global error estimate (3.50) but it will emerge later that lower bounds on these parameters are used in estimating the local L_2 and pointwise errors.   ♣

Remark 3.38. Optimal global L_2-error estimates for bilinear finite elements on structured meshes have been considered in [Zho97]. An alternative technique for proving optimal global L_2-error estimates for linear and bilinear finite elements on special meshes under weaker regularity assumptions can be found in [Näv82].   ♣

All error estimates so far are meaningful only for smooth solutions, i.e., for solutions where the norm $|u|_{k+1}$ is of moderate size. This norm will be large

if boundary or interior layers are present in the solution u. We therefore turn to the investigation of local errors.

In subdomains that exclude layers, uniform local error estimates will be derived. It follows that the SDFEM is able to identify the layer regions. Asymptotic analysis tells us that boundary layers will appear in the solution of (3.43) along the sides $x = 1$, $y = 0$ and $y = 1$. It is therefore natural to consider the local error in the fixed subdomain

$$\Omega' = \{ (x, y) \in \Omega : 0 < x < x_1 < 1, \ 0 < y_1 < y < y_2 < 1 \}$$

where x_1, y_1 and y_2 are some fixed constants. To simplify the presentation assume that the boundary of Ω' coincides with lines of the mesh. Introduce the cut-off function

$$\varphi(x, y) := \exp\left(-d\left(\frac{x - x_1}{\sigma_x}\right) \right) \exp\left(-d\left(\frac{y_1 - y}{\sigma_y}\right) \right) \exp\left(-d\left(\frac{y - y_2}{\sigma_y}\right) \right),$$
$$(3.51)$$

where $d : \mathbb{R} \to \mathbb{R}$ is defined by $d(t) := \max\{0, t\}$, while σ_x and σ_y are positive parameters that will be chosen later. The derivation of local error estimates uses the coercivity of the bilinear form, some approximation theory properties and the stability of the interpolant with respect to the weighted discrete norm

$$|||w^h|||_\varphi := \left[(\varepsilon + \delta)\|w_x^h\|_\varphi^2 + \varepsilon_m \|w_y^h\|_\varphi^2 + \frac{1 - \delta}{2} \left\| \sqrt{|\varphi_x|} \, w^h \right\|_0^2 + \|w^h\|_\varphi^2 \right]^{1/2},$$

where we recall that $0 < \delta < 1$. Here $\| \cdot \|_\varphi$ denotes the φ-weighted L_2 norm defined by

$$\|w\|_\varphi^2 := \int_\Omega \varphi w^2 \, dx.$$

First we prove the coercivity of the bilinear form with respect to $||| \cdot |||_\varphi$.

Lemma 3.39. *Let $0 < \delta \leq 1/2$. Define σ_x and σ_y in (3.51) by $\sigma_x = (\varepsilon + \delta)M$ and $\sigma_y = M\sqrt{\varepsilon_m}$, where $M \geq 4$ is constant. Then*

$$b_h(w^h, \varphi w^h) \geq \frac{1}{2}|||w^h|||_\varphi^2 \qquad \text{for all } w^h \in V_h. \qquad (3.52)$$

Proof. The definition of b_h gives

$$b_h(w^h, \varphi w^h) = |||w^h|||_\varphi^2 - \frac{1 - \delta}{2} \left\| \sqrt{|\varphi_x|} \, w^h \right\|_0^2 + (1 - \delta)(w_x^h, \varphi w^h)$$
$$+ (\varepsilon + \delta)(w_x^h, \varphi_x \, w^h) + \varepsilon_m(w_y^h, \varphi_y \, w^h).$$

Integrating by parts and invoking $\varphi_x \leq 0$, the second and third terms cancel:

$$(w_x^h, \varphi w^h) = \frac{1}{2}\left(-\varphi_x, (w^h)^2 \right) = \frac{1}{2} \left\| \sqrt{|\varphi_x|} \, w^h \right\|_0^2.$$

The last two terms can be absorbed into the others, since

$$|(\varepsilon + \delta)(w_x^h, \varphi_x\, w^h)| \le \frac{\varepsilon + \delta}{\sigma_x^{1/2}} \|w_x^h\|_\varphi \, \big\| \sqrt{|\varphi_x|}\, w^h \big\|_0$$
$$\le \frac{2(\varepsilon + \delta)}{M} \|w_x^h\|_\varphi^2 + \frac{1 - \delta}{4} \big\| \sqrt{|\varphi_x|}\, w^h \big\|_0^2$$

as $\delta \le 1/2$, and

$$|\varepsilon_m(w_y^h, \varphi_y\, w^h)| \le \frac{\varepsilon_m}{\sigma_y} \|w_y^h\|_\varphi \|w^h\|_\varphi \le \frac{\varepsilon_m}{2M^2} \|w_y^h\|_\varphi^2 + \frac{1}{2}\|w^h\|_\varphi^2.$$

Taking $M \ge 4$, we finally obtain (3.52).   □

The next lemma gives an approximation theory estimate. For ease of reading we shall use the notation $I_h(gh)$ instead of $(gh)_I$ for the nodal interpolant to the product of two functions g and h.

Lemma 3.40. *Let the parameters σ_x and σ_y of the cut-off function φ be as in Lemma 3.39. Define the streamline-diffusion parameter δ and the artificial crosswind diffusion parameter ε_m by $\delta = Ch$ and $\varepsilon_m = Ch^{3/2}$. Then for each $\theta \in (0,1)$, one can choose a sufficiently large $M \ge 4$ in the definition of σ_x and σ_y such that*

$$|b_h(w^h, \varphi w^h - I_h(\varphi w^h))| \le \theta |||w^h|||_\varphi^2 \qquad \text{for all } w^h \in V_h.$$

Proof. Set $E = \varphi w^h - I_h(\varphi w^h)$. A detailed analysis is given only for the term (w_x^h, E) as the other terms in $b_h(w^h, E)$ are handled similarly. First, the interpolation property of I_h yields

$$(w_x^h, E) = \sum_{T \in \mathcal{T}_h} (w_x^h, E)_T$$
$$\le \sum_{T \in \mathcal{T}_h} \frac{1}{\min\limits_{z \in T} \varphi^{1/2}(z)} \|\varphi^{1/2} w_x^h\|_{0,T} \|E\|_{0,T}$$
$$\le \frac{\theta(\varepsilon + \delta)}{2} \|w_x^h\|_\varphi^2 + C \sum_{T \in \mathcal{T}_h} \frac{h^4}{(\varepsilon + \delta)\min\limits_{z \in T} \varphi(z)} |\varphi w^h|_{2,T}^2. \qquad (3.53)$$

Now on each triangle T,

$$(\varphi w^h)_{xx} = \varphi_{xx} w^h + 2\varphi_x w_x^h, \qquad |(\varphi w^h)_{xx}|^2 \le C\varphi \left(\frac{|\varphi_x|}{\sigma_x^3} |w^h|^2 + \frac{\varphi}{\sigma_x^2} |w_x^h|^2 \right).$$

Consequently, on each T,

$$\int_T \frac{h^4}{(\varepsilon + \delta) \min\limits_{z \in T} \varphi(z)} |(\varphi w^h)_{xx}|^2 \, dx \, dy$$

$$\leq C \frac{h^4}{\sigma_x^2 (\varepsilon + \delta)^2} \left(\frac{\max\limits_{z \in T} \varphi}{\min\limits_{z \in T} \varphi} \right) \left(\frac{\varepsilon + \delta}{\sigma_x} \| \sqrt{|\varphi_x|} w^h \|_{0,T}^2 + (\varepsilon + \delta) \| w_x^h \|_{\varphi,T}^2 \right)$$

$$\leq \frac{C}{M^2} \left(\| \sqrt{|\varphi_x|} \, w^h \|_{0,T}^2 + (\varepsilon + \delta) \| w_x^h \|_{\varphi,T}^2 \right).$$

Here we used $\sigma_x = (\varepsilon + \delta)M$, $\delta = Ch$ and $\max_{z \in T} \varphi(z) / \min_{z \in T} \varphi(z) \leq C$, where C is independent of T. Analogously, one gets

$$(\varphi w^h)_{yy} = \varphi_{yy} w^h + 2 \varphi_y w_y^h, \qquad |(\varphi w^h)_{yy}|^2 \leq C \varphi \left(\frac{\varphi}{\sigma_y^4} |w^h|^2 + \frac{\varphi}{\sigma_y^2} |w_y^h|^2 \right),$$

and on each T,

$$\int_T \frac{h^4}{(\varepsilon + \delta) \min\limits_{z \in T} \varphi(z)} |(\varphi w^h)_{yy}|^2 \, dx \, dy$$

$$\leq C \frac{h^4}{\delta \sigma_y^4} \left(\| w^h \|_{\varphi,T}^2 + \sigma_y^2 \| w_y^h \|_{\varphi,T}^2 \right)$$

$$\leq C \frac{h^3}{M^2 \varepsilon_m^2} \left(\frac{1}{M^2} \| w^h \|_{\varphi,T}^2 + \varepsilon_m \| w_y^h \|_{\varphi,T}^2 \right).$$

To continue, recall that $\varepsilon_m = Ch^{3/2}$ and obtain

$$\int_T \frac{h^4}{(\varepsilon + \delta) \min\limits_{z \in T} \varphi(z)} |(\varphi w^h)_{yy}|^2 \, dx \, dy \leq \frac{C}{M^2} \left(\| w^h \|_{\varphi,T}^2 + \varepsilon_m \| w_y^h \|_{\varphi,T}^2 \right).$$

The mixed derivative $(\varphi w^h)_{xy}$ can be estimated using the same ideas. Summing, one obtains an estimate for the final term in (3.53). One then has

$$|(w_x^h, E)| \leq \frac{\theta(\varepsilon + \delta)}{2} \| w_x^h \|_\varphi^2 + \frac{C}{M^2} ||| w^h |||_\varphi^2 \leq \theta ||| w^h |||_\varphi^2$$

for sufficiently large M. Analogous estimates can be derived for the other terms in $b_h(w^h, E)$. $\square$

We wish to describe a local L_2-error estimate for the modified SDFEM in the convection-dominated case $\varepsilon \ll h$. Recall that $\Omega' = (0, x_1) \times (y_1, y_2)$, where $0 < x_1 < 1$ and $0 < y_1 < y_2 < 1$. Define the enlargement Ω'' of Ω' to be the union of all triangles that lie entirely in the set

$$\Big\{ (x, y) \in \Omega : x \leq x_1 + K \sigma_x | \log h |,$$

$$y_1 - K \sigma_y | \log h | \leq y \leq y_2 + K \sigma_y | \log h | \Big\}, \tag{3.54}$$

where the constant K is specified later. In the complement $\Omega_c := \Omega \backslash \Omega''$, the cut-off function (3.51) is exponentially small, i.e., $\varphi \le Ch^\kappa$, where a suitably large positive value of κ can be achieved by choosing K large enough.

Theorem 3.41. *Choose the constant K in (3.54) so that $\kappa > 3/2$. Suppose that $u \in L_\infty(\Omega)$ and that for this fixed enlargement Ω'' the solution u of (3.43) satisfies $u \in H^2(\Omega'')$. Take $\delta = C_0 h$ and choose the artificial crosswind diffusion ε_m to satisfy $\varepsilon \le \varepsilon_m = C_1 h^{3/2}$, with arbitrary but fixed positive constants C_0 and C_1. Then there is a positive constant C such that the solution u^h of (3.45) satisfies the local error estimate*

$$\|u - u^h\|_{0,\Omega'} \le Ch^{3/2}\|u\|_{2,\Omega''}. \tag{3.55}$$

Proof. We start with the usual splitting

$$\begin{aligned}
\|u - u^h\|_{0,\Omega'} &\le \|u - u^I\|_{0,\Omega'} + \|u^I - u^h\|_{0,\Omega'} \\
&\le Ch^2|u|_{2,\Omega'} + \|u^I - u^h\|_\varphi \\
&\le Ch^2|u|_{2,\Omega'} + |||u^I - u^h|||_\varphi. \tag{3.56}
\end{aligned}$$

To estimate $w^h := u^I - u^h$, appeal to Lemmas 3.39 and 3.40 and the quasi-orthogonality relation (3.47a), obtaining

$$\begin{aligned}
\frac{1}{2}|||w^h|||_\varphi^2 &\le b_h(w^h, \varphi w^h) \\
&= b_h(w^h, \varphi w^h - I_h(\varphi w^h)) + b_h(w^h, I_h(\varphi w^h)) \\
&\le \theta|||w^h|||_\varphi^2 + b_h(u^I - u, I_h(\varphi w^h)) + b_h(u - u^h, I_h(\varphi w^h)),
\end{aligned}$$

so

$$\left(\frac{1}{2} - \theta\right)|||w^h|||_\varphi^2 \le b_h(u^I - u, I_h(\varphi w^h)) + \text{Per}(u, I_h(\varphi w^h)). \tag{3.57}$$

As in the proof of Lemma 3.40, one can establish the estimate

$$\begin{aligned}
(\varepsilon + \delta)\|\varphi^{-1/2}(I_h(\varphi w^h))_x\|_0^2 &+ \varepsilon_m\|\varphi^{-1/2}(I_h(\varphi w^h))_y\|_0^2 \\
&+ \|\varphi^{-1/2}I_h(\varphi w^h)\|_0^2 \le 6|||w^h|||_\varphi^2.
\end{aligned}$$

Next, each term in (3.57) is estimated by considering separately the two subdomains Ω'' and Ω_c, invoking interpolation properties in Ω'' and the exponential smallness of the cut-off function φ in Ω_c. Thus, taking for example the term

$$((u - u^I)_x, I_h(\varphi w^h)) = (u^I - u, (I_h(\varphi w^h))_x),$$

in Ω'' one has

$$\begin{aligned}
|(u^I - u, (I_h(\varphi w^h))_x)_{\Omega''}| &\le \theta\delta\|(I_h(\varphi w^h))_x\|_0^2 + C\delta^{-1}h^4|u|_{2,\Omega''}^2 \\
&\le \theta|||w^h|||_\varphi^2 + Ch^3|u|_{2,\Omega''}^2,
\end{aligned}$$

while in Ω_c we get

$$|(u^I - u, (I_h(\varphi w^h))_x)_{\Omega_c}| \leq C\|u\|_{\infty,\Omega_c} \max_{z \in \Omega_c} \varphi^{1/2}(z) \|\varphi^{-1/2}(I_h(\varphi w^h))_x\|_0$$

$$\leq \theta|||w^h|||_\varphi^2 + Ch^{\kappa-1}\|u\|_{\infty,\Omega_c}^2.$$

The other terms in (3.57) are handled similarly. One finally obtains

$$|||w^h|||_\varphi^2 \leq Ch^3|u|_{2,\Omega''}^2 + Ch^{\kappa-1}\|u\|_{\infty,\Omega_c}^2.$$

Substituting this result into (3.56) yields (3.55).   □

Remark 3.42. The assumption that the boundary of Ω' coincides with mesh lines is not essential; it merely enables us to use a simple cut-off function. For the general case see [JSW87].   ♣

Remark 3.43. The global optimal-order L_2-error estimate (3.50) for a three-directional mesh also holds true locally, but unlike Theorem 3.41 one chooses $\varepsilon_m = Mh^2$; see [Zho95, ZR96].   ♣

We summarize the known local pointwise error estimates for the SDFEM. The first local pointwise error estimate for the modified SDFEM (3.45) was given in [JSW87] under local smoothness conditions. For piecewise linear elements on quasi-uniform meshes with $\varepsilon_m = h^{3/2}$ and $\delta = h$, they prove for each $(x_0, y_0) \in \Omega$ the estimate

$$|(u - u^h)(x_0, y_0)| \leq Ch^{5/4}|\log h|^{3/2}\|u\|_{2,\Omega_0} + Ch^\kappa, \tag{3.58}$$

where $\kappa \geq 2$ and

$$\Omega_0 := \{(x,y) \in \Omega : x - x_0 \leq Ch|\log h|, |y - y_0| \leq Ch^{3/4}|\log h|\}.$$

The requirement that the mesh be quasi-uniform, i.e., there is also a positive constant C such that $h \leq Ch_T$, is stronger than our usual shape-regular assumption. The proof of (3.58) exploits local bounds on the associated discrete Green's function. In [Nii90] this pointwise error estimate is sharpened to

$$|(u - u^h)(x_0, y_0)| \leq Ch^{11/8}|\log h| \|u\|_{2,\Omega_0} + Ch^\kappa, \tag{3.59}$$

again on quasi-uniform triangular meshes. Furthermore, using improved estimates for the approximation error on rectangular streamline-oriented uniform meshes (similar to those given in Lemma 3.35 for a triangular three-directional mesh), the estimate (3.59) is sharpened in [ZR96] for $\varepsilon_m = h^2$ and $\delta = h$ to

$$|(u - u^h)(x_0, y_0)| \leq Ch^2|\log h| \|u\|_{3,\Omega_0} + Ch^\kappa, \tag{3.60}$$

where

$$\Omega_0 := \{(x,y) \in \Omega : x - x_0 \leq Ch|\log h|, |y - y_0| \leq Ch|\log h|\}.$$

(A rectangular mesh is called streamline-oriented if the streamline direction coincides with mesh lines. Such a mesh is called uniform if the meshsizes are quasi-uniform in each coordinate direction.) All these estimates were proved for differential operators with constant coefficients and under the assumption that $\varepsilon < h^{3/2}$. Note that in (3.60) the order of convergence has been improved and also the local subdomain Ω_0 has reduced width.

Remark 3.44. The standard duality arguments of classical finite element analyses do not yield an improved L_2-error estimate here because negative powers of ε appear in the analysis. Regarding pointwise estimates, a detailed theoretical and experimental analysis of the convergence rate on specially constructed meshes is given in [Zho97]. Starting with a triangular streamline-directed uniform mesh and inserting additional lines that are parallel to the streamline direction, pointwise convergence rates of $\mathcal{O}(h^\alpha)$ with $3/2 \le \alpha \le 2$ are obtained for a smooth function u; here α depends on the number of lines inserted. Thus the SDFEM does not in general give local convergence of $\mathcal{O}(h^2)$. ♣

Finally, we return to the drawback that in general the streamline diffusion method does not satisfy a discrete maximum principle – recall Example 3.33. As a consequence, overshoots or undershoots in the computed solution can be observed near sharp layers. Several proposals for overcoming this unpleasant behaviour are made in the literature. All are based on adding so-called shock-capturing terms that lead to additional numerical viscosity [BE02, BE05, Cod93, HMM86, INSB96, JK06b, JSW87, KLR02, LR06, MH85, SE99, TP86]. A thorough review of methods that are designed to reduce spurious oscillations at layers, including extensive numerical tests, is given in [JK07a, JK07b, JK07c].

Let $V_h \subset H_0^1(\Omega)$ be a finite element space consisting of piecewise polynomials. Following [KLR02, LR06] we consider a shock-capturing variant of the streamline-diffusion method:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a_h(u_h, v_h) + a_{\mathrm{sc}}(u_h; u_h, v_h) = f_h(v_h) \qquad (3.61)$$

where $a_h(\cdot, \cdot)$ and $f_h(\cdot)$ are the bilinear and linear form associated with the streamline-diffusion method (3.34)–(3.36) and $a_{\mathrm{sc}}(\cdot; \cdot, \cdot)$ is a general shock-capturing term of the form

$$a_{\mathrm{sc}}(w; u, v) := \sum_{T \in \mathcal{T}_h} (\tau_T(w) D_{\mathrm{sc}} \nabla u, \nabla v)_T. \qquad (3.62)$$

Here $D_{\mathrm{sc}} : \Omega \to \mathbb{R}^{2\times2}$ is some symmetric positive semi-definite matrix function with $\|D_{\mathrm{sc}}\|_{L^\infty(\Omega)^{2\times2}} \le 1$. The non-negative limiter function τ_T is introduced to restrict the effect of shock capturing to subregions where the residual $Lu - f$ is too large. A common feature of most proposals is that τ_T is a function of the scaled residual

$$\tau_T(w) := \tau_T^*(R_T^*(w)), \qquad R_T^*(w) := \frac{\|Lw - f\|_{0,T}}{\kappa_T + \|w\|_{1,T}}, \tag{3.63}$$

where a regularization parameter $\kappa_T > 0$ has been introduced. (Older shock-capturing schemes [Cod93, GdC88, SE00, HMM86] with $\kappa_T = 0$ lead to ill-posed nonlinear problems.) Note that in general the scheme (3.61) is nonlinear.

Theorem 3.45. *Let the SD parameter δ_T satisfy*

$$0 \le \delta_T \le \frac{1}{2} \min\left\{ \frac{\omega}{c_T^2}, \frac{h_T^2}{\varepsilon \mu_{inv}^2} \right\}$$

for each $T \in \mathcal{T}_h$. Assume that the limiter function $\tau_T^* : \mathbb{R} \to \mathbb{R}$ is continuous. Then the streamline diffusion method with shock-capturing stabilization (3.61)–(3.63) has at least one solution u_h. Moreover, the a priori *estimate*

$$|||u_h|||_{SD}^2 + \sum_{T \in \mathcal{T}_h} \left\| \sqrt{\tau_T^*(u_h)}\, D_{sc}^{1/2} \nabla u_h \right\|_{0,T}^2 \le C \sup_{v_h \in V_h, |||v_h|||_{SD} = 1} (f, v_h)$$

holds true for each solution.

Proof. The proof relies on a variant of Brouwer's fixed-point theorem; for the h and hp version of the FEM see [KLR02] and [LR06] respectively. □

Remark 3.46. All extant uniqueness proofs make strong assumptions on τ_T^* that are usually not satisfied by the limiter functions used in practice.   ♣

Examples of the shock-capturing term (3.62) are now given, with $\mathbb{I}$ used to denote the unit tensor:

Artificial viscosity method [Ike83]

$$\tau_T(w) = \tau_T, \qquad D_{sc} = \mathbb{I}.$$

Artificial crosswind diffusion method [JSW87], see also (3.45)

$$\tau_T(w) = \tau_T, \qquad D_{sc} = \begin{cases} \mathbb{I} - \dfrac{b \otimes b}{|b|^2} & \text{if } b \ne 0, \\[2mm] 0 & \text{if } b = 0. \end{cases}$$

Nonlinear isotropic diffusion

$$\tau_T(w) = \sigma_T(w) \left(\frac{\|Lw - f\|_{0,T}}{\kappa_T + \|w\|_{1,T}} \right)^2, \qquad D_{sc} = \mathbb{I}$$

with appropriate σ_T, see [GdC88] or [Cod93].

Nonlinear crosswind diffusion

$$\tau_T(w) = \sigma_T(w)\frac{\|Lw - f\|_{0,T}}{\kappa_T + \|w\|_{1,T}}, \qquad D_{\text{sc}} = \begin{cases} \mathbb{I} - \dfrac{b \otimes b}{|b|^2} & \text{if } b \neq 0, \\[2ex] 0 & \text{if } b = 0, \end{cases}$$

with appropriate σ_T, see [Cod93, CS99].

Although these methods have been successfully used numerically with higher-order finite elements, all theoretical analysis for shock-capturing is up to now confined to establishing the existence of solutions that allow qualitatively the same error estimates as for stabilized methods without shock-capturing terms [KLR02, LR06]. In particular, a discrete maximum principle (which would explain the non-oscillatory behaviour of their solutions) has been rigorously established only for first-order finite elements on certain simplicial meshes [Cod93, Ike83, BE02, BE05].

Following [BE05], we now derive a discrete maximum principle for a method with a special nonlinear shock-capturing term applied to piecewise linear discretizations. To simplify the notation we will restrict ourselves to meshes satisfying the assumption of Xu and Zikatanov (3.5); recall that this assumption is satisfied for Delaunay triangulations and for weakly acute meshes (see Section 3.1). When proving a discrete maximum principle for discretizations that include nonlinear shock-capturing terms, the M-matrix framework of Section 3.1 is inapplicable and needs to be extended. To find a suitable generalization, consider first discretizations of $-\Delta$. As in Section 3.1, we use the standard piecewise linear basis functions φ_i, $i = 1, \ldots, N + M$, satisfying $\varphi_i(p_j) = \delta_{ij}$, where p_i, $i = 1, \ldots, N$, denote the inner vertices of the triangulation and p_i, $i = N + 1, \ldots, N + M$, are the boundary nodes. The set of all indices j of vertices p_j that are neighbours of p_i is

$$\Lambda_i := \{j \neq i \,:\, \exists\, T \in \mathcal{T}_h \text{ with } p_i,\, p_j \in T\}.$$

Suppose that u_h has a local minimum at an inner vertex p_i, i.e. $u_h(p_j) \geq u_h(p_i)$ for all $j \in \Lambda_i$. By [XZ99, Lemma 2.1] the discretization of $-\Delta$ is an M-matrix such that $(\nabla\varphi_j, \nabla\varphi_i) \leq 0$ for $j \neq i$. But

$$(\nabla u_h, \nabla\varphi_i) = u_h(p_i)\,(\nabla\varphi_i, \nabla\varphi_i) + \sum_{j \in \Lambda_i} u_h(p_j)\,(\nabla\varphi_j, \nabla\varphi_i),$$

so

$$(\nabla u_h, \nabla\varphi_i) \leq u_h(p_i)\left(\nabla\left[\varphi_i + \sum_{j \in \Lambda_i}\varphi_j\right], \nabla\varphi_i\right) = 0. \tag{3.64}$$
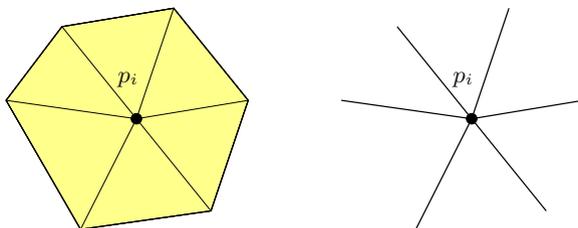
Fig. 3.6. Patch Ω_i (left) and set of edges $\mathcal{E}(p_i)$ (right) associated with an inner vertex $p_i \in \Omega$

In fact this inequality can be sharpened: let Ω_i, for $i = 1, \ldots, N$, be the union of all cells $T \in \mathcal{T}_h$ that have p_i as a vertex, let n_T be the outer unit normal on ∂T, and let $\mathcal{E}(p_i)$ be the set of all edges E to which p_i belongs; see Figure 3.6. Using the notation

$$[\nabla u_h]_E := \Big((\nabla u_h)|_T \cdot n_T + (\nabla u_h)|_{T'} \cdot n_{T'} \Big)\Big|_E$$

for the scalar jump of ∇u_h across the edge $E = T \cap T'$ (see Figure 3.7), and writing h_E for the length of edge E, elementwise integration by parts yields
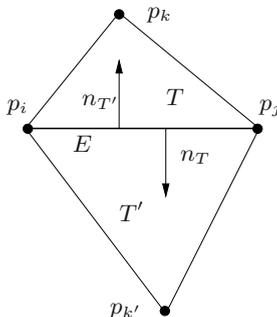
Fig. 3.7. Illustration for calculating the scalar jump $[\nabla u_h]_E$ across the edge E

$$(\nabla u_h, \nabla \varphi_i) = \sum_{T \subset \Omega_i} (\nabla u_h, \nabla \varphi_i)_T = \sum_{T \subset \Omega_i} \langle \nabla u_h \cdot n_T, \varphi_i \rangle_{\partial T}$$

$$= \sum_{E \in \mathcal{E}(p_i)} \langle [\nabla u_h]_E, \varphi_i \rangle_E = \sum_{E \in \mathcal{E}(p_i)} \frac{h_E}{2} [\nabla u_h]_E,$$

since $[\nabla u_h]_E$ is constant along E.

The assumption that u_h has a local minimum at the inner vertex $p_i \notin \partial \Omega$ implies that the term $[\nabla u_h]_E$ is non-positive for all $E \in \mathcal{E}(p_i)$. To see this, in the notation of Figure 3.7 one has $\varphi_i + \varphi_j + \varphi_k = 1$ on T, so

$$\nabla u_h|_T = (u_h(p_j) - u_h(p_i))\nabla\varphi_j + (u_h(p_k) - u_h(p_i))\nabla\varphi_k$$

and similarly

$$\nabla u_h|_{T'} = (u_h(p_j) - u_h(p_i))\nabla\varphi_j + (u_h(p_{k'}) - u_h(p_i))\nabla\varphi_{k'}.$$

As

$$n_T = \frac{\nabla\varphi_{k'}}{|\nabla\varphi_{k'}|} = -\frac{\nabla\varphi_k}{|\nabla\varphi_k|} = -n_{T'},$$

it follows that

$$[\nabla u_h]_E = \Big((\nabla u_h)|_T \cdot n_T + (\nabla u_h)|_{T'} \cdot n_{T'}\Big)\Big|_E$$
$$= -\Big[(u_h(p_k) - u_h(p_i))|\nabla\varphi_k| + (u_h(p_{k'}) - u_h(p_i))|\nabla\varphi_{k'}|\Big] \le 0.$$

To summarize, if u_h has a local minimum at an inner vertex $p_i \notin \partial\Omega$, then

$$(\nabla u_h, \nabla\varphi_i) = -\sum_{E\in\mathcal{E}(p_i)} \frac{h_E}{2}\left|[\nabla u_h]_E\right|.$$

This relation is stronger than the inequality (3.64) and forms the basis for the non-linear generalization of the discrete maximum principle (DMP) that we shall apply to the following problem:

Find $u_h \in V_{gh}$ such that for all $v_h \in V_{0h}$ one has

$$\tilde{a}_h(u_h; v_h) = (f, v_h). \tag{3.65}$$

Here $\tilde{a}(\,\cdot\,;\cdot\,)$ is a semilinear form that is linear in the second argument, and V_{gh} (V_{0h}) is the finite element space associated with inhomogeneous (homogeneous) boundary values g.

Definition 3.47. *The semilinear form $\tilde{a}_h(\,\cdot\,;\cdot\,)$ is said to satisfy the strong DMP property, if for all $u_h \in V_{gh}$ and for all inner vertices $p_i \in \Omega$ such that u_h is locally minimal at the vertex p_i over the patch Ω_i, there are positive quantities α_E such that*

$$\tilde{a}_h(u_h; \varphi_i) \le -\sum_{E\in\mathcal{E}(p_i)} \alpha_E\left|[\nabla u_h]_E\right|;$$

here φ_i is the piecewise linear basis function that satisfies $\varphi_i(p_j) = \delta_{ij}$.

Definition 3.48. *The semilinear form $\tilde{a}_h(\cdot\,;\cdot\,)$ is said to satisfy the weak DMP property, if for all $u_h \in V_{gh}$ and for all inner vertices $p_i \in \Omega$ such that u_h has a negative local minimum at the vertex p_i over the patch Ω_i, there are positive quantities α_E such that*

$$\tilde{a}_h(u_h; \varphi_i) \le -\sum_{E\in\mathcal{E}(p_i)} \alpha_E\left|[\nabla u_h]_E\right|;$$

here φ_i is the piecewise linear basis function that satisfies $\varphi_i(p_j) = \delta_{ij}$.

Following [BE05], we show that the strong/weak DMP property implies a strong/weak minimum principle for the discrete solution of (3.65).

Theorem 3.49. *Assume that the semilinear form $\tilde{a}_h(\cdot\,;\cdot)$ has the strong DMP property and that $(f, \varphi_i) \geq 0$ for $i = 1, \ldots, N$. Then the piecewise linear solution u_h of (3.65) reaches its minimum at a boundary node, i.e.,*

$$u_h(p_i) \geq \min_{p_j \in \partial\Omega} u_h(p_j) \quad \text{for } i = 1, \ldots, N + M.$$

If instead $\tilde{a}_h(\cdot\,;\cdot)$ has the weak DMP and $(f, \varphi_i) \geq 0$ for $i = 1, \ldots, N$, then the solution u_h of (3.65) satisfies

$$u_h(p_i) \geq \min_{p_j \in \partial\Omega} \left\{0, u_h(p_j)\right\} \quad \text{for } i = 1, \ldots, N + M.$$

Proof. Assume that $\tilde{a}_h(\cdot\,;\cdot)$ has the strong DMP property. Suppose that u_h attains its minimum at an inner vertex $p_i \in \Omega$. Then there are positive α_E such that

$$0 \leq (f, \varphi_i) = \tilde{a}_h(u_h; \varphi_i) \leq - \sum_{E \in \mathcal{E}(p_i)} \alpha_E \left|[\nabla u_h]_E\right|.$$

Consequently ∇u_h is constant on Ω_i and the minimum is attained also at a boundary node of Ω_i. One can continue this argument until the boundary of Ω is reached.

Now assume instead that $\tilde{a}_h(\cdot\,;\cdot)$ has the weak DMP property. To prove the second assertion of the theorem, observe that if u_h has a nonnegative minimum, or a negative minimum at a boundary vertex, then the result is immediate. Thus we need only consider the case where u_h attains a negative minimum at an inner vertex, and the above argument can be used again. □

Remark 3.50. If $\tilde{a}_h(-u_h; \varphi_i) = -\tilde{a}_h(u_h; \varphi_i)$, then a multiplication by -1 of the inequalities of Theorem 3.49 gives corresponding maximum principles. ♣

Remark 3.51. Define the discrete nonlinear operator $\tilde{L}_h$ associated with the semilinear form $\tilde{a}_h$ by

$$(\tilde{L}_h u_h, v_h) := \tilde{a}_h(u_h; v_h) \qquad \text{for all } u_h, v_h \in V_h.$$

The weak DMP property guarantees that $\tilde{L}_h$ is inverse-monotone; see (3.3). In particular for $f \equiv 0$ one has the discrete version of the maximum principle (3.4), viz.,

$$\min_{p_j \in \partial\Omega} \left\{0, u_h(p_j)\right\} \leq u_h(p_i) \leq \max_{p_j \in \partial\Omega} \left\{0, u_h(p_j)\right\} \qquad \text{for } i = 1, \ldots, N + M.$$

The strong DMP property implies the sharper estimate

$$\min_{p_j \in \partial\Omega} u_h(p_j) \leq u_h(p_i) \leq \max_{p_j \in \partial\Omega} u_h(p_j) \qquad \text{for } i = 1, \ldots, N + M.$$

As we saw on pages 321–323, one has the strong DMP property for the discretization of the Laplacian by piecewise linear finite elements on meshes that satisfy the Xu and Zikatanov condition (3.5).   ♣

Example 3.52. Consider for $c \geq 0$ the continuous problem

$$-u'' + cu = 0 \quad \text{in } (-1, +1), \qquad u(-1) = u(+1) = 1,$$

whose solution is $u(x) = \cosh \sqrt{c}x / \cosh \sqrt{c}$. For $c \geq 0$ one has a weak maximum principle:

$$0 = \min\{0, 1\} \leq u(x) \leq \max\{0, 1\} = 1.$$

But if $c > 0$ there can be no strong maximum principle because it would force u to be constant on $[-1, 1]$. Thus in the discrete case one cannot expect that the associated operator satisfies a strong DMP property for $c > 0$.   ♣

The following technical lemma will be needed later.

Lemma 3.53. *If $u_h \in V_{gh}$ has a local minimum at the vertex p_i, then*

$$\left| (\nabla u_h)|_T \right| \leq \sum_{E \in \mathcal{E}(p_i)} \left| [\nabla u_h]_E \right| \quad \text{for all } T \subset \Omega_i.$$

Proof. See [BE05, Lemma 2.7].   □

Now consider the model problem

$$-\varepsilon \Delta u + b \nabla u = 0 \quad \text{in } \Omega, \qquad u = g \quad \text{on } \Gamma = \partial \Omega,$$

where b is smooth. Let g_h be the piecewise linear interpolation of g. Let V_{gh} and $V_{0,h}$ be the finite element spaces comprising piecewise linear functions v_h that satisfy $v_h = g_h$ and $v_h = 0$ respectively on the boundary Γ. The streamline diffusion method would be based on the bilinear form

$$a_h(w, v) := \varepsilon (\nabla w, \nabla v) + (b \cdot \nabla w, v) + \sum_{T \in \mathcal{T}_h} \delta_T (b \cdot \nabla w, b \cdot \nabla v)_T.$$

We seek a shock-capturing term $a_{\mathrm{sc}}(w; v)$ such that the augmented semilinear form $\tilde{a}_h(w; v) := a_h(w, v) + a_{\mathrm{sc}}(w; v)$ enjoys the strong or weak DMP property. Then the SDFEM with shock-capturing for solving the model problem is:

Find $u_h \in V_{gh}$ such that for all $v_h \in V_{0h}$ one has

$$a_h(u_h, v_h) + a_{\mathrm{sc}}(u_h; v_h) = 0. \tag{3.66}$$

To construct a suitable shock-capturing term $a_{\mathrm{sc}}(\cdot; \cdot)$ one needs sharp estimates of $a_h(u_h; \varphi_i)$ when u_h has a local minimum at p_i relative to the patch Ω_i; recall Figure 3.6. It has already been shown that

$$(\nabla u_h, \nabla \varphi_i) = - \sum_{E \in \mathcal{E}(p_i)} \frac{h_E}{2} \left| [\nabla u_h]_E \right|. \tag{3.67}$$

The shape-regularity of the mesh implies the existence of a positive constant ρ such that

$$\max_{T \subset \Omega_i} |T| \leq \rho \min_{E \in \mathcal{E}(p_i)} h_E^2$$

and that there is a fixed maximum number of cells T in Ω_i, independently of the mesh size h. As $\nabla u_h|_T$ is constant, an appeal to Lemma 3.53 yields

$$\big|(b \cdot \nabla u_h, \varphi_i)\big| = \sum_{T \subset \Omega_i} |(b \cdot \nabla u_h, \varphi_i)_T| \leq \sum_{T \subset \Omega_i} \frac{|T| \, \|b\|_{0,\infty,T}}{3} \big|(\nabla u_h)|_T\big|$$

$$\leq \sum_{E \in \mathcal{E}(p_i)} h_E^2 \big|[\nabla u_h]_E\big| \left\{ \sum_{T \subset \Omega_i} \frac{|T|}{3 h_E^2} \|b\|_{0,\infty,\Omega_i} \right\}$$

$$\leq d_1 \sum_{E \in \mathcal{E}(p_i)} h_E^2 \, \|b\|_{0,\infty,\omega(E)} \big|[\nabla u_h]_E\big|, \tag{3.68}$$

where $\omega(E) = \Omega_i \cup \Omega_j$ when E is the edge joining the vertices p_i and p_j, and d_1 is some fixed constant. Analogously, the stabilizing term satisfies

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T (b \cdot \nabla u_h, b \cdot \nabla \varphi_i)_T \right| = \left| \sum_{T \subset \Omega_i} \delta_T (b \cdot \nabla u_h, b \cdot \nabla \varphi_i)_T \right|$$

$$\leq \sum_{T \subset \Omega_i} C \delta_T \|b\|_{0,\infty,T}^2 |T|^{1/2} \big|(\nabla u_h)|_T\big|$$

$$\leq d_2 \sum_{E \in \mathcal{E}(p_i)} h_E^2 \, \|b\|_{0,\infty,\omega(E)} \big|[\nabla u_h]_E\big|, \tag{3.69}$$

where Lemma 3.53 was used again, the standard choice of δ_T in the convection-dominated regime was made, viz., $\|b\|_{0,\infty,T}\delta_T \leq C h_T$, and the shape-regularity of the mesh gives $h_T |T|^{1/2} \leq C h_E^2$. In (3.69), d_2 is a fixed constant.

Consider the term

$$\psi_E(u_h; v_h) = \left(\mathrm{sgn}\left(\frac{\partial u_h}{\partial t_E} \right) \right) \frac{\partial v_h}{\partial t_E} h_E$$

on any edge $E \in \mathcal{E}$, where $\partial/\partial t_E$ denotes the tangential derivative along E. It is not difficult to see that for a basis function $v_h = \varphi_i$ the term depends only on the restriction of u_h onto Ω_i, which means that it is local, and $\psi_E(u_h; \varphi_i) = -1$ for u_h locally minimal at p_i. Thus, one possible shock-capturing term is

$$a_{\mathrm{sc}}(u_h; v_h) := c_\rho \sum_{E \in \mathcal{E}} h_E^2 \, \|b\|_{0,\infty,\omega(E)} \big|[\nabla u_h]_E\big| \psi_E(u_h; v_h) \tag{3.70}$$

with a positive constant c_ρ that depends on the shape-regularity constant. Indeed, we have the following

Theorem 3.54. *Let the SD parameter δ_T satisfy $\delta_T \leq C h_T / \|b\|_{0,\infty,T}$. Define the shock-capturing term by (3.70). Then for sufficiently large c_ρ, the semilinear form corresponding to the SDFEM with shock-capturing satisfies the strong DMP.*

Proof. Collecting the estimates (3.67)–(3.69), one gets

$$\tilde{a}_h(u_h; \varphi_i) \leq - \sum_{E \in \mathcal{E}(p_i)} \left[\frac{\varepsilon}{2} + (c_\rho - 1)(d_1 + d_2) h_E \|b\|_{0,\infty,\omega(E)} \right] h_E \left| [\nabla u_h]_E \right|$$

which establishes the strong DMP property of the semilinear form $\tilde{a}_h$. $\square$

Remark 3.55. For clarity in our presentation we did not compute precise formulas specifying how d_1 and d_2 depend on the data and on the mesh parameters in (3.68) and (3.69). More detailed information and an extension to the three-dimensional case as well as to the continuous interior penalty (CIP) method of Section 3.3.2 below can be found in [BE05].   ♣

Remark 3.56. One has $f \equiv 0$ in the model problem above and the shock-capturing term (3.70) satisfies $a_{sc}(-u_h, v_h) = -a_{sc}(u_h, v_h)$. Hence the discrete solution u_h of (3.66) achieves its minimum and its maximum at the boundary if c_ρ is chosen large enough. That is, the discrete maximum principle suppresses excessive oscillations.   ♣

3.2.2 Galerkin Least Squares Finite Element Method (GLSFEM)

Two important features of the streamline diffusion finite element method are:

(a) the standard Galerkin method is augmented by the addition of terms that represent the residual of the original differential equation on each mesh element;

(b) since the residual of the exact solution is zero, the method is automatically consistent, unlike some other upwind methods.

The above features can be preserved in a more general framework, the Galerkin least squares method, which tries to combine certain advantages of the Galerkin and least squares methods: the Galerkin method has the projection property and can use elements that lie only in $C^0(\Omega)$, while the least squares method can be applied to a large class of problems.

Let us briefly describe the application of the classical least squares method to solving the problem

$$Lu := -\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad u\big|_{\partial\Omega} = 0. \tag{3.71}$$

Begin by choosing a finite dimensional space $W_h \subset H^2(\Omega)$ that comprises functions v_h satisfying the homogeneous boundary condition $v_h\big|_{\partial\Omega} = 0$. Then we seek a solution $u_h \in W_h$ of the minimization problem

$$\| Lu_h - f \|_0^2 = \min_{v_h \in W_h} \| Lv_h - f \|_0^2.$$

This is equivalent to the problem:

Find $u_h \in W_h$ such that for all $v_h \in W_h$ one has

$$(Lu_h - f, Lv_h) = 0. \tag{3.72}$$

A drawback of this method – compared with the standard Galerkin finite element method – is that when using piecewise polynomials, the assumption that $W_h \subset H^2(\Omega)$ requires the use of $C^1(\Omega)$ elements, but the construction of $C^1(\Omega)$ elements on arbitrary triangulations is not easy. A second drawback is that the condition number of the matrix associated with the discrete problem (3.72) is larger than the condition number encountered in the standard Galerkin approach (but see [CLMM94, FMM98], where the original elliptic problem is transformed into a first-order system to avoid this difficulty). On the other hand, the least squares method does not have restricted stability properties like the standard Galerkin finite element method in the singularly perturbed case; furthermore, the matrix associated with (3.72) is symmetric and positive definite.

The aim of this section is to combine the best features of the Galerkin and least squares methods. Let us introduce the residual

$$\sum_{T \in \mathcal{T}_h} \delta_T \left(Lu_h - f, Lv_h \right)_T \tag{3.73}$$

of the equation (3.71), which is evaluated element by element. This permits use of C^0 elements from any space $V_h \subset H_0^1(\Omega)$. The basic idea of the *Galerkin least squares finite element method* (GLSFEM) ([HS88, HFH89]) is to add this term to the standard Galerkin finite element method:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$\varepsilon(\nabla u_h, \nabla v_h) + (b \cdot \nabla u_h + cu, v_h) + \sum_{T \in \mathcal{T}_h} \delta_T \left(Lu_h, Lv_h \right)_T$$

$$= (f, v_h) + \sum_{T \in \mathcal{T}_h} \delta_T \left(f, Lv_h \right)_T . \tag{3.74}$$

Remark 3.57. (Generalized GLSFEM) More generally, instead of (3.73), one could add the term

$$\sum_{T \in \mathcal{T}_h} \delta_T \left(Lu_h - f, \psi(v_h) \right)_T$$

to the standard Galerkin finite element method, where ψ is some user-chosen operator. If $\psi(v_h) = Lv_h$ one recovers the GLSFEM, while if $\psi(v_h) = b \cdot \nabla v_h$ the SDFEM is generated. Other choices are possible, e.g., in [FFH92] $\psi(v_h) = b \cdot \nabla v_h + \varepsilon \Delta v_h$ was studied in the case $c = 0$.   ♣

We now analyse the GLSFEM. Assume that b, c and f are sufficiently smooth with $c - \frac{1}{2} \nabla \cdot b \geq \omega > 0$. The discrete solution is sought in the space

$$V_h := \{ v_h \in V_h : v_h |_T \in P_k(T) \quad \text{for all} \quad T \in \mathcal{T}_h \}$$

of piecewise polynomials of degree $k \geq 1$. Introduce the mesh-dependent norm

$$|||v_h|||_{GLS} := \left(\varepsilon |v_h|_1^2 + \omega \|v_h\|_0^2 + \sum_{T \in \mathcal{T}_h} \delta_T \|Lv_h\|_{0,T}^2 \right)^{1/2},$$

the bilinear form

$$a_h(w,v) := \varepsilon (\nabla w, \nabla v) + (b \cdot \nabla w + cw, v) + \sum_{T \in \mathcal{T}_h} \delta_T (Lw, Lv)_T,$$

and the linear form

$$f_h(v) := (f, v) + \sum_{T \in \mathcal{T}_h} \delta_T (f, Lv)_T.$$

The next result shows that the GLSFEM is more stable than the standard Galerkin method. Unlike the SDFEM of Section 3.2.1, no upper bound on δ_T is needed in the proof.

Lemma 3.58. *Let the Galerkin least squares parameter δ_T be positive. Then the discrete bilinear form a_h is coercive on $V_h \times V_h$, i.e.,*

$$a_h(v_h, v_h) \geq |||v_h|||_{GLS}^2 \quad \textit{for all } v_h \in V_h,$$

and the linear form f_h is continuous on V_h, i.e.,

$$|f_h(v_h)| \leq C \left[\|f\|_0 + \left(\sum_{T \in \mathcal{T}_h} \delta_T \|f\|_{0,T}^2 \right)^{1/2} \right] |||v_h|||_{GLS} \quad \textit{for all } v_h \in V_h.$$

Proof. For each $v_h \in V_h$, one has

$$a_h(v_h, v_h) = \varepsilon |v_h|_1^2 + (c - \frac{1}{2} \nabla \cdot b, v_h^2) + \sum_{T \in \mathcal{T}_h} \delta_T \|Lv_h\|_{0,T}^2$$

$$\geq |||v_h|||_{GLS}^2$$

and

$$|f_h(v_h)| \leq \|f\|_0 \|v_h\|_0 + \left(\sum_{T \in \mathcal{T}_h} \delta_T \|f\|_{0,T}^2 \right)^{1/2} \left(\sum_{T \in \mathcal{T}_h} \delta_T \|Lv_h\|_{0,T}^2 \right)^{1/2}$$

$$\leq C \left[\|f\|_0 + \left(\sum_{T \in \mathcal{T}_h} \delta_T \|f\|_{0,T}^2 \right)^{1/2} \right] |||v_h|||_{GLS}.$$

$\square$

Remark 3.59. Lemma 3.58 implies the *a priori* estimate

$$|||u_h|||_{GLS} \leq C \left[\|f\|_0 + \left(\sum_{T \in \mathcal{T}_h} \delta_T \|f\|_{0,T}^2 \right)^{1/2} \right].$$

Moreover, we have the stability inequality

$$|||u_h|||_{GLS} \leq |||A_h u_h|||_* \quad \text{for all } u_h \in V_h,$$

where the discrete operator $A_h : V_h \to V_h^*$ is defined by

$$< A_h v_h, w_h >:= a_h(v_h, w_h) \quad \forall v_h, w_h \in V_h$$

and the norm on the dual space V_h^* is

$$|||g_h|||_* := \sup_{w_h \in V_h} \frac{< g_h, w_h >}{|||w_h|||_{GLS}} \quad \forall g_h \in V_h^*.$$

Thus the GLSFEM has additional stability compared with the standard Galerkin finite element method, where $\delta_T = 0$. ♣

If an upper bound is imposed on δ_T, then the GLSFEM can also control the derivatives in the streamline direction. To demonstrate this, introduce the mesh-dependent norm

$$|||v_h|||_{SDGLS} := \left(|||v_h|||_{SD}^2 + \sum_{T \in \mathcal{T}_h} \delta_T \|cv_h - \varepsilon \Delta v_h\|_{0,T}^2 \right)^{1/2},$$

where $||| \cdot |||_{SD}$ was defined in Section 3.2.1.

Lemma 3.60. *Let the GLS parameter δ_T satisfy*

$$0 < \delta_T < \frac{1}{8} \min \left\{ \frac{\omega}{c_T^2}, \frac{h_T^2}{\varepsilon \mu_{\mathrm{inv}}^2} \right\},$$

where μ_{inv} is defined in (3.32) and $c_T := \max_{x \in T} |c(x)|$. Then the discrete bilinear form is coercive with respect to $||| \cdot ||||_{SDGLS}$, i.e.,

$$a_h(v_h, v_h) \geq \frac{1}{2} |||v_h|||_{SDGLS}^2 \quad \text{for all } v_h \in V_h.$$

Furthermore, one has the a priori *estimate*

$$|||u_h|||_{SDGLS} \leq 2\|f\|_0 + 4 \left(\sum_{T \in \mathcal{T}_h} \delta_T \|f\|_{0,T}^2 \right)^{1/2}.$$

Proof. For all $v_h \in V_h$, we have

$$a_h(v_h, v_h) = \varepsilon |v_h|_1^2 + (c - \frac{1}{2}\nabla \cdot b, v_h^2) + \sum_{T \in \mathcal{T}_h} \delta_T \|b \cdot \nabla v_h\|_{0,T}^2$$

$$+ 2 \sum_{T \in \mathcal{T}_h} \delta_T (cv_h - \varepsilon \Delta v_h, b \cdot \nabla v_h)_T + \sum_{T \in \mathcal{T}_h} \delta_T \|cv_h - \varepsilon \Delta v_h\|_{0,T}^2$$

$$\geq |||v_h|||_{SDGLS}^2 - \sum_{T \in \mathcal{T}_h} \frac{\delta_T}{2} \|b \cdot \nabla v_h\|_{0,T}^2$$

$$- 4 \sum_{T \in \mathcal{T}_h} \delta_T \left(\|cv_h\|_{0,T}^2 + \|\varepsilon \Delta v_h\|_{0,T}^2 \right).$$

Using the local inverse inequality $\|\Delta v_h\|_{0,T} \leq \mu_{\text{inv}}\, h_T^{-1}|v_h|_{1,T}$ and the upper bound for δ_T gives

$$a_h(v_h, v_h) \geq |||v_h|||^2_{SDGLS} - \sum_{T \in \mathcal{T}_h} \frac{\delta_T}{2}\|b \cdot \nabla v_h\|^2_{0,T} - \frac{\omega}{2}\|v_h\|^2_0 - \frac{\varepsilon}{2}|v_h|^2_1$$

$$\geq \frac{1}{2}\,|||v_h|||^2_{SDGLS}\,.$$

$\square$

The convergence properties of the GLSFEM are studied next.

Theorem 3.61. *Assume that $u \in H^1_0(\Omega) \cap H^{k+1}(\Omega)$, where $k \geq 1$. Let the GLS parameter δ_T be positive. Then for the solution u_h of the GLSFEM (3.74) one has the global error estimate*

$$|||u - u_h|||_{GLS} \leq C\, h^k \left(\sum_{T \in \mathcal{T}_h} \lambda(\varepsilon, \delta_T, h_T)|u|^2_{k+1} \right)^{1/2},$$

where

$$\lambda(\varepsilon, \delta_T, h_T) := \varepsilon + \varepsilon^2 \delta_T h_T^{-2} + \delta_T + h_T^2 + \delta_T^{-1} h_T^2.$$

Proof. The solution u belongs to the space $H^1_0(\Omega) \cap H^{k+1}(\Omega)$ with $k \geq 1$, so the scheme is consistent, i.e., $a_h(u, v_h) = f_h(v_h)$ for all $v_h \in V_h$. Hence, writing u^I for the interpolant of u from V_h, one obtains

$$|||u^I - u_h|||^2_{GLS} \leq a_h(u^I - u_h, u^I - u_h) = a_h(u^I - u, u^I - u_h).$$

Invoking the interpolation properties (3.31) and estimating separately each term on the right-hand side, we have

$$|\varepsilon(\nabla(u^I - u), \nabla(u^I - u_h))| \leq C\varepsilon^{1/2}h^k|u|_{k+1}\,|||u^I - u_h|||_{GLS}\,,$$

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T\big(L(u^I - u), L(u^I - u_h)\big)_T \right|$$

$$\leq \left(\sum_{T \in \mathcal{T}_h} \delta_T\|L(u^I - u)\|^2_{0,T} \right)^{1/2} |||u^I - u_h|||_{GLS}$$

$$\leq C\, h^k \left(\sum_{T \in \mathcal{T}_h} \delta_T(\varepsilon^2 h_T^{-2} + 1 + h_T^2) \right)|u|_{k+1}\,|||u^I - u_h|||_{GLS}$$

and

$$|(c(u^I - u), u^I - u_h)| \leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2k+2}|u|^2_{k+1,T} \right)^{1/2} |||u^I - u_h|||_{GLS}.$$

The term

$$(b \cdot \nabla(u^I - u), u^I - u_h).$$

has not yet been estimated. If we were to bound this term by

$$|(b \cdot \nabla(u^I - u), u^I - u_h)| \leq C \, h^k \, |u|_{k+1} \, |||u^I - u_h|||_{GLS},$$

it is then impossible to extract any extra power of h in the convection-dominated case $\varepsilon \leq h$. Therefore, integrate by parts then complete the term to get $L(u^I - u_h)$:

$$
\begin{aligned}
- &(b \cdot \nabla(u^I - u), u^I - u_h) \\
&= (u^I - u, b \cdot \nabla(u^I - u_h)) + (u^I - u, \nabla \cdot b \, (u^I - u_h)) \\
&= \sum_{T \in \mathcal{T}_h} (u^I - u, L(u^I - u_h))_T + \sum_{T \in \mathcal{T}_h} (u^I - u, (\nabla \cdot b - c)(u^I - u_h))_T \\
&\quad + \sum_{T \in \mathcal{T}_h} (u^I - u, \varepsilon \Delta(u^I - u_h))_T \, .
\end{aligned}
$$

The first term on the right-hand side can now be bounded by

$$
\left| \sum_{T \in \mathcal{T}_h} (u^I - u, \, L(u^I - u_h))_T \right|
$$

$$
\leq C \left(\sum_{T \in \mathcal{T}_h} \delta_T^{-1} h_T^{2k+2} |u|_{k+1,T}^2 \right)^{1/2} |||u^I - u_h|||_{GLS}
$$

$$
\leq C h^k \left(\sum_{T \in \mathcal{T}_h} \delta_T^{-1} h_T^2 |u|_{k+1,T}^2 \right)^{1/2} |||u^I - u_h|||_{GLS} \, .
$$

The second term is dealt with in a standard way:

$$
| \sum_{T \in \mathcal{T}_h} (u^I - u, (\nabla \cdot b - c)(u^I - u_h))_T |
$$

$$
\leq C \, h^{k+1} |u|_{k+1} \|u^I - u_h\|_0
$$

$$
\leq C (\sum_{T \in \mathcal{T}_h} h_T^{2k+2} |u|_{k+1,T}^2)^{1/2} |||u^I - u_h|||_{GLS}.
$$

To bound the third term, one appeals to the local inverse inequality (3.32):

$$
\left| \sum_{T \in \mathcal{T}_h} (u^I - u, \varepsilon \Delta(u^I - u_h))_T \right| \leq C \sum_{T \in \mathcal{T}_h} h_T^{k+1} |u|_{k+1,T} \, \varepsilon \mu_{\mathrm{inv}} h_T^{-1} |u^I - u_h|_{1,T}
$$

$$
\leq C \varepsilon^{1/2} h^k |u|_{k+1} \, |||u^I - u_h|||_{GLS} \, .
$$

Finally, combining all these inequalities yields

$$|||u^I - u_h|||_{GLS}$$

$$\leq Ch^k \left[\sum_{T \in \mathcal{T}_h} \left(\varepsilon + \varepsilon^2 \delta_T h_T^{-2} + \delta_T + h_T^2 + \delta_T^{-1} h_T^2 \right) |u|_{k+1,T}^2 \right]^{1/2}$$

from which, via a triangle inequality, the error estimate follows.   □

Remark 3.62. The choice

$$\delta_T \sim \frac{h_T}{\sqrt{1 + (\varepsilon/h_T)^2}} \tag{3.75}$$

minimizes the expression $\varepsilon^2 \delta_T h_T^{-2} + \delta_T + \delta_T^{-1} h_T^2$ in Theorem 3.61. Thus, taking the GLS parameter

$$\delta_T = \delta_0 \frac{h_T}{\sqrt{1 + (\varepsilon/h_T)^2}}$$

with some user-chosen constant δ_0, we have

$$|||u^I - u_h|||_{GLS} \leq C(\varepsilon^{1/2} + h^{1/2}) \, h^k |u|_{k+1} \,.$$

This choice of δ_T corresponds asymptotically to the choice of the SD parameter in Section 3.2.1: from (3.75), for $\varepsilon \leq Ch_T$ we see that $\delta_T \sim h_T$ and for $\varepsilon \geq Ch_T$, we have $\delta_T \sim h_T^2/\varepsilon$; compare (3.38).   ♣

Remark 3.63. The optimal choice of the GLS parameter is, as for the SDFEM, an open question. If $c \equiv 0$, b and f are constant and $h_T = h$ for all T, then a nodally exact solution can be obtained in the one-dimensional case by setting

$$\delta_T = \frac{h}{2|b|} \left(\coth Pe - \frac{1}{Pe} \right) \quad \text{with} \quad Pe := \frac{|b|h}{2\varepsilon} \,.$$

In the asymptotic limit case $\varepsilon \ll h$, i.e., $Pe \gg 1$, one then has $\delta_T \sim h$, while for $\varepsilon \gg h$ we have instead $\delta_T \sim h^2/\varepsilon$.   ♣

Remark 3.64. (Galerkin gradient least squares method) For reaction-diffusion problems where $b \equiv 0$, a so-called Galerkin gradient least squares method is proposed in [FdC89]. Here a stabilization term of the form

$$\sum_{T \in \mathcal{T}_h} \delta_T (\nabla(Lu_h - f), \nabla(Lv_h))_T$$

is used instead of $\sum_{T \in \mathcal{T}_h} \delta_T (Lu_h - f, \psi(v_h))_T$.   ♣

3.2.3 Residual-Free Bubbles

A new characterization of the streamline diffusion method is presented in [BR94]. Its key observation is that the SDFEM for piecewise linear finite

elements is equivalent to a Galerkin approach using standard finite element spaces enriched by "bubble functions" where static condensation of the bubble component of the solution yields the SDFEM [BBF93]. A comparison of both approaches is carried out by Russo [Rus06], who writes "the importance of these ideas lies in the recognition that the variational framework should be used as a 'safe guide' in the development of new numerical methods".

To elucidate the bubble function approach, let us consider the problem (3.30) with piecewise constant functions b and f and $c \equiv 0$. Assume that V_h consists of piecewise linear functions; enrich this space by a *bubble space* B_h defined by

$$B_h := \operatorname{span}\{b_T \in H_0^1(T),\ \forall T \in \mathcal{T}_h\,\}, \qquad \dim(B_h) < \infty.$$

This definition of B_h permits very general bubble functions b_T which will be restricted later. Now consider the standard Galerkin FEM on the enriched space $V_h \oplus B_h$:

Find $u_h \in V_h \oplus B_h$ such that for all $v_h \in V_h \oplus B_h$ one has

$$\varepsilon(\nabla u_h, \nabla v_h) + (b \cdot \nabla u_h, v_h) = (f, v_h). \tag{3.76}$$

The dimension of this system of equations can be reduced by static condensation of the bubble component of the solution. To do this, write the solution u_h as $u_h = u_L + u_B$, with $u_L \in V_h$ and $u_B \in B_h$, and apply the test functions $v_h = v_L \in V_h$ and $v_h = b_T \in B_h$. Then (3.76) can be reformulated as:

Find $u_L \in V_h$ and $u_B \in B_h$ such that for all $v_L \in V_h$ and all $v_B \in B_h$,

$$\varepsilon(\nabla(u_L + u_B), \nabla v_L) + (b \cdot \nabla(u_L + u_B), v_L) = (f, v_L), \tag{3.77a}$$

$$\varepsilon(\nabla(u_L + u_B), \nabla v_B) + (b \cdot \nabla(u_L + u_B), v_B) = (f, v_B). \tag{3.77b}$$

Now $u_B = \sum_{T \in \mathcal{T}_h} d_T b_T$ where the d_T are unknown constants, so (3.77b) is equivalent to:

Given $u_L \in V_h$, find $\{d_T : d_T \in R\}$ such that for each T one has

$$\varepsilon\big(\nabla(u_L + d_T b_T), \nabla b_T\big)_T + \big(b \cdot \nabla(u_L + d_T b_T), b_T\big)_T = (f, b_T)_T. \tag{3.78}$$

An integration by parts then gives

$$\varepsilon(\nabla u_L, \nabla b_T)_T = -\varepsilon(\Delta u_L, b_T)_T + \left\langle \varepsilon \frac{\partial u_L}{\partial n}, b_T \right\rangle_{\partial T} = 0,$$

$$d_T(b \cdot \nabla b_T, b_T)_T = \frac{d_T}{2}\langle b \cdot n, b_T^2 \rangle_{\partial T} = 0,$$

so one can solve (3.78) for d_T, obtaining

$$d_T = \frac{(1, b_T)_T}{\varepsilon |b_T|_{1,T}^2}\, (f - b \cdot \nabla u_L)\big|_T.$$

Similarly $\varepsilon(\nabla u_B, \nabla v_L) = 0$ and one can reduce (3.77a) to

$$\varepsilon(\nabla u_L, \nabla v_L) + (b \cdot \nabla u_L, v_L) + \sum_{T \in \mathcal{T}_h} d_T (b \cdot \nabla b_T, v_L)_T = (f, v_L).$$

The term $\sum_{T \in \mathcal{T}_h} \cdots$ does not appear in the standard Galerkin finite element method applied on the space V_h. It can be rewritten as

$$\sum_{T \in \mathcal{T}_h} d_T (b \cdot \nabla b_T, v_L)_T = - \sum_{T \in \mathcal{T}_h} d_T (b \cdot \nabla v_L, b_T)_T$$

$$= \sum_{T \in \mathcal{T}_h} \gamma_T (b \cdot \nabla u_L - f, b \cdot \nabla v_L)_T,$$

where

$$\gamma_T = \frac{1}{|T|} \frac{|(1, b_T)_T|^2}{\varepsilon |b_T|_{1,T}^2} . \tag{3.79}$$

The bubble component has now been eliminated from (3.76), giving

$$\varepsilon(\nabla u_L, \nabla v_L) + (b \cdot \nabla u_L, v_L) + \sum_{T \in \mathcal{T}_h} \gamma_T (b \cdot \nabla u_L, b \cdot \nabla v_L)_T$$

$$= (f, v_L) + \sum_{T \in \mathcal{T}_h} \gamma_T (f, b \cdot \nabla v_L)_T \quad \text{for all } v_L \in V_h. \tag{3.80}$$

This is the SDFEM with the SD parameter $\delta_T = \gamma_T$ specified by (3.79). Clearly the choice of bubble function b_T determines the value of the SD parameter γ_T.

Remark 3.65. The simplest bubble function is the product of barycentric coordinates λ_i^T, $i = 1, 2, 3$. A scaling argument then shows that $\gamma_T \sim h_T^2/\varepsilon$, which corresponds to the choice of δ_T in the diffusion-dominated case of (3.38). ♣

Now let us consider the largest possible bubble space

$$B_h := \{v \in H_0^1(\Omega) : v|_T \in H_0^1(T) \ \forall T \in \mathcal{T}_h\}, \qquad \dim(B_h) = \infty.$$

As above one can split the problem (3.76) into (3.77a) and (3.77b), but in contrast to the previous case (3.77b) is now an infinite-dimensional problem. Integrating by parts over each $T \in \mathcal{T}_h$, one obtains

$$(\nabla u_L, \nabla v_B)_T = (\nabla u_B, \nabla v_L)_T = 0, \qquad \forall u_L, v_L \in V_L, \quad \forall u_B, v_B \in B_h$$

and hence the local problems:

Given $u_L \in V_h$, find $u_B \in B_h$ such that for each T and all $v_B \in B_h$,

$$\varepsilon(\nabla u_B, \nabla v_B)_T + (b \cdot \nabla u_B, v_B)_T = (f - b \cdot \nabla u_L, v_B)_T. \tag{3.81}$$

Observe that (3.81) is the weak formulation of the problem

$$-\varepsilon \Delta u_B + b \cdot \nabla u_B = (f - b \cdot \nabla u_L)\big|_T \quad \text{in } T, \quad u_B = 0 \quad \text{on } \partial T.$$

Thus, choosing the *residual-free bubble function* b_T to be the solution of

$$-\varepsilon \Delta b_T + b \cdot \nabla b_T = 1 \quad \text{in } T, \quad b_T = 0 \quad \text{on } \partial T, \tag{3.82}$$

the component $u_B \in B_h$ can be represented as

$$u_B = \sum_{T \in \mathcal{T}_h} (f - b \cdot \nabla u_L)\big|_T \, b_T.$$

Note that the choice of b_T in (3.82) is a natural generalization of the L-spline ψ_i used in (I.2.27). Eliminating u_B from (3.77a) again produces (3.80) with γ_T specified by (3.79). Multiplying (3.82) by b_T and integrating by parts shows that

$$\varepsilon |b_T|_{1,T}^2 = (1, b_T)_T,$$

so in order to determine γ_T from (3.79), one must compute

$$\gamma_T = \frac{1}{|T|}(1, b_T)_T$$

for the solution b_T of (3.82). In the one-dimensional case, (3.82) can be solved explicitly so γ_T can be computed; this yields the Il'in-Allen-Southwell scheme. But in higher dimensions the exact solution of (3.82) seems to be impossible. Nevertheless, one can generalize the method to an abstract setting by relaxing the assumptions that the data (b and f) are constant and that piecewise linear elements are used, as we now demonstrate.

Let us start again with the Galerkin formulation of (3.76): we seek $u_h \in V_h$, $u_B \in B_h$ such that

$$\begin{align}
a(u_h, v_h) + a(u_B, v_h) &= (f, v_h) & \text{for all } v_h \in V_h, \tag{3.83a} \\
a(u_h, v_B) + a(u_B, v_B) &= (f, v_B) & \text{for all } v_B \in B_h, \tag{3.83b}
\end{align}$$

where V_h is some finite element space of piecewise polynomials of degree $k \geq 1$, the space B_h is a suitable bubble space with $V_h \cap B_h = \emptyset$, and the bilinear form is

$$a(u, v) := \varepsilon(\nabla u, \nabla v) + (b \cdot \nabla, v).$$

Two new functions in B_h are now introduced.

Find $M(u_h)$, $F(f) \in B_h$ such that for all $v_B \in B_h$ one has

$$a(M(u_h), v_B) = -a(u_h, v_B), \qquad a(F(f), v_B) = (f, v_B).$$

Then $u_B = M(u_h) + F(f)$ can be eliminated from (3.83), yielding the *exact residual-free bubble* (RFB) method:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$a(u_h + M(u_h), v_h) = (f, v_h) - a(F(f), v_h).$$

The stability properties of the exact RFB method are identical to those of the SDFEM when piecewise linears are used, but for piecewise bilinears on rectangular meshes certain mesh configurations must be avoided to ensure that the RFB method is as stable as the SDFEM; for more details see [FT02].

The terms $a(M(u_h), v_h)$ and $a(F(f), v_h)$ must be evaluated to implement the exact RFB method, but this means solving an infinite-dimensional problem so in practice some type of approximation is used. Various possibilities are the pseudo-residual-free bubble method [BMR98], the stabilizing subgrid method [BMR05], and two-level and three-level approaches [FN01, FNS98, GWR04, GWR05].

Remark 3.66. A priori error estimates for the exact residual-free bubble method are given for finite elements on simplicial meshes in [BHM$^+$99, BMS00, ARS04]. Bilinears on quadrilateral meshes without the assumption of shape regularity are studied in [Ris01]. The case of higher- order finite elements on anisotropic quadrilateral meshes is investigated in [CS07]. Local estimates are discussed in [San00, San01]. *A priori* error estimates for some approximate residual-free bubble methods are given in [ARS04, FJMT07]. ♣

Remark 3.67. The residual-free bubble method is closely related to the variational multiscale method [Hug95, HS07, HFMQ98] that was examined in Section I.2.2.4. The essential difference between these methods is that in the residual-free bubble approach one has $V_h \oplus B_h \subset V$ but $V_h \oplus B_h \neq V$, whereas in the variational multiscale method we make the splitting $V = V_h \oplus V^\diamond$ into resolvable and unresolvable scales. ♣

3.3 Adding Symmetric Stabilizing Terms

When the residual-based stabilization methods of the previous section are applied to systems of convection-diffusion-reaction problems, this engenders couplings between the dependent variables but in general these couplings do not have any physical counterpart. In optimal control problems, residual-based stabilization methods lead to different discrete adjoint equations depending on whether the discretization of the problem or the construction of the adjoint is carried out first [BV07, BL08]. It has been observed that the asymmetry of the stabilizing term means that the computed control is significantly affected by the way in which the discrete optimality condition is defined. Moreover, in the case of transient problems, this asymmetric stabilization does not lead to diagonal matrices for the reaction term when a lumping technique (nodal quadrature) is applied; this is awkward for convection-dominated flows with zones of strong reaction. In the next two subsections we consider symmetric stabilization methods that avoid these failings.

3.3.1 Local Projection Stabilization

In residual-based stabilization methods with a given finite element space Y_h, several terms are added to the standard Galerkin method. For example, the streamline diffusion method adds

$$\sum_{T \in \mathcal{T}_h} \delta_T (-\varepsilon \Delta u + b \cdot \nabla u + cu - f, \, b \cdot \nabla v)_T,$$

but an inspection of how stabilization is achieved reveals that only the term

$$\sum_{T \in \mathcal{T}_h} \delta_T (b \cdot \nabla u, b \cdot \nabla v)_T \tag{3.84}$$

is responsible for the increased stability and consequent improved convergence properties. Thus it is natural to ask: in order to reduce the costs of assembling the discrete system, it is enough to add only a term like (3.84)? But with such a replacement, the consistency property of the method is lost. To retain the stability properties of the SDFEM, in the convection-dominated case choose $\delta_T = \mathcal{O}(h_T)$ in (3.84); then

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T (b \cdot \nabla u, b \cdot \nabla v)_T \right| \le C h^{1/2} |u|_1 \left(\sum_{T \in \mathcal{T}_h} \delta_T \| b \cdot \nabla v \|_{0,T}^2 \right)^{1/2}$$

shows that the consistency error is $\mathcal{O}(\sqrt{h}\,)$ and the method will be suboptimal. The remedy presented here is to introduce a projection $\pi_h : L_2(\Omega) \to D_h$ into a second finite element space D_h, then to replace $b \cdot \nabla u$ by its fluctuations $\kappa_h(b \cdot \nabla u)$, where $\kappa_h := \mathrm{id} - \pi_h$ with $\mathrm{id} : L_2(\Omega) \to L_2(\Omega)$ the identity

operator. The order of the consistency error can now be tuned by choosing an appropriate projection space D_h. Indeed, if π_h is the L_2 projection and D_h the space of discontinuous, piecewise polynomials of degree $k-1$ with $k \geq 1$, then

$$\|\kappa_h(b \cdot \nabla u)\|_{0,T} \leq Ch_T^k \|u\|_{k+1,T},$$

and for $\delta_T = \mathcal{O}(h_T)$ it follows that

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T (\kappa_h(b \cdot \nabla u), \kappa_h(b \cdot \nabla v))_T \right|$$

$$\leq Ch^{k+1/2} \|u\|_{k+1} \left(\sum_{T \in \mathcal{T}_h} \delta_T \|\kappa_h(b \cdot \nabla v)\|_{0,T}^2 \right)^{1/2}.$$

Later we shall learn that the $\mathcal{O}(h^{k+1/2})$ estimation of the convection term for an approximation space Y_h with piecewise polynomials of degree k (which is already known for the SDFEM) can be preserved if there is an interpolant $j_h : H^2(\Omega) \to Y_h$ such that $w - j_h w$ is orthogonal to D_h.

This local projection stabilization (LPS) method is introduced for the Stokes problem in [BB01], extended to the transport equation in [BB04], and analysed for the lowest order ($r \leq 2$) discretizations of the Oseen equations in [BB06]. In all these papers a two-level approach is used where the projection space D_h lives on a mesh that is coarser than the mesh used by the approximation space Y_h. This has the disadvantage that the LPS scheme produces a stencil that is less compact than for the SDFEM stabilization. To overcome this difficulty, an alternative technique based on enrichment of the approximation space Y_h is proposed in [MST07]. We shall explain both approaches in a unified framework.

In the following the notation $\alpha \sim \beta$ means that there exist positive constants C_1 and C_2, which are independent of the meshsize h and of ε, such that

$$C_1 \alpha \leq \beta \leq C_2 \alpha.$$

Let $\mathcal{M}_h$ be a shape-regular decomposition of Ω into d-dimensional simplices, quadrilaterals or hexahedra. Each cell $M \in \mathcal{M}_h$ is called a macro-element and its diameter is denoted by h_M. Each macro-element M will be decomposed into one or more cells $T \in \mathcal{T}_h$, such that $\mathcal{T}_h$ also is shape-regular – one could for example generate $\mathcal{T}_h$ from $\mathcal{M}_h$ by some refinement rule. Then the projection space D_h will be a discontinuous finite element space defined on the macro-decomposition $\mathcal{M}_h$ while the approximation space $Y_h \subset H^1(\Omega)$ comprises continuous piecewise polynomial functions defined on $\mathcal{T}_h$. The case $\mathcal{T}_h = \mathcal{M}_h$ is permitted. We assume that the partitions $\mathcal{T}_h$ and $\mathcal{M}_h$ satisfy

$$h_T \sim h_M \qquad \forall T \subset M, \quad \forall M \in \mathcal{M}_h.$$

Let $D_h(M) := \{q_h|_M : q_h \in D_h\}$ be the local projection space. Define the global projection $\pi_h : L_2(\Omega) \to D_h$ by $(\pi_h w)|_M := \pi_M(w|_M)$, where

$\pi_M : L_2(M) \to D_h(M)$ is a local projection. Associate with the projection π_h the fluctuation operator $\kappa_h : L_2(\Omega) \to L_2(\Omega)$ defined by $\kappa_h := \mathrm{id} - \pi_h$, where $\mathrm{id} : L_2(\Omega) \to L_2(\Omega)$ is the identity.

Now we are ready to formulate the local projection stabilization (LPS) method for the convection-diffusion-reaction problem

$$-\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega \subset \mathbb{R}^d, \quad u = 0 \quad \text{on } \Gamma, \tag{3.85}$$

where $\Gamma = \partial \Omega$, $d \geq 2$, the data b, c, f are sufficiently smooth, and $0 < \varepsilon \ll 1$ is a given small positive parameter. Assume that

$$c - \frac{1}{2} \operatorname{div} b \geq \omega > 0$$

which guarantees the unique solvability of the problem. Let $V_h = Y_h \cap H_0^1(\Omega)$ be the finite element space for approximating the weak solution $u \in H_0^1(\Omega)$ of (3.85). The corresponding stabilized discrete problem is:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ one has

$$\varepsilon(\nabla u_h, \nabla v_h) + (b \cdot \nabla u_h + cu_h, v_h) + S_h(u_h, v_h) = (f, v_h), \tag{3.86a}$$

where the stabilizing term S_h is given by

$$S_h(u_h, v_h) := \sum_{M \in \mathcal{M}_h} \tau_M \Big(\kappa_h(b \cdot \nabla)u_h, \kappa_h(b \cdot \nabla)v_h \Big)_M \tag{3.86b}$$

with user-chosen constants τ_M. Define the mesh-dependent norm

$$|||v|||_{LPS} := \left(\varepsilon |v|_1^2 + \omega \|v\|_0^2 + \sum_{M \in \mathcal{M}_h} \tau_M \|\kappa_h(b \cdot \nabla)v\|_{0,M}^2 \right)^{1/2} \tag{3.87}$$

associated with the discrete bilinear form implicitly defined by the left-hand side of (3.86a).

Remark 3.68. There is a close relation to stabilization by subgrid modelling [EG04, Gue99a], as we shall see in Section IV.4.5, but in subgrid modelling the stabilizing term uses gradients of fluctuations instead of fluctuations of gradients. ♣

The stability and convergence properties of the LPS method (3.86) will now be studied under the following assumptions.

Assumption A1: The approximation space Y_h is of order $r \in \mathbb{N}$. That is, there exists an interpolation operator $i_h : H^2(\Omega) \to Y_h$ with the properties that $i_h : H_0^1(\Omega) \cap H^2(\Omega) \to V_h$ and

$$\|w - i_h w\|_{0,T} + h_T |w - i_h w|_{1,T} \leq C\, h_T^l \|w\|_{l,T} \tag{3.88}$$

for all $w \in H^l(T)$, all $T \in \mathcal{T}_h$, and $2 \leq l \leq r + 1$.

Assumption A2: The fluctuation operator κ_h has the approximation property

$$\|\kappa_h q\|_{0,M} \leq C \, h_M^l |q|_{l,M} \qquad \forall q \in H^l(M), \ \forall M \in \mathcal{M}_h, \ 0 \leq l \leq r. \quad (3.89)$$

Remark 3.69. Let π_h be the L_2 projection in D_h and let the space $D_h(M)$ contain the space $P_{r-1}(M)$ of polynomials of degree at most $r - 1$, where $r \geq 1$. Since D_h is allowed to be discontinuous across macro-element faces, the projection $\pi_M : L_2(M) \to D_h(M)$ is defined locally by

$$(\pi_M w - w, w_h)_M = 0 \qquad \forall w_h \in D_h(M), \ w \in L_2(M).$$

Then the L_2 projection $\pi_M : L_2(M) \to D_h(M)$ reduces to the identity mapping on the subspace $P_{r-1}(M) \subset H^l(M)$, and the Bramble-Hilbert lemma gives the approximation property of Assumption A2. ♣

Let $Y_h(M) := \{w_h|_M \ : \ w_h \in Y_h\} \cap H_0^1(M)$.

Assumption A3: There exists a constant $\beta_1 > 0$ such that for all $h > 0$ and all $M \in \mathcal{M}_h$ one has

$$\inf_{q_h \in D_h(M)} \sup_{v_h \in Y_h(M)} \frac{(v_h, q_h)_M}{\|v_h\|_{0,M} \|q_h\|_{0,M}} \geq \beta_1 > 0. \quad (3.90)$$

Remark 3.70. To satisfy Assumption A3, clearly $Y_h(M)$ has to be sufficiently rich compared with $D_h(M)$. In particular, it is necessary that

$$\dim Y_h(M) \geq \dim D_h(M). \quad (3.91)$$

On the other hand one cannot choose $D_h(M)$ too small to satisfy Assumption A3 since Assumption A2 should also be met. Later we try to fulfill both requirements for a given approximation space Y_h on $\mathcal{T}_h$ by choosing the projection space D_h as a discontinuous finite element space on the coarser mesh $\mathcal{M}_h$, where the dimension of $D_h(M)$ is small enough to satisfy Assumption A3 yet big enough to fulfil Assumption A2. A different strategy is used in the one-level approach where both spaces are defined on the same mesh: $D_h(M)$ is chosen such that Assumption A2 holds, then $Y_h(M)$ is enriched by additional functions in order to verify Assumption A3. ♣

Theorem 3.71. *Let Assumptions A1 and A3 be satisfied. Then there is an interpolation operator $j_h : H^2(\Omega) \to Y_h$, with $j_h : H_0^1(\Omega) \cap H^2(\Omega) \to V_h$, that has the following orthogonality and approximation properties:*

$$(w - j_h w, q_h) = 0 \quad (3.92a)$$

for all $q_h \in D_h$ and all $w \in H^2(\Omega)$, and

$$\|w - j_h w\|_{0,M} + h_M |w - j_h w|_{1,M} \leq C \, h_M^l \|w\|_{l,M} \quad (3.92b)$$

for all $w \in H^l(\Omega)$ with $2 \leq l \leq r + 1$, and all $M \in \mathcal{M}_h$.

Proof. Let $D_h(M)'$ denote the dual space of $D_h(M)$. Define the continuous linear operator $B_h : Y_h(M) \to D_h(M)'$ by

$$\langle B_h v_h, q_h \rangle_{D_h(M)} := (v_h, q_h)_M \quad \forall v_h \in Y_h(M), \ q_h \in D_h(M).$$

Set

$$W_h(M) := \{v_h \in Y_h(M) \ : \ (v_h, q_h) = 0 \ \forall q_h \in D_h(M)\},$$

and let $W_h(M)^\perp$ be the L_2-orthogonal complement of $W_h(M)$ in $Y_h(M)$. By [GR86, Lemma I.4.1], B_h is an isomorphism from $W_h(M)^\perp$ onto $D_h(M)'$ with

$$\beta_1 \|v_h\|_{0,M} \le \|B_h v_h\|_{D_h(M)'} \qquad \forall v_h \in W_h(M)^\perp$$

if and only if Assumption A3 holds true. Now, given $w \in H^2(\Omega)$, the mapping

$$q_h \mapsto (w - i_h w, q_h)_M$$

is linear and continuous on $D_h(M)$; hence for each $w \in H^2(\Omega)$ there is a unique $z_h(w) \in W_h(M)^\perp$ such that

$$\langle B_h z_h(w), q_h \rangle_{D_h(M)} = (w - i_h w, q_h)_M \quad \forall q_h \in D_h(M),$$

$$\|z_h(w)\|_{0,M} \le \frac{1}{\beta_1} \sup_{q_h \in D_h(M)} \frac{\langle B_h(z_h(w)), q_h \rangle_{D_h(M)}}{\|q_h\|_{0,M}} .$$

The definition of $B_h : Y_h(M) \to D_h(M)'$ yields

$$(z_h(w), q_h)_M = (w - i_h w, q_h)_M \qquad \forall w \in H^2(\Omega), \forall q_h \in D_h(M), \quad (3.93a)$$

$$\|z_h(w)\|_{0,M} \le \frac{1}{\beta_1} \|w - i_h w\|_{0,M} \qquad \forall w \in H^2(\Omega). \quad (3.93b)$$

Set $j_h w\big|_M := i_h w\big|_M + z_h(w)$ for all $M \in \mathcal{M}_h$. Since $\bigoplus_{M \in \mathcal{M}_h} Y_h(M) \subset Y_h$, we then have a global interpolation operator $j_h : H^2(\Omega) \to Y_h$ such that

$$\|w - j_h w\|_{0,M} \le \left(1 + \frac{1}{\beta_1}\right) \|w - i_h w\|_{0,M} \le C \, h_M^l \|w\|_{l,M}$$

for all $M \in \mathcal{M}_h$, for all $w \in H^l(\Omega)$, $2 \le l \le r + 1$. That is, the L_2 approximation property of (3.92b) is verified.

The orthogonality property (3.92a) follows from (3.93a) and the definition of j_h. It remains to show the approximation property for the H^1 seminorm. To this end, apply an inverse inequality and (3.93b) to get

$$|z_h(w)|_{1,M} \le C h_M^{-1} \|z_h(w)\|_{0,M} \le C h_M^{-1} \|w - i_h w\|_{0,M} .$$

This inequality and the approximation property (3.88) then give

$$|w - j_h w|_{1,M} \le |w - i_h w|_{1,M} + |z_h(w)|_{1,M} \le C \, h_M^{l-1} \|w\|_{l,M} .$$

$\square$

Remark 3.72. Following the analysis of [Ste99] and assuming a family of macro-elements that are equivalent to a reference macro-element, Assumption A3 reduces to showing that

$$N_M := \{q_h \in D_h(M) \ : \ (q_h, v_h)_M = 0 \quad \forall v_h \in V_h(M)\} = \{0\}.$$

♣

Example 3.73. Consider the case $\mathcal{T}_h = \mathcal{M}_h$. Let the approximation space Y_h comprise continuous piecewise linear functions enriched element by element with the bubble function b_T that is the product of the barycentric coordinates. Let the projection space D_h be discontinuous piecewise constant functions on $\mathcal{T}_h$. The usual piecewise linear nodal interpolation i_h satisfies the approximation property of Assumption A1 with $r = 1$, but it fails to satisfy (3.92a). Since $D_h(T) = \operatorname{span}(1)$ and $Y_h(T) = \operatorname{span}(b_T)$, Assumption A3 can be established by transforming the integrals in (3.90) to a reference cell. Thus there does exist an interpolation operator $j_h : H^2(\Omega) \to Y_h$ with the properties (3.92). It is given explicitly by a local definition on each cell T:

$$(j_h w)|_T(p_i) = w(p_i) \text{ for all vertices } p_i \in T, \quad (j_h w, 1)_T = (w, 1)_T \quad \forall T \in \mathcal{T}_h.$$

♣

Theorem 3.74. *Let the data of the problem be sufficiently smooth. Let Assumptions A1–A3 be fulfilled. If $\tau_M \sim h_M$ for all $M \in \mathcal{M}_h$, then there is a positive constant C, which is independent of ε and the mesh, such that*

$$|||u - u_h|||_{LPS} \leq C \left(\varepsilon^{1/2} + h^{1/2}\right) h^r \|u\|_{r+1}.$$

Proof. The argument is standard: one demonstrates coercivity of the underlying discrete bilinear form

$$a_h(w, v) := \varepsilon(\nabla w, \nabla v) + (b \cdot \nabla w + cw, v) + S_h(w, v)$$

then estimates the approximation and consistency errors. Coercivity with respect to the $||| \cdot |||_{LPS}$ norm, i.e.,

$$a_h(v_h, v_h) \geq |||v_h|||^2_{LPS} \qquad \forall v_h \in V_h,$$

follows by integration by parts for all nonnegative τ_M. (This differs from the streamline diffusion method where an upper bound for δ_T is needed; compare the proof of Lemma 3.25.) Then for the interpolant $j_h u$ of the weak solution u of (3.85) and the solution u_h of the discrete problem (3.86) we have

$$|||j_h u - u_h|||^2_{LPS} \leq a_h(j_h u - u, j_h u - u_h) + a_h(u - u_h, j_h u - u_h)$$

whence

$$|||j_h u - u_h|||_{LPS} \leq \sup_{w_h \in V_h} \frac{a_h(j_h u - u, w_h)}{|||w_h|||_{LPS}} + \sup_{w_h \in V_h} \frac{a_h(u - u_h, w_h)}{|||w_h|||_{LPS}}.$$

The first term here is the approximation error, the second term the consistency error. (The consistency error of a consistent method is zero.)

The tricky part in the estimation of the approximation error is the convection term which is split into two terms:

$$(b \cdot \nabla(j_h u - u), w_h) = -(j_h u - u, b \cdot \nabla w_h) - (\text{div } b \, (j_h u - u), w_h)$$

using integration by parts. For the first term, use the orthogonality and approximation properties of the special interpolant and $\tau_M \sim h_M$ to get

$$\left|(j_h u - u, b \cdot \nabla w_h)\right| = \left|(j_h u - u, \kappa_h(b \cdot \nabla)w_h)\right|$$

$$\leq C \left(\sum_{M \in \mathcal{M}_h} \tau_M^{-1} h_M^{2r+2} |u|_{r+1,M}^2 \right)^{1/2} \left(\sum_{M \in \mathcal{M}_h} \tau_M \|\kappa_h(b \cdot \nabla)w_h\|_{0,M}^2 \right)^{1/2}$$

$$\leq C \, h^{r+1/2} |u|_{r+1} \, |||w_h|||_{LPS}.$$

The estimation of the second term uses the approximation properties and the definition of the $||| \cdot |||_{LPS}$ norm:

$$\left|(\text{div } b \, (j_h u - u), w_h)\right| \leq C h^{r+1} |u|_{r+1} \|w_h\|_0 \leq C h^{r+1} |u|_{r+1} \, |||w_h|||_{LPS}.$$

Using (3.89), $\tau_M \sim h_M$, and $a_h(u - u_h, w_h) = S_h(u, w_h)$, the consistency error bound follows from

$$|S_h(u, w_h)| \leq \sum_{M \in \mathcal{M}_h} \tau_M \|\kappa_h(b \cdot \nabla u)\|_{0,M} \, \|\kappa_h(b \cdot \nabla w_h)\|_{0,M}$$

$$\leq C \sum_{M \in \mathcal{M}_h} \tau_M h_M^r |b \cdot \nabla u|_{r,M} \, \|\kappa_h(b \cdot \nabla w_h)\|_{0,M}$$

$$\leq C h^{r+1/2} \|u\|_{r+1} |||w_h|||_{LPS}.$$

It is now straightforward to finish the proof.   ♣

Remark 3.75. An analogous theorem can be proved when the stabilizing term (3.86b) and the norm (3.87) are replaced by

$$S_h(u_h, v_h) := \sum_{M \in \mathcal{M}_h} \tau_M \left(\kappa_h(\nabla u_h), \kappa_h(\nabla v_h) \right)_M$$

and

$$|||v|||_{LPS} := \left(\varepsilon |v|_1^2 + \omega \|v\|_0^2 + \sum_{M \in \mathcal{M}_h} \tau_M \|\kappa_h(\nabla v)\|_{0,M}^2 \right)^{1/2}$$

respectively.   ♣

Fulfillment of Assumptions A1–A3 depends on the selections of the approximation space Y_h and the projection space D_h. Assumption A1 is satisfied for common finite element spaces that contain continuous piecewise polynomials of degree r. Assumption A2 can be easily satisfied by choosing the projection space D_h sufficiently large but Assumption A3 restricts the size of D_h for a given approximation space Y_h. Below we discuss examples of pairs of finite element spaces (Y_h, D_h) that satisfy Assumptions A1–A3 of Theorem 3.74 while referring the reader to [MST07] for the proofs.

Local Projection as a Two-level Approach

Consider the case where the partition $\mathcal{T}_h$ is formed by a suitable refinement of a macro-mesh $\mathcal{M}_h$. This is indicated by the notation $\mathcal{M}_h = \mathcal{T}_{2h}$. First we discuss simplicial elements in $\mathbb{R}^d$. A macro-element $M \in \mathcal{T}_{2h}$ is refined into $d + 1$ elements $T \in \mathcal{T}_h$ by connecting the $d + 1$ vertices of M with its barycentre; see Figure 3.8 for the cases $d = 2$ and $d = 3$. For the approximation
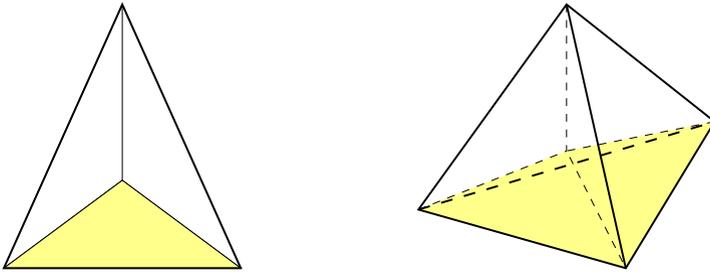
Fig. 3.8. Refinement of a macro-simplex $M \in \mathcal{T}_{2h}$ into cells $T \in \mathcal{T}_h$

space Y_h we choose a finite element space of continuous piecewise polynomials of degree $r \geq 1$. Let the projection space D_h comprise discontinuous piecewise polynomials of degree $r - 1$ on $\mathcal{T}_{2h}$. This is summarized by writing $(Y_h, D_h) = (P_{r,h}, P_{r-1,2h}^{\mathrm{disc}})$. Here and in what follows the superscript 'disc' indicates that the finite element space contains discontinuous functions. Then on shape-regular meshes Assumptions A1–A3 are satisfied [MST07].

Consider now hexahedral elements such as bricks. Let $\widehat{M} = (-1, 1)^d$ denote the reference hyper-cube with 2^d vertices. This is refined into 2^d congruent cubes $\widehat{T}_i$, where $i = 1, \ldots, 2^d$. The multilinear mapping $F_M : \widehat{M} \to M$ maps $\widehat{M}$ onto a macro-cell $M \in \mathcal{T}_{2h}$ and induces a refinement of M into 2^d cells $T_i = F_M(\widehat{T}_i)$; see Figure 3.9 for the two-dimensional case. Furthermore, there is a bijective linear mapping $G_i : \widehat{T} \to \widehat{T}_i$ of the reference cell $\widehat{T} = (0, 1)^d$ onto $\widehat{T}_i$ for $i = 1, \ldots, 2^d$. Now for each $T \in \mathcal{T}_h$ there are a unique $M \in \mathcal{M}_h$ and a unique $i \in \{1, \ldots, 4\}$ such that $T = T_i \subset M$ and $T = (F_M \circ G_i)(\widehat{T})$.
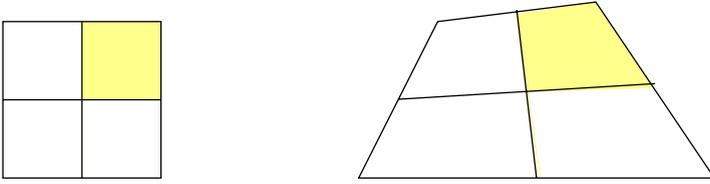
Fig. 3.9. Refinement of a macro-cell $M \in \mathcal{T}_{2h}$ (right) induced by a congruent refinement of the reference hyper-cube $\widehat{M}$ (left)

We write the bijective multilinear mapping $F_M \circ G_i$ as F_T for brevity. For the approximation space Y_h choose the standard space of mapped continuous piecewise polynomials of degree at most r in each variable, i.e., $Y_h = Q_{r,h}$. The projection space D_h lives on the coarser mesh $\mathcal{T}_{2h}$ and can be defined in two different ways, namely as an image of a space living on the reference macro-cell $\widehat{M}$ or directly on the macro-cell M. In general, this leads to two different finite element spaces. The mapped version of D_h has the advantage that the projection space defined locally on the reference macro-cell is always the same when moving from one element to another, but the approximation property of Assumption A2 is not satisfied on arbitrary families of shape-regular meshes [ABF02, Mat01]. This is apparently a great disadvantage but in practice the family of macro-element meshes is often generated by successively refining a given initial mesh, and for such a (restricted) mesh family Assumption A2 does hold true [Mat01]. The unmapped version of D_h satisfies Assumption A2 for any family of shape-regular meshes but the associated finite element spaces on the reference macro-cell differ from element to element. To distinguish between these two spaces we shall use the superscript 'unm' for the unmapped version of the finite element space D_h, with the understanding that all spaces lacking this superscript are mapped spaces.

The finite element pair $(Y_h, D_h) = (Q_{r,h}, Q_{r-1,2h}^{\text{disc}})$ is our first example on hexahedral meshes; here

$$Q_{r,h} := \{ v \in H^1(\Omega) \ : \ v|_T \circ F_T \in Q_r(\widehat{T}) \quad \forall T \in \mathcal{T}_h \},$$
$$Q_{r-1,2h}^{\text{disc}} := \{ v \in L_2(\Omega) \ : \ v|_M \circ F_M \in Q_{r-1}(\widehat{M}) \quad \forall M \in \mathcal{T}_{2h} \}.$$

Assumption A1 is clearly satisfied [Ape99, Clé75, SZ90]. Furthermore, since $P_{r-1}(M) \subset Q_{r-1}^{\text{disc}}(M)$, one can verify Assumption A2 on arbitrary shape-regular families of meshes. For the proof of Assumption A3 see [MST07].

Alternatively, one can choose a smaller projection space by taking D_h to be

$$P_{r-1,2h}^{\text{disc}} := \{ v \in L_2(\Omega) \ : \ v|_M \circ F_M \in P_{r-1}(\widehat{M}) \quad \forall M \in \mathcal{T}_{2h} \}.$$

This produces more stabilization in the sense that the stabilizing term vanishes on the smaller subset $P_{r-1,2h}^{\text{disc}} \subset Q_{r-1,2h}^{\text{disc}}$. Assumptions A1 and A3 are

still valid but Assumption A2 can be guaranteed only on a restricted family of shape-regular meshes, e.g., on uniformly-refined families of meshes; see [ABF02] for quadrilateral meshes and [Mat01] for hexahedral meshes.

One could also investigate a choice of projection space D_h that is larger than $Q_{r-1,2h}^{\mathrm{disc}}$ in order to minimize the stabilizing effect. Indeed, a dimensional analysis indicates that the inequality (3.91) is still satisfied for larger spaces D_h. For the choice $(Y_h, D_h) = (Q_{r,h}, Q_{r-1,2h}^{\mathrm{disc}})$ one has

$$\dim Y_h(M) = (2r-1)^d \geq r^d = \dim D_h(M)$$

and only for $r = 1$ do the dimensions of both spaces coincide. In the case $r \geq 2$ a possible choice might be $D_h = Q_{r,2h}^{\mathrm{disc}}$ since

$$\dim Y_h(M) = (2r-1)^d \geq (r+1)^d = \dim D_h(M), \qquad r \geq 2.$$

Now Assumption A1 still holds true without any change and Assumption A2 would be satisfied with a higher order of approximation than necessary. It is unclear however whether the inf-sup condition of Assumption A3 is valid.

Unmapped finite element spaces satisfy Assumption A2 on arbitrary shape-regular meshes. For example, take the approximation space to be again the space $Y_h = Q_{r,h}$ but for the projection space D_h select the space of discontinuous, piecewise polynomials of degree $r-1$ posed directly on the macro-cells $M \in \mathcal{M}_h$. That is, we choose

$$(Y_h, D_h) = (Q_{r,h}, P_{r-1,2h}^{\mathrm{disc,unm}})$$

where

$$Q_{r,h} := \{v \in H^1(\Omega) \,:\, v|_T \circ F_T \in Q_r(\widehat{T}) \quad \forall T \in \mathcal{T}_h\},$$
$$P_{r-1,2h}^{\mathrm{disc,unm}} := \{v \in L_2(\Omega) \,:\, v|_M \in P_{r-1}(M) \quad \forall M \in \mathcal{T}_{2h}\}.$$

Then Assumptions A1–A3 are satisfied on families of shape-regular meshes [MST07].

Local Projection by Enrichment of Approximation Spaces

One disadvantage of the local projection onto coarser meshes is that the support of the projected gradient $\kappa_h(b \cdot \nabla\varphi)$ of a basis function φ is in general larger than the support of $\nabla\varphi$, which leads to an increase in the stencil size that might not suit the data structure of an existing computer code. Bearing in mind that the key ingredient of the local projection method is the existence of an interpolation with the additional orthogonality property (3.92a), one can try to define the approximation and projection space on the same mesh $\mathcal{M}_h = \mathcal{T}_h$ and to satisfy Assumption A3 by an enrichment of the approximation space Y_h. This approach has been developed successfully in [MST07], as we now describe.

Use simplicial elements and set

$$\hat{b}(\hat{x}) := (d+1)^{d+1} \prod_{i=1}^{d+1} \hat{\lambda}_i(\hat{x}),$$

where $\hat{\lambda}_i$, $i = 1, \ldots, d+1$, are barycentric coordinates on $\widehat{T}$. This bubble function $\hat{b}$ takes the value 1 at the barycentre of the reference simplex $\widehat{T}$. Then define the enriched space of continuous piecewise polynomials of degree r by

$$P_r^{\text{bubble}}(\widehat{T}) := P_r(\widehat{T}) + \hat{b} \cdot P_{r-1}(\widehat{T}).$$

We choose the approximation and projection spaces

$$(Y_h, D_h) := (P_{r,h}^{\text{bubble}}, P_{r-1,h}^{\text{disc}})$$

to be the pair of finite element spaces defined via reference mappings by

$$P_{r,h}^{\text{bubble}} := \{v \in H^1(\Omega) : v|_T \circ F_T \in P_r^{\text{bubble}}(\widehat{T}) \quad \forall T \in \mathcal{T}_h\},$$
$$P_{r-1,h}^{\text{disc}} := \{v \in L_2(\Omega) : v|_T \circ F_T \in P_{r-1}(\widehat{T}) \quad \forall T \in \mathcal{T}_h\}.$$

Clearly Assumptions A1 and A2 are fulfilled. At first sight the enriched space seems large, but in fact

$$P_r(\widehat{T}) + \hat{b} \cdot P_{r-1}(\widehat{T}) = P_r(\widehat{T}) \oplus \left(\hat{b} \cdot \sum_{i=1}^{d} \tilde{P}_{r-i}(\widehat{T}) \right)$$

where

$$\tilde{P}_r(\widehat{T}) = \text{span} \left\{ \prod_{i=1}^{d} \hat{x}_i^{\alpha_i}, \quad \sum_{i=1}^{d} \alpha_i = r, \quad (\hat{x}_1, \ldots, \hat{x}_d) \in \widehat{K} \right\}.$$

The enrichment is minimal with respect to the required inequality (3.91). For, since the bubble part of the space $P_r(\widehat{T})$ is $\hat{b} \cdot P_{r-(d+1)}(\widehat{T})$, we have

$$\dim \widehat{Y}(\widehat{T}) = \binom{r - (d+1) + d}{d} + \sum_{i=1}^{d} \left[\binom{r-i+d}{d} - \binom{r-i+d-1}{d} \right]$$
$$= \binom{r-1}{d} + \binom{r-1+d}{d} - \binom{r-1}{d}$$
$$= \dim D_h(\widehat{T}).$$

When $(Y_h, D_h) := (P_{r,h}^{\text{bubble}}, P_{r-1,h}^{\text{disc}})$, Assumption A3 is satisfied [MST07].

If the mesh is quadrilateral or hexahedral, then the reference mapping $F_T : \widehat{T} \to T$ is in general no longer affine. Thus one has two different options for the projection space, the mapped version

$$P_{r-1,h}^{\mathrm{disc}} := \{v \in L_2(\Omega) : v|_T \circ F_T \in P_{r-1}(\widehat{T}) \quad \forall T \in \mathcal{T}_h\}$$

and the unmapped version

$$P_{r-1,h}^{\mathrm{disc,unm}} := \{v \in L_2(\Omega) : v|_T \in P_{r-1}(T) \quad \forall T \in \mathcal{T}_h\}.$$

To ensure the approximation property of Assumption A2 for the mapped version of the projection space, only families of uniformly-refined meshes will be considered [ABF02, Mat01]. For the unmapped version, Assumption A2 holds true on general shape-regular meshes. Choosing as approximation space $Y_h = Q_{r,h}$, i.e., the usual space of continuous piecewise mapped polynomials of degree at most r in each variable, one obtains the approximation property Assumption A1 but not the local inf-sup condition of Assumption A3. Therefore we search for suitable enrichments of the approximation space Y_h. Let

$$\hat{b}(\hat{x}) = \prod_{i=1}^{d}(1 - \hat{x}_i^2) \in Q_2(\widehat{T}), \quad \hat{x} = (\hat{x}_1, \ldots, \hat{x}_d) \in \widehat{T}, \quad d = 2, 3,$$

be a bubble function associated with the reference cell $\widehat{T} := (-1, 1)^d$. Our first enriched finite element space is

$$Q_r^{\mathrm{bubble,1}}(\widehat{T}) := Q_r(\widehat{T}) \oplus \mathrm{span}\left\{\hat{b}\,\hat{x}_i^{r-1} : i = 1, \ldots, d\right\}.$$

Select the finite element spaces

$$(Y_h, D_h) := (Q_{r,h}^{\mathrm{bubble,1}}, P_{r-1,h}^{\mathrm{disc}})$$

where

$$Q_{r,h}^{\mathrm{bubble,1}} := \{v \in H^1(\Omega) : v|_T \circ F_T \in Q_r^{\mathrm{bubble,1}}(\widehat{T}) \quad \forall T \in \mathcal{T}_h\}.$$

Note that in general $Q_{r,h}^{\mathrm{bubble,1}}$ and $P_{r-1,h}^{\mathrm{disc}}$ are not polynomial spaces. Since $Q_r(\widehat{T}) \subset Q_r^{\mathrm{bubble,1}}(\widehat{T})$, Assumption A1 is clearly satisfied. Assumption A2 holds on uniformly refined meshes – see [ABF02, Mat01]. For the proof of Assumption A3 we refer to [MST07].

Comparing the dimensions of the spaces $Y_h(T)$ and $D_h(T)$, one has

$$\dim \widehat{Y}(\widehat{T}) = (r-1)^d + d \geq \binom{r-1+d}{d} = \dim P_{r-1}(\widehat{T}) \quad \text{for all } r, d \in \mathbb{N}.$$

In particular the enrichment is minimal with respect to (3.91) for biquadratic and bicubic elements on quadrilaterals and for triquadratic elements on hexahedra.

Remark 3.76. It is remarkable that the space $Q_r^{\mathrm{bubble,1}}(\widehat{T})$ has, for all $r \geq 2$, precisely d basis functions more than $Q_r(\widehat{T})$. That is, the amount of enrichment is independent of the polynomial degree r. ♣

To satisfy Assumption A2 on arbitrary families of shape-regular (non-simplicial) meshes, we propose a second version of the enriched finite element space: set

$$Q_r^{\text{bubble},2}(\widehat{T}) := Q_r(\widehat{T}) + \hat{b} \cdot Q_{r-1}(\widehat{T})$$

with the bubble function $\hat{b} \in Q_2(\widehat{T})$ and use the mapped enriched space

$$Q_{r,h}^{\text{bubble},2} := \{ v \in H^1(\Omega) \ : \ v|_T \circ F_T \in Q_r^{\text{bubble},2}(\widehat{T}) \quad \forall T \in \mathcal{T}_h \} .$$

Thus

$$(Y_h, D_h) := (Q_{r,h}^{\text{bubble},2}, P_{r-1,h}^{\text{disc,unm}}).$$

and Assumptions A1–A3 are fulfilled [MST07].

Remark 3.77. The space $Q_{r,h}^{\text{bubble},2}$ is more enriched than the space $Q_{r,h}^{\text{bubble},1}$. Comparing the dimensions of the spaces $Y_h(T)$ and $D_h(T)$, one can surmise that the enriched space could be made smaller, but the validity of the local inf-sup condition Assumption A3 is then unresolved.   ♣

Relationship to the Streamline Diffusion Method (SDFEM)

In Section 3.2.3 we started from the standard Galerkin finite element method with piecewise linears enriched by bubble functions on simplices and showed that elimination of the bubble part yields the streamline diffusion finite element method [BBF93, BR94]. Moreover, the shape of the bubble defined the SD parameter uniquely, but the symmetric version of the bubble

$$b_T := \prod_{i=1}^{d+1} \lambda_i^T, \qquad \lambda_i^T \text{ barycentric coordinates of } T,$$

as we saw in Remark 3.65, generated the SD parameter for the diffusion-dominated instead of the convection-dominated case. Several ideas have been developed to overcome this problem, ranging from the pseudo-residual-free bubble to the residual-free bubble method, where the bubbles are local solutions of the problem under consideration.

Here we shall examine the idea of eliminating the bubble part from the local projection method (3.86) for enriched approximation spaces. In problem (3.85) assume that one has piecewise constant functions b and f, and $c \equiv 0$. As in Section 3.2.3 suppose that V_h consists of piecewise linear functions and enrich this space by a bubble space B_h defined by

$$B_h := \text{span} \{ b_T \ : \ T \in \mathcal{T}_h \} .$$

Consider the local projection method on the enriched space $V_h \oplus B_h$ where the projection space D_h is the space of discontinuous piecewise constant functions on a triangulation $\mathcal{T}_h$:

Find $u_h \in V_h \oplus B_h$ such that for all $v_h \in V_h \oplus B_h$ one has

$$\varepsilon(\nabla u_h, \nabla v_h) + (b \cdot \nabla u_h, v_h) + S_h(u_h, v_h) = (f, v_h). \tag{3.94}$$

Here the stabilizing term S_h is given by (3.86b) with $\mathcal{M}_h = \mathcal{T}_h$. The dimension of the corresponding algebraic system of equations can be reduced by static condensation of the bubble part of the solution. To do this, write the solution as $u_h = u_L + u_B$, with $u_L \in V_h$ and $u_B \in B_h$, and use the test functions $v_h = v_L \in V_h$ and $v_h = v_B \in B_h$. As ∇v_L is piecewise constant, we get $\kappa_h(b \cdot \nabla)v_L = 0$ for all $v_L \in V_h$. Moreover, element-by-element integration by parts shows that $(\nabla v_L, \nabla v_B) = 0$ for all $v_L \in V_h$, $v_B \in V_B$. Hence (3.94) can be reformulated as:

Find $u_L \in V_h$ and $u_B \in B_h$ such that for all $v_L \in V_h$ and all $v_B \in B_h$,

$$\varepsilon(\nabla u_L, \nabla v_L) + (b \cdot \nabla(u_L + u_B), v_L) = (f, v_L), \tag{3.95a}$$
$$\varepsilon(\nabla u_B, \nabla v_B) + (b \cdot \nabla(u_L + u_B), v_B) + S_h(u_B, v_B) = (f, v_B). \tag{3.95b}$$

Now from the representation $u_B = \sum_{T \in \mathcal{T}_h} d_T b_T$, where the d_T, $T \in \mathcal{T}_h$, are unknown constants, (3.95b) becomes:

Given $u_L \in V_h$, find $\{d_T \in \mathbb{R} : T \in \mathcal{T}_h\}$ such that for each $T \in \mathcal{T}_h$,

$$\varepsilon(\nabla d_T b_T, \nabla b_T)_T + (b \cdot \nabla(u_L + d_T b_T), b_T)_T$$
$$+ S_h(d_T b_T, b_T) = (f, b_T)_T. \tag{3.96}$$

An integration by parts gives, using $\langle \cdot, \cdot, \rangle$ to denote the $L_2(\Gamma)$ inner product,

$$d_T(b \cdot \nabla b_T, b_T)_T = \frac{d_T}{2} \langle b \cdot n, b_T^2 \rangle_{\partial T} = 0,$$
$$\pi_T(b \cdot \nabla)b_T = \frac{1}{|T|} b \cdot \int_T \nabla b_T \, dx = \frac{1}{|T|} b \cdot \int_{\partial T} b_T \, n \, d\gamma = 0$$

and (3.96) reduces to:

Given $u_L \in V_h$, find $\{d_T \in \mathbb{R} : T \in \mathcal{T}_h\}$ such that for each T,

$$d_T \left(\varepsilon |b_T|_{1,T}^2 + \tau_T \|b \cdot \nabla b_T\|_{0,T}^2 \right) = (f - b \cdot \nabla u_L, b_T)_T.$$

This has the solution

$$d_T = \frac{(1, b_T)_T}{\varepsilon |b_T|_{1,T}^2 + \tau_T \|b \cdot \nabla b_T\|_{0,T}^2} \left. (f - b \cdot \nabla u_L) \right|_T. \tag{3.97}$$

Then (3.95a) can be rewritten as

$$\varepsilon(\nabla u_L, \nabla v_L) + (b \cdot \nabla u_L, v_L) + \sum_{T \in \mathcal{T}_h} d_T(b \cdot \nabla b_T, v_L)_T = (f, v_L).$$

The term $\sum_{T \in \mathcal{T}_h} \cdots$ does not appear in the standard Galerkin finite element method applied on the space V_h. One can rearrange it as

$$\sum_{T \in \mathcal{T}_h} d_T (b \cdot \nabla b_T, v_L)_T = - \sum_{T \in \mathcal{T}_h} d_T (b \cdot \nabla v_L, b_T)_T$$

$$= \sum_{T \in \mathcal{T}_h} \gamma_T (b \cdot \nabla u_L - f, b \cdot \nabla v_L)_T,$$

where, using (3.97), one sees that

$$\gamma_T = \frac{1}{|T|} \frac{|(1, b_T)_T|^2}{\varepsilon |b_T|_{1,T}^2 + \tau_T \|b \cdot \nabla b_T\|_{0,T}^2} \ . \qquad (3.98)$$

We have now eliminated the bubble component from (3.94), arriving at

$$\varepsilon (\nabla u_L, \nabla v_L) + (b \cdot \nabla u_L, v_L) + \sum_{T \in \mathcal{T}_h} \gamma_T (b \cdot \nabla u_L, b \cdot \nabla v_L)_T$$

$$= (f, v_L) + \sum_{T \in \mathcal{T}_h} \gamma_T (f, b \cdot \nabla v_L)_T \quad \text{for all } v_L \in V_h.$$

This is the streamline diffusion method (3.36) with the SD parameter $\delta_T \equiv \gamma_T$ given by (3.98). A scaling argument shows that $(1, b_T) \sim |T|$, $|b_T|_{1,T}^2 \sim |T|/h_T^2$, and $\|b \cdot \nabla b_T\|_{0,T}^2 \sim |T| \|b\|^2/h_T^2$, so $\gamma_T \sim h_T^2/(\varepsilon + \tau_T \|b\|^2)$. For $\tau_T = 0$ one has $\gamma_T \sim h_T^2/\varepsilon$ which corresponds to the diffusion-dominated case. Clearly γ_T is decreasing for increasing τ_T. The choice $\gamma_T \sim h_T/\|b\|$ in the convection-dominated case $\|b\| h_T/\varepsilon \gg 1$ corresponds to $\tau_T \sim h_T/\|b\|$. Letting $\tau_T \to \infty$, we obtain the standard Galerkin method that corresponds to $\gamma_T = 0$.

Comparing the residual-free bubble method with the local projection methods applied to the model problem (piecewise constant b and f, $c \equiv 0$), we see that via static condensation both methods recover the streamline diffusion method. But to generate the correct SD parameter, the RFB method needs to solve (at least approximately) local subproblems to find the correct bubble functions whereas for LPS the use of the simple bubble function $b_T = \prod_{i=1}^{d+1} \lambda_i^T$ suffices.

3.3.2 Continuous Interior Penalty Stabilization

Now we move on to the continuous interior penalty (CIP) stabilization method for the convection-diffusion problem

$$-\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \Gamma, \qquad (3.99)$$

where $\Gamma = \partial \Omega$, $\Omega \subset \mathbb{R}^d$ with $d = 2$ or 3, the data b, c, f are sufficiently smooth, and $0 < \varepsilon \ll 1$ is a given small positive parameter. Assume as usual that

$$c - \frac{1}{2} \operatorname{div} b \geq \omega > 0,$$

which guarantees existence and uniqueness of a solution to (3.99). Let $\mathcal{T}_h$ be a shape-regular triangulation of the domain Ω into cells $T \in \mathcal{T}_h$ with $\mathcal{E}_h$ the

set of all inner edges (faces in the three-dimensional case). Let $Y_h \subset H^1(\Omega)$ be a finite element space of piecewise polynomials of degree $r \geq 1$.

In the continuous interior penalty stabilization method, a symmetric term will be added to the Galerkin finite element discretization. Unlike other stabilization methods the Dirichlet boundary conditions are not incorporated into the finite element space Y_h but are imposed weakly on the discrete problem. We first discuss how Dirichlet-type boundary condition are implemented in a weak sense and address the CIP stabilization later.

Multiplying the differential equation $-\varepsilon\Delta u + b \cdot \nabla u + cu = f$ by a test function v, integrating over Ω and integrating by parts, we get

$$\varepsilon(\nabla u, \nabla v) + (b \cdot \nabla u + cu, v) - \varepsilon\left\langle \frac{\partial u}{\partial n}, v \right\rangle_\Gamma = (f, v)$$

where $\langle \cdot, \cdot \rangle_\Gamma$ denotes the inner product in $L_2(\Gamma)$. To obtain a lower bound like

$$(b \cdot \nabla v + cv, v) \geq \omega\|v\|_0^2 \qquad \forall v \in H_0^1(\Omega)$$

on the larger space $H^1(\Omega)$, subtract the term $\langle b \cdot n\, u, v \rangle_{\Gamma_-}$, which vanishes for $u \in H_0^1(\Omega)$ but not for $u \in H^1(\Omega)$. Here $\Gamma_- = \{x \in \Gamma : (b \cdot n)(x) < 0\}$ is the inflow part of the boundary. Then

$$(b \cdot \nabla v + cv, v) - \langle b \cdot n\, v, v \rangle_{\Gamma_-}$$

$$= \left(c - \frac{1}{2}\operatorname{div} b, v^2\right) + \frac{1}{2}\langle b \cdot n, v^2 \rangle_\Gamma - \langle b \cdot n, v^2 \rangle_{\Gamma_-}$$

$$\geq \omega\|v\|_0^2 + \frac{1}{2}\| |b \cdot n|^{1/2} v\|_{0,\Gamma}^2.$$

Furthermore, we add the term $\varepsilon\langle u, \frac{\partial v}{\partial n}\rangle_\Gamma$ to preserve the symmetry on $H^1(\Omega)$ of the diffusion term contribution and also add a penalty term to ensure coercivity. Then the statement of the *standard Galerkin method with weakly imposed boundary conditions* is:

Find $u_h \in Y_h$ such that for all $v_h \in Y_h$ one has

$$a_h(u_h, v_h) = (f, v_h)$$

where

$$a_h(u, v) = \varepsilon(\nabla u, \nabla v) + (b \cdot \nabla u + cu, v) - \varepsilon\left\langle \frac{\partial u}{\partial n}, v \right\rangle_\Gamma$$

$$- \varepsilon\left\langle u, \frac{\partial v}{\partial n} \right\rangle_\Gamma - \langle b \cdot nu, v \rangle_{\Gamma_-} + \sum_{E \subset \Gamma} \frac{\varepsilon\gamma}{h_E}\langle u, v \rangle_E. \qquad (3.100)$$

Lemma 3.78. *For all $v_h \in Y_h$, the bilinear form a_h given in (3.100) satisfies*

$$a_h(v_h, v_h) \geq \frac{1}{2}\left(\varepsilon|v_h|_1^2 + \omega\|v_h\|_0^2 + \| |b \cdot n|^{1/2} v\|_{0,\Gamma}^2 + \sum_{E \subset \Gamma} \frac{\varepsilon}{h_E}\|v_h\|_{0,E}^2 \right)$$

provided that $\gamma \geq \gamma_0$ where γ_0 is sufficiently large (independently of ε and h).

Proof. It is already clear that

$$a_h(v_h, v_h) \geq \varepsilon |v_h|_1^2 + \omega \|v_h\|_0^2 - 2\varepsilon \left\langle \frac{\partial v_h}{\partial n}, v_h \right\rangle_\Gamma$$

$$+ \frac{1}{2} \| \, |b \cdot n|^{1/2} \, v\|_{0,\Gamma}^2 + \varepsilon \gamma \sum_{E \subset \Gamma} \frac{1}{h_E} \|v_h\|_{0,E}^2.$$

For $E \subset \partial T$, the Cauchy-Schwarz inequality and a trace inequality yield

$$2\varepsilon \left| \left\langle \frac{\partial v_h}{\partial n}, v_h \right\rangle_E \right| \leq 2\varepsilon \, C h_E^{-1/2} |v_h|_{1,T} \|v_h\|_{0,E} \leq \frac{\varepsilon}{2} |v_h|_{1,T}^2 + \frac{2\varepsilon C^2}{h_E} \|v_h\|_{0,E}^2.$$

Summing over all edges (faces) $E \subset \Gamma$ and taking $\gamma \geq 1/2 + 2C^2$ gives the desired result. □

The above derivation of the bilinear form a_h shows that the standard Galerkin method with weakly imposed boundary condition is consistent, i.e., for a solution $u \in H_0^1(\Omega) \cap H^2(\Omega)$ of (3.99) one has

$$a_h(u, v_h) = (f, v_h) \qquad \forall v_h \in Y_h.$$

The CIP stabilized discrete problem is now defined to be:

Find $u_h \in Y_h$ such that for all $v_h \in Y_h$ one has

$$a_h(u_h, v_h) + J_h(u_h, v_h) = (f, v_h), \tag{3.101a}$$

where the stabilizing term J_h has the form

$$J_h(u, v) := \sum_{E \in \mathcal{E}_h} \tau_E \left\langle b_h \cdot [\nabla u]_E, b_h \cdot [\nabla v]_E \right\rangle_E. \tag{3.101b}$$

Here for each $E \in \mathcal{E}_h$ the τ_E are user-chosen parameters, $[w]_E$ is the jump of w across $E \in \mathcal{E}_h$ in a fixed direction n_E, i.e.,

$$([w]_E)(x) = \lim_{t \to +0} \{w(x + tn_E) - w(x - tn_E)\} \quad \text{for } x \in E,$$

and b_h is a continuous piecewise linear approximation of b that satisfies

$$\|b - b_h\|_{0,\infty,T} \leq C h_T \|b\|_{1,\infty,T}.$$

The form of the stabilizing term means that CIP stabilization is also called *edge stabilization*. For $u \in H^2(\Omega)$ one has $[u]_E = 0$ for all $E \in \mathcal{E}_h$ so CIP stabilization is consistent and enjoys the Galerkin orthogonality property.

Remark 3.79. Modifications of the stabilizing term are possible; see [BH04, Bur05, BFH06]   ♣

A discrete bilinear form is associated with the left-hand side of (3.101a) in the usual way. To analyse this bilinear form we introduce the mesh-dependent norm

$$|||v|||_{CIP} := \left(\varepsilon|v|_1^2 + \omega\|v\|_0^2 + J_h(v,v) + \||b\cdot n|^{1/2}v\|_{0,\Gamma}^2 + \sum_{E\subset\Gamma} \frac{\varepsilon}{h_E}\|v\|_{0,E}^2 \right)^{1/2}.$$

Let

$$H^2(\mathcal{T}_h) := \{v : \Omega \to \mathbb{R} : v|_T \in H^2(T) \quad \forall T \in \mathcal{T}_h\}$$

be the space of piecewise H^2 functions. Then the key step in analysing the CIP stabilization is the following lemma.

Lemma 3.80. *There exists an interpolation operator $\pi_h^* : H^2(\mathcal{T}_h) \to Y_h$ and a positive constant C (independent of the mesh size) such that for all $v_h \in Y_h$ and all $T \in \mathcal{T}_h$ one has*

$$h_T \|b_h \cdot \nabla v_h - \pi_h^*(b_h \cdot \nabla v_h)\|_{0,T}^2 \le C \sum_{E\in\mathcal{E}_h(T)} \int_E h_E^2 |b_h \cdot [\nabla v_h]_E|^2 \, d\gamma, \quad (3.102)$$

where $\mathcal{E}_h(T) := \{E \in \mathcal{E}_h : E \cap T \ne \emptyset\}$.

Proof. Let $\mathcal{N}$ be the set of all nodes, i.e., those points p_i that are associated with the degrees of freedom $v_h(p_i)$ of Y_h. Thus each $v_h \in Y_h$ is uniquely defined by prescribing its values $v_h(p_i)$ for all $p_i \in \mathcal{N}$. For each node $p_i \in \mathcal{N}$, let m_i be the number of cells that contain p_i as a node. If $m_i = 1$ we call p_i an *inner node* – so a point $p_i \in \Gamma \cap \mathcal{N}$ that does not lie on an intersection of mesh lines is an 'inner' node. As in [Osw91, Sch00, Bur05, BFH06] introduce the quasi-interpolant $\pi_h^* v \in Y_h$ defined by

$$(\pi_h^* v)(p_i) := \frac{1}{m_i} \sum_{\{T:p_i\in T\}} v|_T(p_i) \qquad v \in H^2(\mathcal{T}_h).$$

Choose a discontinuous piecewise polynomial function Φ by setting

$$\Phi|_T := \Phi_T = (b_h \cdot \nabla v_h - \pi_h^*(b_h \cdot \nabla v_h))\Big|_T \in P_r(T).$$

Then $\Phi_T(p_j) = 0$ at all inner nodes p_j of T, owing to the definition of π_h^*. Hence, applying the norm equivalence of finite-dimensional spaces on the reference cell and using the scaling property, for shape-regular meshes one gets

$$\|\Phi_T\|_{0,T} \le C h_T^{1/2} \|\Phi_T\|_{0,\partial T} \qquad \forall T \in \mathcal{T}_h.$$

Next, define the (scaled) ℓ_1 norm of each $q_h \in P_r(E)$ by

$$\|q_h\|_{\ell_1,E} := |E|^{1/2} \sum_{\{j:p_j\in E\}} |q_h(p_j)|.$$

Appealing to norm equivalence on a reference edge (face), we find that there are positive constants C_1 and C_2 such that

$$C_1 \|q_h\|_{0,E} \leq \|q_h\|_{\ell_1,E} \leq C_2 \|q_h\|_{0,E} \qquad \forall q_h \in P_r(E), \; \forall E \in \mathcal{E}_h.$$

The continuity of b_h and the definition of the quasi-interpolant π_h^* imply that for all nodes $p_j \in E \subset \partial T$ we have

$$\Phi_T(p_j) = \frac{1}{m_j} \sum_{\{T':p_j \in T'\}} b_h(p_j) \cdot \left(\nabla v_h\big|_T(p_j) - \nabla v_h\big|_{T'}(p_j)\right),$$

so

$$|\Phi_T(p_j)| \leq \frac{1}{m_j} \sum_{\{T':p_j \in T'\}} \sum_{E' \in P(T,T')} |b_h(p_j) \cdot [\nabla v_h]_{E'}(p_j)|,$$

where $P(T,T')$ denotes the set of all edges (faces) between T and T' (the shortest path); see Figure 3.10. If there are two paths with the same number of edges, choose one of them to make the definition of $P(T,T')$ unique. On
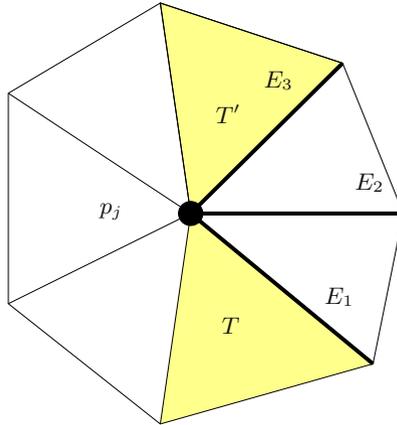
Fig. 3.10. Set of all edges E belonging to the shortest path $P(T,T') = \{E_1, E_2, E_3\}$

the skeleton $\mathcal{E}_h$ define the piecewise polynomial function

$$\Psi_E := b_h \cdot [\nabla v_h]_E \in P_r(E) \qquad \forall E \in \mathcal{E}_h$$

and denote the subset of edges containing the node p_j by

$$\mathcal{E}_{h,j} := \{E \in \mathcal{E}_h : p_j \in E\}.$$

The assumption that the family of meshes is shape-regular ensures that the sets $\mathcal{E}_{h,j}$ and $\mathcal{E}_h(T)$ each contain a bounded number of edges (faces). Moreover, $h_E \sim h_T$ for all $E \in \mathcal{E}_h(T)$ and $|E'| \sim |E|$ for all $E', E \in \mathcal{E}_h(T)$. Since

$$|\Phi_T(p_j)| \leq C \sum_{E' \in \mathcal{E}_{h,j}} |\Psi_{E'}(p_j)|$$

one obtains the estimate

$$\|\Phi_T\|_{\ell_1,E} \leq C \sum_{E' \in \mathcal{E}_h(T)} \|\Psi_{E'}\|_{\ell_1,E'} \qquad \forall E \subset \partial T.$$

Collecting the various inequalities, for each $T \in \mathcal{T}_h$ we deduce that

$$h_T \|\Phi_T\|_{0,T}^2 \leq Ch_T^2 \sum_{E \subset \partial T} \|\Phi_T\|_{0,E}^2 \leq Ch_T^2 \sum_{E \subset \partial T} \|\Phi_T\|_{\ell_1,E}^2$$

$$\leq Ch_T^2 \left(\sum_{E' \in \mathcal{E}_h(T)} \|\Psi_{E'}\|_{\ell_1,E'} \right)^2$$

$$\leq C \sum_{E' \in \mathcal{E}_h(T)} h_{E'}^2 \|\Psi_{E'}\|_{\ell_1,E'}^2 \leq C \sum_{E' \in \mathcal{E}_h(T)} h_{E'}^2 \|\Psi_{E'}\|_{0,E'}^2$$

where the inequality $(\sum a_i)^2 \leq C \sum a_i^2$ – valid for a bounded number of summands – was used. Recalling the definitions of Φ_T and $\Psi_{E'}$, the proof is complete. $\square$

Remark 3.81. It can be shown (see for example [BFH06]) that a positive constant C^* exists such that the lower bound

$$C^* \sum_{E \in \mathcal{E}_h(T)} \int_E h_E^2 |b_h \cdot [\nabla v_h]_E|^2 \, d\gamma \leq h_T \|b_h \nabla \cdot v_h - \pi_h^*(b_h \nabla \cdot v_h)\|_{0,T}^2$$

also holds true. Summing this inequality and (3.102) over T we get

$$C_1 J_h(v_h, v_h) \leq \sum_{T \in \mathcal{T}_h} h_T \|b_h \nabla \cdot v_h - \pi_h^*(b_h \nabla \cdot v_h)\|_{0,T}^2 \leq C_2 J_h(v_h, v_h)$$

when the parameter in (3.101b) is chosen so that $\tau_E \sim h_E^2$. ♣

Remark 3.82. In local projection stabilization we added a stabilizing term of the form
$$S_h(u_h, v_h) = \sum_{T \in \mathcal{T}_h} \tau_T(\kappa_h(b_h \cdot \nabla u_h), \kappa_h(b_h \cdot \nabla v_h))_T$$

where $\kappa_h = \mathrm{id} - \pi_h$ is the fluctuation operator and π_h a local projection onto the (discontinuous) projection space D_h. If π_h is replaced by the quasi-interpolant $\pi_h^* : H^2(\mathcal{T}_h) \to Y_h$, then Lemma 3.80 enables us to replace the stabilizing term $S_h(\cdot,\cdot)$ on the discrete space Y_h by the stabilizing term

$$J_h(u_h, v_h) = \sum_{E \in \mathcal{E}_h} \tau_E \langle b_h \cdot [\nabla u_h]_E, b_h \cdot [\nabla v_h]_E \rangle_E .$$

The advantage of this replacement is the consistency of the CIP stabilization method; the LPS method is not consistent. ♣

Before investigating the convergence properties of the CIP stabilization method we describe the approximation properties of the global L_2 projection $i_h : L_2(\Omega) \to Y_h$.

Lemma 3.83. *The L_2 projection $i_h : L_2(\Omega) \to Y_h$ satisfies the global approximation properties*

$$\sum_{T \in \mathcal{T}_h} h_T^{2m} |u - i_h u|_{m,T}^2 \le C \sum_{T \in \mathcal{T}_h} h_T^{2r+2} |u|_{r+1,T}^2 \qquad \forall u \in H^{r+1}(\Omega),$$

$$\sum_{E \subset \Gamma} h_E |u - i_h u|_{0,E}^2 \le C \sum_{T \in \mathcal{T}_h} h_T^{2r+2} |u|_{r+1,T}^2 \qquad \forall u \in H^{r+1}(\Omega),$$

on shape-regular meshes $\mathcal{T}_h$ where $0 \le m \le r + 1$ with $r \ge 1$.

Proof. Let $u^I \in Y_h$, $u \in H^2(\Omega)$, be the usual nodal interpolant that satisfies

$$h_T^m |u - u^I|_{m,T} \le C h_T^{r+1} |u|_{r+1,T} \qquad \forall u \in H^{r+1}(T)$$

where $0 \le m \le r + 1$ and $r \ge 1$. Applying the Cauchy-Schwarz inequality to $(u - i_h u, u - i_h u) = (u - i_h u, u - u^I)$ yields the $L_2(\Omega)$ estimate

$$\|u - i_h u\|_0 \le \|u - u^I\|_0 \le C \sum_{T \in \mathcal{T}_h} h_T^{2r+2} |u|_{r+1,T}^2.$$

Estimates for the derivatives can then be deduced via a triangle inequality and an inverse estimate:

$$\begin{aligned}
h_T^m |u - i_h u|_{m,T} &\le h_T^m |u - u^I|_{m,T} + h_T^m |u^I - i_h u|_{m,T} \\
&\le C h_T^{r+1} |u|_{r+1,T} + C \|u^I - u\|_{0,T} + C \|u - i_h u\|_{0,T} \\
&\le C h_T^{r+1} |u|_{r+1,T} + C \|u - i_h u\|_{0,T}.
\end{aligned}$$

Squaring, summing and applying the above L_2 bound, we get the first of the desired estimates. For the second, a scaled version of a trace theorem gives

$$h_E^{1/2} \|v\|_{0,E} \le C(\|v\|_{0,T} + h_T |v|_{1,T}) \qquad \text{for all } v \in H^1(T). \tag{3.103}$$

Again square, sum, and apply the global L_2 and H^1 bounds.   □

Remark 3.84. Lemma 3.83 does *not* imply that

$$\|u - i_h u\|_m^2 \le C \sum_{T \in \mathcal{T}_h} h_T^{2(r-m+1)} |u|_{r+1,T}^2 \qquad \forall u \in H^{r+1}(\Omega) \qquad (3.104\text{a})$$

$$\|u - i_h u\|_{0,\Gamma}^2 \le C \sum_{T \in \mathcal{T}_h} h_T^{2r+1} |u|_{r+1,T}^2 \qquad \forall u \in H^{r+1}(\Omega) \qquad (3.104\text{b})$$

for $m = 1, \ldots, r+1$, but on quasi-uniform meshes where $ch \le h_T \le h$ these inequalities are valid. ♣

Remark 3.85. Assume that the L_2 projection i_h is H^m stable, i.e.,

$$\|i_h u\|_m \le C_S \|u\|_m \qquad \forall u \in H^m(\Omega).$$

Then for the nodal interpolant u^I one has

$$\|u - i_h u\|_m \le \|u - u^I\|_m + \|i_h(u^I) - i_h u\|_m \le (1 + C_S)\|u - u^I\|_m$$

and (3.104a) follows. The L_2 projection is H^1 stable on quasi-uniform meshes and in [BPS01] this stability has been proved for the more general case of shape-regular meshes that satisfy a certain mesh condition. ♣

Theorem 3.86. *Let the data of the problem be sufficiently smooth, let γ be sufficiently large and assume that $\tau_E \sim h_E^2$. Then there is a positive constant C, which is independent of ε and the mesh, such that on quasi-uniform meshes one has*

$$|||u - u_h|||_{CIP} \le C\,(\varepsilon^{1/2} + h^{1/2})h^r \|u\|_{r+1}\,.$$

Proof. The proof follows a familiar pattern: demonstrate the coercivity of the underlying discrete bilinear form on Y_h with respect to the norm $||| \cdot |||_{CIP}$ then estimate the approximation error. By Lemma 3.78 one has

$$a_h(v_h, v_h) + J_h(v_h, v_h) \ge \frac{1}{2} |||v_h|||_{CIP}^2 \qquad \forall v_h \in Y_h,$$

for all nonnegative τ_E and γ large enough. Then, for any interpolant $i_h u \in Y_h$ of the weak solution u, with u_h the solution of the discrete problem, we get

$$\frac{1}{2} |||u_h - i_h u|||_{CIP}^2 \le a_h(u - i_h u, u_h - i_h u) + J_h(u - i_h u, u_h - i_h u)$$

whence

$$|||u_h - i_h u|||_{CIP} \le 2 \sup_{w_h \in Y_h} \frac{a_h(u - i_h u, w_h)}{|||w_h|||_{CIP}} + 2 \sup_{w_h \in Y_h} \frac{J_h(u - i_h u, w_h)}{|||w_h|||_{CIP}}\,.$$

Consider the individual terms in $a_h(u - i_h u, w_h)$ for all $w_h \in Y_h$. Integration by parts of the convection term gives

$$a_h(u - i_h u, w_h) = \varepsilon \left(\nabla(u - i_h u), \nabla w_h \right) + ((c - \operatorname{div} b)(u - i_h u), w_h)$$
$$+ \langle b \cdot n \, (u - i_h u), w_h \rangle_{\Gamma_+} - (u - i_h u, b \cdot \nabla w_h)$$
$$- \varepsilon \left\langle \frac{\partial(u - i_h u)}{\partial n}, w_h \right\rangle_\Gamma - \varepsilon \left\langle u - i_h u, \frac{\partial w_h}{\partial n} \right\rangle_\Gamma$$
$$+ \sum_{E \subset \Gamma} \frac{\varepsilon \gamma}{h_E} \langle u - i_h u, w_h \rangle_E. \tag{3.105}$$

Here the fourth term is the most troublesome and we estimate it first. Adding and subtracting $b_h \cdot \nabla w_h$ gives

$$(u - i_h u, b \cdot \nabla w_h) = (u - i_h u, (b - b_h) \cdot \nabla w_h) + (u - i_h u, b_h \cdot \nabla w_h);$$

for the first term here an inverse inequality gives

$$|(u - i_h u, (b - b_h) \cdot \nabla w_h)| \le C \sum_{T \in \mathcal{T}_h} \|u - i_h u\|_{0,T} \, h_T \, |w_h|_{1,T}$$
$$\le C h^{r+1} \|u\|_{r+1} \, |||w_h|||_{CIP},$$

while for the second term choose i_h to be the global L_2 projection in Y_h, and then the orthogonality of $u - i_h u$ with respect to Y_h and Lemma 3.80 imply that

$$|(u - i_h u, b_h \cdot \nabla w_h)| = |(u - i_h u, b_h \cdot \nabla w_h - \pi_h^*(b_h \cdot \nabla w_h))|$$
$$\le C \left(\sum_{T \in \mathcal{T}_h} h_T^{-1} \|u - i_h u\|_{0,T}^2 \right)^{1/2} |||w_h|||_{CIP}$$
$$\le C h^{r+1/2} \|u\|_{r+1} \, |||w_h|||_{CIP}.$$

The other terms in (3.105) are bounded by means of standard arguments:

$$\varepsilon |(\nabla(u - i_h u), \nabla w_h)| \le C \varepsilon^{1/2} h^r \|u\|_{r+1} \, |||w_h|||_{CIP},$$
$$|((c - \nabla \cdot b)(u - i_h u), w_h)| \le C h^{r+1} \|u\|_{r+1} \, |||w_h|||_{CIP},$$
$$|\langle |b \cdot n|(u - i_h u), w_h \rangle_{\Gamma_+}| \le C h^{r+1/2} \|u\|_{r+1} \, |||w_h|||_{CIP},$$

where the scaled trace inequality (3.103) was used in deriving the last estimate. The Cauchy-Schwarz inequality shows that

$$\left| \varepsilon \left\langle \frac{\partial(u - i_h u)}{\partial n}, w_h \right\rangle_\Gamma \right| \le C \varepsilon^{1/2} \left(\sum_{E \subset \Gamma} h_E \left\| \frac{\partial(u - i_h u)}{\partial n} \right\|_{0,E}^2 \right)^{1/2} |||w_h|||_{CIP}.$$

An invocation of the scaled trace inequality (3.103) gives

$$h_E^{1/2} \left\| \frac{\partial(u - i_h u)}{\partial n} \right\|_{0,E} \le C \left(|u - i_h u|_{1,T} + h_T |u - i_h u|_{2,T} \right) \qquad \forall E \subset \partial T;$$

squaring then summing, we get

$$\left| \varepsilon \left\langle \frac{\partial (u - i_h u)}{\partial n}, w_h \right\rangle_\Gamma \right| \leq C \varepsilon^{1/2} h^r \|u\|_{r+1} \, |||w_h|||_{CIP}.$$

For the penultimate term in (3.105) one proceeds similarly, using an inverse inequality:

$$h_E^{1/2} \left\| \frac{\partial w_h}{\partial n} \right\|_{0,E} \leq C \left(|w_h|_{1,T} + h_T |w_h|_{2,T} \right) \leq C |w_h|_{1,T} \qquad \forall E \subset \partial T.$$

It follows that

$$\left| \varepsilon \left\langle (u - i_h u), \frac{\partial w_h}{\partial n} \right\rangle_\Gamma \right| \leq \sum_{E \subset \Gamma} \left(\frac{\varepsilon}{h_E} \right)^{1/2} \|u - i_h u\|_{0,E} \, (\varepsilon h_E)^{1/2} \left\| \frac{\partial w_h}{\partial n} \right\|_{0,E}$$

$$\leq \left(\sum_{E \subset \Gamma} \frac{\varepsilon}{h_E} \|u - i_h u\|_{0,E}^2 \right)^{1/2} \left(\varepsilon \sum_{T \in \mathcal{T}_h} |w_h|_{1,T}^2 \right)^{1/2}$$

$$\leq C \varepsilon^{1/2} h^r \|u\|_{r+1} \, |||w_h|||_{CIP}.$$

The final term in (3.105) is handled by a Cauchy-Schwarz inequality, obtaining

$$\left| \sum_{E \subset \Gamma} \frac{\varepsilon \gamma}{h_E} \langle (u - i_h u, w_h)_E \right| \leq \gamma \left(\sum_{E \subset \Gamma} \frac{\varepsilon}{h_E} \|u - i_h u\|_{0,E}^2 \right)^{1/2} |||w_h|||_{CIP}$$

$$\leq C \varepsilon^{1/2} h^r \|u\|_{r+1} \, |||w_h|||_{CIP}.$$

Finally, using similar arguments to estimate the stabilizing term from the start of the proof, we get

$$|J_h(u - i_h u, w_h)| \leq C h^{r+1/2} \|u\|_{r+1} \, |J_h(w_h, w_h)|^{1/2}$$

$$\leq C h^{r+1/2} \|u\|_{r+1} \, |||w_h|||_{CIP}.$$

Combining the above estimates produces the desired error estimate. □

Remark 3.87. The proof of Theorem 3.86 assumed that the meshes were quasi-uniform. This assumption can be relaxed slightly [BFH06]. An alternative way of avoiding the assumption of quasi-uniformity is to replace the L_2 projection i_h by the standard nodal interpolation u^I. Although one cannot then appeal to an orthogonality property when estimating the convection term, nevertheless an $\mathcal{O}(h^r)$ error estimate (instead of the above $\mathcal{O}(h^{r+1/2})$) can be established; see [Sch07]. ♣

Remark 3.88. The continuous interior penalty approach is generalized to the hp version of the finite element method in [BE07]. In [BH04] the question of a discrete maximum principle is discussed. Local error estimates similar to those stated for the streamline diffusion method in Theorem 3.41 have been established in [BGL07]. ♣

Finally, we wish to point out the close relationship between the LPS and CIP analyses. The essential point in the error estimation of both methods is a special treatment of the convection term.

For the LPS method, after integrating by parts, the orthogonality property of a special interpolant j_h with respect to the projection space D_h is used:

$$(u - j_h u, b \cdot \nabla w_h) = (u - j_h u, b \cdot \nabla w_h - \pi_h(b \cdot \nabla w_h))$$

where $\pi_h : L_2(\Omega) \to D_h$ is a local projection into the discontinuous projection space D_h. Control over $\kappa_h(b \cdot \nabla w_h) = b \cdot \nabla w_h - \pi_h(b \cdot \nabla w_h)$ is achieved by adding a stabilizing term like (3.86b) which causes a consistency error, but this is sufficiently small provided that the projection space D_h is sufficiently large.

In the CIP stabilization method, the special interpolant j_h is replaced by the standard (global) L_2 projection $i_h : L_2(\Omega) \to V_h$ into the continuous finite element space Y_h and the L_2 projection π_h of LPS is replaced by the quasi-interpolant π_h^* into Y_h. The special construction of the quasi-interpolant π_h^* permits an L_2 control of $b_h \cdot \nabla w_h - \pi_h^*(b_h \cdot \nabla w_h)$ by (appropriately scaled) jumps in the gradient of w_h – see Lemma 3.80 above.

3.4 The Discontinuous Galerkin Finite Element Method

In 1973, Reed and Hill [RH73] introduced the first discontinuous Galerkin method for hyperbolic first-order equations. Since then there has been an active development of discontinuous Galerkin methods for hyperbolic problems; the technique is also used in the time discretization of unsteady problems, as we saw in Section II.4.2.2.

In the 1970s Galerkin methods using discontinuous finite elements were proposed also for elliptic equations, but it is only recently that the discontinuous Galerkin finite element method (dGFEM) for elliptic problems has attracted the attention of many researchers. When applied to convection-diffusion problems, like SDFEM and other stabilizations the dGFEM is much more stable than the standard Galerkin FEM, but the details of the construction of the dGFEM bilinear form are very different from these other methods. It has even been claimed that dGFEM has the advantage that it needs no special stabilization for convection-diffusion problems, but in fact it includes a natural upwinding that is equivalent to some stabilization.

The name of the method comes from its use of a standard polynomial trial space on every element that is not required to be continuous across element boundaries. Thus nonstandard meshes can be used: in two dimensions, triangles and rectangles can be combined arbitrarily, nonconvex quadrilaterals are allowed and in the case of mesh refinement it is not necessary to avoid hanging nodes. The local nature of the method means it is more readily parallelizable than the standard FEM and it clearly permits the use of polynomials of different degrees on different elements (one can in two dimensions use linears on quadrilaterals and bilinears on triangles). This flexibility can be exploited to gain increased accuracy when the solution of a problem is quite smooth on a part of the domain – as is usually the case for convection-diffusion problems. On the other hand, a drawback of the dGFEM is its much larger number of degrees of freedom compared with the standard Galerkin FEM.

The following subsections introduce the dGFEM. We restrict ourselves to simple elements (linears and bilinears) and conforming meshes. The method can be extended to nonconforming meshes and to the hp-version of the finite element method [ABCM02, BO99, Coc03, HSS02]. For an application to compressible flow problems, see [FFS03].

3.4.1 The Primal Formulation for a Reaction-Diffusion Problem

Consider the boundary value problem

$$-\varepsilon \Delta u + cu = f \quad \text{in } \Omega \subset \mathbb{R}^2, \tag{3.106a}$$

$$u = 0 \quad \text{on } \Gamma := \partial \Omega, \tag{3.106b}$$

assuming that $c > 0$ and Ω is a polygonal bounded domain.

Let $\mathcal{T}$ be a general partition of the domain Ω into elements (disjoint closed triangles or rectangles) κ such that

$$\bar{\Omega} = \bigcup_{\kappa \in \mathcal{T}} \kappa \,.$$

Assume that there are no hanging nodes; see [HSS02] where one hanging node per element is allowed.

To each element $\kappa \in \mathcal{T}$ assign a nonnegative integer s_κ and define the broken Sobolev space of composite order $\mathbf{s} = \{s_\kappa : \kappa \in \mathcal{T}\}$ by

$$H^{\mathbf{s}}(\Omega, \mathcal{T}) = \left\{ v \in L^2(\Omega) : v|_\kappa \in H^{s_\kappa}(\kappa), \forall \kappa \in \mathcal{T} \right\} \,.$$

The corresponding broken Sobolev norm and seminorm are

$$|||v|||_{\mathbf{s},\mathcal{T}} = \left(\sum_{\kappa \in \mathcal{T}} \|v\|^2_{H^{s_\kappa}(\kappa)} \right)^{\frac{1}{2}} \qquad \text{and} \qquad |v|_{\mathbf{s},\mathcal{T}} = \left(\sum_{\kappa \in \mathcal{T}} |v|^2_{H^{s_\kappa}(\kappa)} \right)^{\frac{1}{2}} .$$

If $s_\kappa = s$ for all $\kappa \in \mathcal{T}$, we then write $H^s(\Omega, \mathcal{T})$, $|||v|||_{s,\mathcal{T}}$ and $|v|_{s,\mathcal{T}}$. For each $v \in H^1(\Omega, \mathcal{T})$, the broken gradient $\nabla_{\mathcal{T}} v$ of a function v is defined by $(\nabla_{\mathcal{T}} v)|_\kappa = \nabla(v|_\kappa)$, $\kappa \in \mathcal{T}$.

Assume also that each element $\kappa \in \mathcal{T}$ is the affine image of a fixed reference element $\hat{\kappa}$, viz., $\kappa = F_\kappa(\hat{\kappa})$. Then the finite element space is defined to be

$$S(\Omega, \mathcal{T}, \mathbf{F}) = \left\{ v \in L^2(\Omega) : v|_\kappa \circ F_\kappa \in P_1(\hat{\kappa}) \right\} \,,$$

where $\mathbf{F} = \{F_\kappa : \kappa \in \mathcal{T}\}$ and $P_1(\hat{\kappa})$ is the space of linear functions defined on $\hat{\kappa}$. That is, the solution of (3.106) is approximated by a piecewise linear function that is continuous on each element κ but allowed to be discontinuous across interelement edges. (Similarly, one could use bilinears or even mixed linear and bilinear elements.)

Let $\mathcal{E}$ be the set of all open one-dimensional element interfaces associated with the partition $\mathcal{T}$, and write $\mathcal{E}_{int} \subset \mathcal{E}$ for the the set of all edges $e \in \mathcal{E}$ that lie in Ω. Set $\Gamma_{int} = \{x \in \Omega : x \in e \text{ for some } e \in \mathcal{E}_{int}\}$. Number the elements as $\kappa_1, \kappa_2, \ldots$ Then for each $e \in \mathcal{E}_{int}$ there exist indices i and j such that $i > j$ where $\kappa := \kappa_i$ and $\kappa' := \kappa_j$ share the interface e. The (element-numbering-dependent) jump of a function $v \in H^1(\Omega, \mathcal{T})$ across e and the mean value of v on e are defined by

$$[v]_e = v|_{\partial \kappa \cap e} - v|_{\partial \kappa' \cap e} \qquad \text{and} \qquad \langle v \rangle_e = \frac{1}{2} \left(v|_{\partial \kappa \cap e} + v|_{\partial \kappa' \cap e} \right),$$

where $\partial \kappa$ denotes the union of all open edges of the element κ. With each interface $e \in \mathcal{E}_{int}$ associate the unit normal vector ν pointing from κ to κ'; if $e \subset \Gamma$, take ν to be the outward-pointing unit normal vector μ.

To simplify the notation, indices are sometimes omitted from the terms $[v]_e$ and $\langle v \rangle_e$.

In formulating the discrete problem, we assume that the solution u of (3.106) satisfies $u \in H^2(\Omega) \subset H^2(\Omega, \mathcal{T})$. Then

$$[u]_e = 0 \quad \text{and} \quad \langle u \rangle_e = u \quad \forall e \in \mathcal{E}_{int}.$$

As is standard for finite element methods, multiply the differential equation (3.106a) by a test function $v \in H^1(\Omega, \mathcal{T})$ then integrate over the domain Ω, obtaining

$$\int_{\Omega} (-\varepsilon \Delta u + cu) \, v \, dx = \int_{\Omega} fv \, dx. \tag{3.107}$$

Let us now consider the contribution of $-\varepsilon \Delta u$ to (3.107). Integrating by parts over each element κ then summing over all $\kappa \in \mathcal{T}$, some manipulations yield (denote the normal vectors that arise from the application of Green's formula by μ on Γ, μ_κ on the boundary of κ and ν on the set of all interior edges)

$$\int_{\Omega} (-\varepsilon \Delta u) \, v \, dx = \sum_{\kappa \in \mathcal{T}} \varepsilon \int_{\kappa} \nabla u \cdot \nabla v \, dx - \sum_{\kappa \in \mathcal{T}} \varepsilon \int_{\partial \kappa} (\nabla u \cdot \mu_\kappa) \, v \, ds$$

$$= \sum_{\kappa \in \mathcal{T}} \varepsilon \int_{\kappa} \nabla u \cdot \nabla v \, dx - \sum_{e \in \mathcal{E} \cap \Gamma} \varepsilon \int_{e} (\nabla u \cdot \mu) \, v \, ds$$

$$- \sum_{e \in \mathcal{E}_{int}} \varepsilon \int_{e} \left[((\nabla u \cdot \mu_\kappa) v)|_{\partial \kappa \cap e} + ((\nabla u \cdot \mu_{\kappa'}) v)|_{\partial \kappa' \cap e} \right] ds,$$

where an interior edge e is common to the elements κ and κ'. The sum of the integrals over all $e \in \mathcal{E}_{int}$ can be written as

$$\sum_{e \in \mathcal{E}_{int}} \varepsilon \int_{e} \left[((\nabla u \cdot \mu_\kappa) v)|_{\partial \kappa \cap e} + ((\nabla u \cdot \mu_{\kappa'}) v)|_{\partial \kappa' \cap e} \right] ds$$

$$= \sum_{e \in \mathcal{E}_{int}} \varepsilon \int_{e} \left[((\nabla u \cdot \nu) v)|_{\partial \kappa \cap e} - ((\nabla u \cdot \nu) v)|_{\partial \kappa' \cap e} \right] ds$$

$$= \sum_{e \in \mathcal{E}_{int}} \varepsilon \int_{e} \left(\langle \nabla u \cdot \nu \rangle_e [v]_e + [\nabla u \cdot \nu]_e \langle v \rangle_e \right) ds$$

$$= \sum_{e \in \mathcal{E}_{int}} \varepsilon \int_{e} \langle \nabla u \cdot \nu \rangle_e [v]_e \, ds.$$

Introducing the notation

$$\sum_{e \in \mathcal{E}_{int}} \varepsilon \int_{e} \langle \nabla u \cdot \nu \rangle_e [v]_e \, ds = \varepsilon \int_{\Gamma_{int}} \langle \nabla u \cdot \nu \rangle [v] \, ds,$$

$$\sum_{e \in \mathcal{E} \cap \Gamma} \varepsilon \int_{e} (\nabla u \cdot \mu) \, v \, ds = \varepsilon \int_{\Gamma} (\nabla u \cdot \mu) \, v \, ds,$$

we obtain

$$\int_\Omega (-\varepsilon\Delta u)\, v\, dx = \sum_{\kappa\in\mathcal{T}} \varepsilon \int_\kappa \nabla u \cdot \nabla v\, dx$$

$$- \varepsilon \int_\Gamma (\nabla u \cdot \mu)\, v\, ds - \varepsilon \int_{\Gamma_{int}} \langle\nabla u \cdot \nu\rangle [v]\, ds\,.$$

To the right–hand side of this expression add or subtract the zero terms

$$\varepsilon \int_\Gamma u(\nabla v \cdot \mu)\, ds \qquad \text{and} \qquad \varepsilon \int_{\Gamma_{int}} [u]\langle\nabla v \cdot \nu\rangle\, ds$$

to symmetrize the formula. Also add the following penalty terms to achieve coercivity of the bilinear form on the discrete space:

$$\int_\Gamma \sigma uv\, ds \qquad \text{and} \qquad \int_{\Gamma_{int}} \sigma [u][v]\, ds\,.$$

Here σ is called the *discontinuity-penalization* parameter and is defined by

$$\sigma|_e = \sigma_e \quad \text{for} \quad e \in \mathcal{E},$$

where $\sigma_e \geq 0$ is a user-chosen constant. Recall that in the edge stabilization or continuous interior penalty method (CIP) of the previous section, similar arguments were used. While in a continuous finite element method it is natural to punish jumps of the gradient, in a discontinuous method it is the jumps of the function values that are penalised. The parallels between the CIP and discontinuous methods allow similar error analyses [Bur05].

Returning to the derivation of the method, we get finally

$$\int_\Omega (-\varepsilon\Delta u)\, v\, dx = \sum_{\kappa\in\mathcal{T}} \varepsilon \int_\kappa \nabla u \cdot \nabla v\, dx$$

$$+ \varepsilon \int_\Gamma (\pm u(\nabla v \cdot \mu) - (\nabla u \cdot \mu)v)\, ds + \int_\Gamma \sigma uv\, ds$$

$$+ \varepsilon \int_{\Gamma_{int}} (\pm [u]\langle\nabla v \cdot \nu\rangle - \langle\nabla u \cdot \nu\rangle [v])\, ds + \int_{\Gamma_{int}} \sigma [u][v]\, ds.$$

Here either both plus or both minus signs are used. Now the primal formulation of the dGFEM with interior penalties can be stated as:

$$\begin{cases} \text{Find} \quad u_h \in S(\Omega, \mathcal{T}, \mathbf{F}) \text{ such that} \\ B_\pm(u_h, v_h) = L(v_h) \text{ for all } v_h \in S(\Omega, \mathcal{T}, \mathbf{F}), \end{cases}$$

with

$$L(w) := \sum_{\kappa\in\mathcal{T}} \int_\kappa fw\, dx.$$

Correspondingly, the bilinear form is defined by

$$B_\pm(v, w) = \sum_{\kappa \in \mathcal{T}} \left(\varepsilon \int_\kappa \nabla v \cdot \nabla w \, dx + \int_\kappa cvw \, dx \right)$$

$$+ \varepsilon \int_\Gamma (\pm v(\nabla w \cdot \mu) - (\nabla v \cdot \mu)w) \, ds + \int_\Gamma \sigma vw \, ds$$

$$+ \varepsilon \int_{\Gamma_{int}} (\pm[v]\langle\nabla w \cdot \nu\rangle - \langle\nabla v \cdot \nu\rangle[w]) \, ds + \int_{\Gamma_{int}} \sigma[v][w] \, ds.$$

The minus sign leads to a symmetric bilinear form, and the method is then called *symmetric with interior penalties* (SIP). With the plus sign we have, surprisingly, an asymmetric bilinear form even though we started from a symmetric problem; this method is called *non-symmetric with interior penalties* (NIP) in the literature. The relative advantages and disadvantages of these two approaches will be discussed later.

Note that the numbering-dependent notation for the jumps and the normal ν does not affect the final formulation since only products of them appear.

Example 3.89. Let us consider the simple problem

$$-u'' = f, \qquad u(0) = u(1) = 0,$$

and its discretization with linear elements on an equidistant mesh of mesh size h, using the NIP method and choosing $\sigma = 1/h$.

On the interval (x_i, x_{i+1}), set $u_h = u_i^+ \phi_i^+ + u_{i+1}^- \phi_{i+1}^-$. Here, for instance, ϕ_i^+ is the restriction to the interval (x_i, x_{i+1}) of the standard hat function associated with the mesh point x_i. Using the test functions ϕ_i^- and ϕ_i^+, at each interior mesh point we obtain the following equations:

$$\frac{1}{h} \left(\frac{1}{2} u_{i-1}^+ - u_i^- + 2u_i^+ - u_{i+1}^- - \frac{1}{2} u_{i+1}^+ \right) = \int_{x_{i-1}}^{x_i} f \phi_i^-,$$

$$\frac{1}{h} \left(-\frac{1}{2} u_{i-1}^- - u_{i-1}^+ + 2u_i^- - u_i^+ + \frac{1}{2} u_{i+1}^- \right) = \int_{x_i}^{x_{i+1}} f \phi_i^+.$$

Thus the dGFEM generates a multi-valued difference stencil. But if one sums these two equations the classical three-point stencil for $(u_i^+ + u_i^-)/2$ is obtained.

This multi-valued difference stencil is valid at each interior mesh point; at the boundary the situation is different – in general the boundary conditions are enforced weakly.   ♣

Remark 3.90. (Flux formulation) Alternatively, one has the flux formulation of the dGFEM, which starts from the formulation of (3.106) as

$$\theta = \nabla u, \qquad -\varepsilon \nabla \cdot \theta + cu = f.$$

A corresponding weak form is

$$\int_\kappa \theta \cdot \tau = -\int_\kappa u \, \nabla \cdot \tau + \int_{\partial\kappa} u \, \mu_\kappa \cdot \tau,$$

$$-\varepsilon \int_\kappa \theta \cdot \nabla v + \int_\kappa cuv = \int_\kappa fv + \int_{\partial\kappa} \theta \cdot \mu_\kappa \, v.$$

This generates the following discretization: find u_h, θ_h such that

$$\int_\kappa \theta_h \cdot \tau_h = -\int_\kappa u_h \, \nabla \cdot \tau_h + \int_{\partial\kappa} \hat{u}_\kappa \, \mu_\kappa \cdot \tau_h,$$

$$-\varepsilon \int_\kappa \theta_h \cdot \nabla v_h + \int_\kappa cu_h v_h = \int_\kappa fv_h + \int_{\partial\kappa} \hat{\theta}_\kappa \cdot \mu_\kappa \, v_h.$$

Here the choice of the numerical fluxes $\hat{\theta}_\kappa$ and $\hat{u}_\kappa$ that approximate $\theta = \nabla u$ and u on $\partial\kappa$ is very important. In [ABCM02] one finds a thorough discussion of 9 variants of the dGFEM that are characterized by different choices of $\hat{\theta}_\kappa$ and $\hat{u}_\kappa$. For each of these methods, the properties of the associated primal formulation that is obtained by eliminating θ_h are discussed.   ♣

Remark 3.91. (Discontinuous Petrov-Galerkin formulation) In the approach just described the numerical fluxes are defined by expressing the element interface fields as suitable averages of the internal fields. Following a slightly different philosophy, the discontinuous Petrov-Galerkin formulation approximates all unknown fields by internal and boundary variables. See [CSB05] for the application of this method to convection-diffusion problems.   ♣

We shall work only with the primal form of the dGFEM. Next, the treatment of the convective term is discussed.

3.4.2 A First-Order Hyperbolic Problem

Consider in this subsection the pure convection problem

$$b \cdot \nabla u + cu = f \quad \text{in} \quad \Omega, \tag{3.108a}$$

$$u = g \quad \text{on} \quad \Gamma_-, \tag{3.108b}$$

with the assumption that $c - (\text{div } b)/2 \geq \omega > 0$. It is well known that the standard Galerkin discretization of (3.108) by linear or bilinear elements gives solutions that are only $\mathcal{O}(h)$ accurate in the L_2 norm [JNP84]; moreover the stability properties of this method are unsatisfactory. Thus the method needs to be improved using, e.g., the streamline diffusion method – or the discontinuous Galerkin approach as we now explain.

When discretizing the convective term, a form of upwinding is used. Define the inflow and outflow parts of the boundary $\partial\kappa$ of each $\kappa \in \mathcal{T}$ by

$$\partial_-\kappa = \{x \in \partial\kappa : b(x) \cdot \mu_\kappa(x) < 0\} \quad \text{and} \quad \partial_+\kappa = \{x \in \partial\kappa : b(x) \cdot \mu_\kappa(x) \geq 0\}$$

respectively, where $\mu_\kappa(x)$ again represents the outward-pointing unit vector normal to $\partial\kappa$ at the point $x \in \partial\kappa$.

For each element $\kappa \in \mathcal{T}$ and $v \in H^1(\kappa)$, denote by v_κ^+ the interior trace of $v|_\kappa$ on $\partial\kappa$. If $\partial_-\kappa \setminus \Gamma \neq \emptyset$ for some $\kappa \in \mathcal{T}$, then for each $x \in \partial_-\kappa \setminus \Gamma$ there exists a unique $\kappa' \in \mathcal{T}$ such that $x \in \partial_+\kappa'$. Now for a function $v \in H^1(\Omega, \mathcal{T})$ and for each $\kappa \in \mathcal{T}$ with the property that $\partial_-\kappa \setminus \Gamma \neq \emptyset$, define the outer trace v_κ^- of v on $\partial_-\kappa \setminus \Gamma$ relative to κ to be the inner trace $v_{\kappa'}^+$ relative to the element κ' such that $\partial_+\kappa' \cap (\partial_-\kappa \setminus \Gamma) \neq \emptyset$. The jump of v across $\partial_-\kappa \setminus \Gamma$ is defined by

$$\lfloor v \rfloor_\kappa = v_\kappa^+ - v_\kappa^-.$$

Note that the jump $\lfloor \cdot \rfloor$ depends on the flow direction b, unlike the previous jump $[\cdot]$ which depends on the numbering of elements.

Now

$$\int_\kappa (b \cdot \nabla u)\, v \, dx = \int_{\partial\kappa} (b \cdot \mu_\kappa) uv \, ds - \int_\kappa u\, \nabla \cdot (b\, v)\, dx$$

$$= \int_{\partial_-\kappa} (b \cdot \mu_\kappa) uv \, ds + \int_{\partial_+\kappa} (b \cdot \mu_\kappa) uv \, ds - \int_\kappa u\, \nabla \cdot (b\, v)\, dx.$$

For a continuous function it is irrelevant whether one writes u, u^+ or u^-. But for a discontinuous function, in all integrals over the boundary one might consider replacing all boundary values by their inner traces. A better idea is to replace u on the inflow part of the boundary by u^-, which introduces a form of upwinding.

Again integrating the last term by parts (which leads to a cancellation of the terms on $\partial_+\kappa$), we get

$$\int_\kappa (b \cdot \nabla u)\, v \, dx$$

$$= \int_{\partial_-\kappa} (b \cdot \mu_\kappa) u^- v^+ \, ds + \int_{\partial_+\kappa} (b \cdot \mu_\kappa) u^+ v^+ \, ds - \int_\kappa u\, \nabla \cdot (b\, v)\, dx$$

$$= \int_\kappa (b \cdot \nabla u)\, v \, dx - \int_{\partial_-\kappa \cap \Gamma_-} (b \cdot \mu_\kappa) u^+ v^+ \, ds - \int_{\partial_-\kappa \setminus \Gamma} (b \cdot \mu_\kappa) \lfloor u \rfloor v^+ \, ds.$$

Thus (3.108) has the following weak formulation:

$$B_0(u, v) := \sum_{\kappa \in \mathcal{T}} \left(\int_\kappa (b \cdot \nabla u + cu)\, v \, dx \right.$$

$$\left. - \int_{\partial_-\kappa \cap \Gamma} (b \cdot \mu_\kappa) u^+ v^+ \, ds - \int_{\partial_-\kappa \setminus \Gamma} (b \cdot \mu_\kappa) \lfloor u \rfloor v^+ \, ds \right)$$

$$= \sum_{\kappa \in \mathcal{T}} \left(\int_\kappa fv \, dx - \int_{\partial_-\kappa \cap \Gamma_-} (b \cdot \mu_\kappa) gv^+ \, ds \right).$$

In simple cases the corresponding discretization is an upwind scheme, as the next example demonstrates.

Example 3.92. Let us discretize the problem

$$u_x + u = f \quad \text{in } (0,1), \quad u(0) = A,$$

using the above dGFEM with a piecewise constant approximation u_i on every subinterval (x_{i-1}, x_i) of length h_i. In the first interval one gets

$$\frac{u_1}{h_1} + u_1 = \frac{1}{h_1} \int_{x_0}^{x_1} f \, dx + \frac{1}{h_1} A.$$

This equation corresponds to a weak enforcement of the initial condition. For all other intervals one has

$$\frac{u_i - u_{i-1}}{h_i} + u_i = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} f \, dx$$

which is an upwind scheme.

If one uses linear elements, the scheme depends as usual on the degrees of freedom chosen. For the dGFEM there are more possibilities than for a continuous Galerkin method; for instance, one could take the function values and their derivatives at the midpoints of each mesh interval.

One can associate the two values u_i^- and u_i^+ with each mesh point x_i, interpreting the method as a multivalued difference scheme. Eliminating u_i^+ while assuming an equidistant mesh and constant f, on writing u_i for u_i^- the scheme becomes

$$\frac{u_i - u_{i-1}}{h} + \frac{1}{2} \frac{u_{i-1} + (1 + h/3)u_i}{1 + h/6} = \frac{1}{2} \frac{f + (1 + h/3)f}{1 + h/6}.$$

Thus the scheme bears some resemblance to the midpoint upwind scheme and consequently to the streamline diffusion method.   ♣

What is the advantage of the bilinear form $B_0(\cdot, \cdot)$ over standard Galerkin? Setting $c_0^2 := c - (\nabla \cdot b)/2$, after integration by parts it is easy to see that

$$B_0(v, v) = \sum_{\kappa \in \mathcal{T}} \Big[\|c_0 v\|_{L^2(\kappa)}^2$$

$$+ \frac{1}{2} \left(\|v^+\|_{\partial_-\kappa \cap \Gamma}^2 + \|v^+ - v^-\|_{\partial_-\kappa \setminus \Gamma}^2 + \|v^+\|_{\partial_+\kappa \cap \Gamma}^2 \right) \Big].$$

Here we used the notation

$$(v, w)_\tau = \int_\tau |b \cdot \mu_\kappa| vw \, ds \quad \text{for } \tau \subset \partial\kappa, \quad \text{and set} \quad \|v\|_\tau^2 = (v, v)_\tau.$$

This shows that in the dGFEM one controls not only the L_2 error but also the error in the stronger norm $B_0(v, v)^{1/2}$. The Galerkin orthogonality property

$$B_0(u - u_h, v_h) = 0$$

then permits us to use standard techniques in the error analysis that will be presented in the next subsection for the full convection-diffusion problem.

Remark 3.93. (Jump formulation of upwinding) Instead of using the upwind idea described above, the upwind effect can be generated equivalently by adding a certain jump penalty term [Coc03]. In [BMS04] it is observed that this approach has two advantages: stability can be proved more elegantly and one has more flexibility in tuning the amount of stabilization used.   ♣

3.4.3 dGFEM Error Analysis for Convection-Diffusion Problems

The ideas of the two previous subsections will now be merged in considering the convection-diffusion problem

$$-\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in} \quad \Omega, \tag{3.109a}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{3.109b}$$

while assuming that $c - (\nabla \cdot b)/2 \geq \omega > 0$ and Ω has a polygonal boundary. The dGFEM bilinear form for the problem (3.109) is

$$
\begin{aligned}
B_{\pm}(v, w) := \sum_{\kappa \in \mathcal{T}} \bigg(& \varepsilon \int_\kappa \nabla v \cdot \nabla w \, dx + \int_\kappa (b \cdot \nabla v + cv) w \, dx \\
& - \int_{\partial_- \kappa \cap \Gamma} (b \cdot \mu) v^+ w^+ \, ds - \int_{\partial_- \kappa \setminus \Gamma} (b \cdot \mu_\kappa) \lfloor v \rfloor w^+ \, ds \bigg) \\
& + \varepsilon \int_\Gamma (\pm v(\nabla w \cdot \mu) - (\nabla v \cdot \mu) w) \, ds + \int_\Gamma \sigma vw \, ds \\
& + \varepsilon \int_{\Gamma_{int}} (\pm [v] \langle \nabla w \cdot \nu \rangle - \langle \nabla v \cdot \nu \rangle [w]) \, ds + \int_{\Gamma_{int}} \sigma [v][w] \, ds
\end{aligned}
$$

for all $v, w \in H^1(\Omega, \mathcal{T})$. The dGFEM with interior penalties is:

$$
\begin{cases}
\text{Find} \quad u_h \in S(\Omega, \mathcal{T}, \mathbf{F}) \text{ such that} \\
B_{\pm}(u_h, v_h) = L(v_h) \text{ for all } v_h \in S(\Omega, \mathcal{T}, \mathbf{F}),
\end{cases}
$$

with

$$L(w) := \sum_{\kappa \in \mathcal{T}} \int_\kappa fw \, dx.$$

Assume that $u \in H^2(\Omega, \mathcal{T})$ and $\nabla u \cdot \nu$ is continuous across each interior edge e. Then one has both consistency of the method and the Galerkin orthogonality property

$$B_{\pm}(u - u_h, v) = 0 \quad \text{for all} \quad v \in S(\Omega, \mathcal{T}, \mathbf{F}).$$

Define the full dGFEM norm by

$$|||v|||_{dG}^2 = \sum_{\kappa \in \mathcal{T}} \left(\varepsilon \|\nabla v\|_{L^2(\kappa)}^2 + \|c_0 v\|_{L^2(\kappa)}^2 \right) + \int_\Gamma \sigma v^2 \, ds + \int_{\Gamma_{int}} \sigma [v]^2 \, ds$$

$$+ \frac{1}{2} \sum_{\kappa \in \mathcal{T}} \left(\|v^+\|_{\partial_- \kappa \cap \Gamma}^2 + \|v^+ - v^-\|_{\partial_- \kappa \backslash \Gamma}^2 + \|v^+\|_{\partial_+ \kappa \cap \Gamma}^2 \right).$$

Is $B_\pm(\cdot, \cdot)$ coercive over $S(\Omega, \mathcal{T}, \mathbf{F}) \times S(\Omega, \mathcal{T}, \mathbf{F})$? For the *asymmetric NIP method* it is easy to see that

$$B_-(v, v) = |||v|||_{dG}^2$$

on *any* mesh. For the symmetric SIP version, assuming a shape-regular mesh, an analysis proves coercivity provided that

$$\sigma = \frac{\varepsilon}{h} \sigma_0$$

with a sufficiently large constant σ_0. As it's our intention to carry out an error analysis later on an anisotropic mesh, we prefer the NIP version for its simpler stability property. See [Geo05] for an examination of the SIP on anisotropic meshes.

Remark 3.94. To prove optimal L_2 error estimates or to apply the DWR method of Section 3.6 in order to control some error functionals, the *adjoint consistency* of the method is important. If the dual problem formed by transposing the arguments in the bilinear form $B_\pm(\cdot, \cdot)$ is based on the formal adjoint of the original differential operator, then the bilinear form $B_\pm(\cdot, \cdot)$ is said to be adjoint consistent. It turns out that SIP is adjoint consistent but NIP lacks this property; see [HHSS03]. ♣

We now sketch the dG-norm error analysis on a shape-regular mesh for NIP; one uses similar arguments for SIP. For convenience, write the bilinear form of NIP as $B(\cdot, \cdot)$. The starting point is the error representation

$$u - u_h = (u - \Pi u) + (\Pi u - u_h) \equiv \eta + \xi,$$

using some projector Π onto the finite element space. Galerkin orthogonality gives

$$|||\xi|||_{dG}^2 = B(\xi, \xi) = -B(\eta, \xi),$$

and one deals with $|B(\eta, \xi)|$ in such a way that eventually only the projection error in various norms needs to be estimated. For the projection operator Π let us take the L_2 projection onto our discontinuous finite element space. Then for linear or bilinear elements on a shape-regular mesh, one obtains as a consequence of the Bramble-Hilbert lemma that

$$\|\eta\|_{L_2} \le C h^2 \|u\|_{H^2(\Omega)} \quad \text{and} \quad \|\eta\|_{H^1(\Omega, \mathcal{T})} \le C h \|u\|_{H^2(\Omega)}.$$

To estimate the convective terms, on applying Cauchy-Schwarz to $|B_0(\eta, \xi)|$ one sees that one must estimate both

$$\sum_{\kappa \in \mathcal{T}} \|\eta\|_{L_2(\kappa)} + \|\eta^-\|_{\partial_- \kappa \backslash \Gamma} + \|\eta^+\|_{\partial_- \kappa \cap \Gamma}$$

and

$$\sum_{\kappa \in \mathcal{T}} \int_\kappa \eta(b \cdot \nabla \xi) \, dx.$$

For the first of these one invokes the *multiplicative trace inequality* [DFS02]

$$\|v\|^2_{L_2(\partial \kappa)} \le C \left(\|v\|_{L_2(\kappa)} |v|_{H^1(\kappa)} + \frac{1}{h_\kappa} \|v\|^2_{L_2(\kappa)} \right) \quad \text{for all } v \in H^1(\kappa).$$

Regarding the second, if

$$b \cdot \nabla_{\mathcal{T}} v \in S(\Omega, \mathcal{T}, \mathbf{F}) \qquad \forall v \in S(\Omega, \mathcal{T}, \mathbf{F}),$$

the contribution is zero owing to the choice of Π. If b is neither piecewise linear nor bilinear, then use the triangle inequality and an inverse inequality to obtain

$$\left| \sum_{\kappa \in \mathcal{T}} \int_\kappa \eta(b \cdot \nabla \xi) \, dx \right| \le C \, h^2 \|u\|_{H^2} \|\xi\|_{L_2}$$

for $b \in W^{1,\infty}(\Omega)$. Summarizing the bounds on the convective part, we get our first important result for the pure convection problem:

Lemma 3.95. *Consider the pure convection problem (3.108). Assume that $c - (\nabla \cdot b)/2 \ge \omega > 0$. Let this problem be discretized on a shape-regular mesh using the discontinuous Galerkin finite element method with linear or bilinear elements. Then the error satisfies*

$$|||u - u_h|||_{dG} \le C \, h^{3/2} \|u\|_{H^2(\Omega)}.$$

For the convection-diffusion problem (3.109) it still remains to estimate the remaining terms in $B(\eta, \xi)$. The first two terms and the penalty terms are bounded via a direct application of the Cauchy-Schwarz inequality. In addition, two types of integrals on Γ and on Γ_{int} can be estimated in a similar way. Let us demonstrate the technique for the integrals on Γ: introducing an auxiliary positive parameter γ, for the expression

$$Z = \int_\Gamma \varepsilon \big(\eta(\nabla \xi \cdot \nu) - (\nabla \eta \cdot \nu)\xi \big) \, ds$$

one obtains the estimate

$$|Z| \leq \left(\sum_{\kappa \in \mathcal{T}} \frac{\varepsilon}{\gamma} \|\eta\|^2_{L_2(\partial\kappa\cap\Gamma)} \right)^{1/2} \left(\sum_{\kappa \in \mathcal{T}} \varepsilon\,\gamma \|\nabla\xi\|^2_{L_2(\partial\kappa\cap\Gamma)} \right)^{1/2}$$

$$+ \left(\sum_{\kappa \in \mathcal{T}} \frac{\varepsilon^2}{\sigma} \|\nabla\eta\|^2_{L_2(\partial\kappa\cap\Gamma)} \right)^{1/2} \left(\sum_{\kappa \in \mathcal{T}} \sigma \|\xi\|^2_{L_2(\partial\kappa\cap\Gamma)} \right)^{1/2}.$$

The second term involving ξ can be directly estimated by $|||\xi|||_{dG}$. In the first ξ term, apply an inverse inequality to replace the integrals over $\partial\kappa$ by integrals over κ, and to compensate for the power of h that arises choose $\gamma|_\kappa = \mathcal{O}(h_\kappa)$. This yields

$$|Z| \leq \left[\left(\sum_{\kappa \in \mathcal{T}} \frac{\varepsilon}{h_\kappa} \|\eta\|^2_{L_2(\partial\kappa\cap\Gamma)} \right)^{1/2} + \left(\sum_{\kappa \in \mathcal{T}} \frac{\varepsilon^2}{\sigma} \|\nabla\eta\|^2_{L_2(\partial\kappa\cap\Gamma)} \right)^{1/2} \right] |||\xi|||_{dG}.$$

One therefore has $\mathcal{O}(\varepsilon^{1/2}h)$ and $\mathcal{O}(\varepsilon\,h^{1/2}/\sigma^{1/2})$ error contributions. As the penalty terms make an error contribution of $\mathcal{O}(\sigma^{1/2}h^{3/2})$, an equilibration of the various terms leads us to the choice $\sigma = \sigma_0\varepsilon/h$ for some positive constant σ_0.

Theorem 3.96. *Let the convection-diffusion problem (3.109) be discretized using the asymmetric (NIP) dGFEM with linear or bilinear elements on a shape-regular mesh. Then the choice*

$$\sigma = \sigma_0\, \frac{\varepsilon}{h}, \tag{3.110a}$$

with some arbitrary positive constant σ_0, yields the error estimate

$$|||u - u_h|||^2_{DG} \leq C \left(\varepsilon h^2 + h^3 \right) \|u\|^2_{H^2(\Omega)}. \tag{3.110b}$$

Error estimates for the *hp* version of the dGFEM are derived in [HHSS03, HSS02]. Detailed numerical studies can be found in [Cas02].

Remark 3.97. (Relationship of the SIP to SDFEM and its conditioning) In [GK03] a modification of the bound (3.110b) is obtained. The authors considered the convection–diffusion problem where for the diffusion part a symmetric version (SIP) of dGFEM is used, based on the bilinear form $B_-(v,w)$. In terms of the slightly stronger norm

$$|||v|||^2_{dG*} = B_+(v,v) + \sum_{\kappa \in \mathcal{T}} h_\kappa \,\|b \cdot \nabla v\|^2_{L^2(\kappa)},$$

it is proved that on shape-regular meshes one has

$$|||u - u_h|||^2_{dG*} \leq C \left(\varepsilon h^2 + h^3 \right) \|u\|^2_{H^2(\Omega)}.$$

This estimate, in a norm similar to the SDFEM norm, shows there is some relationship between the SDFEM and dGFEM stabilizations.

In the same paper it is proved that the condition number of the discrete problem generated is $\mathcal{O}(h^{-2})$ and a multigrid method for solving the discrete problem is described. ♣

Remark 3.98. (Error estimates in L_∞ and local estimates) When the problem is not singularly perturbed (i.e., when $\varepsilon = 1$), error estimates for the SIP method are obtained in the L_∞ norm in [KR02], based on a detailed analysis of an approximate Green's function. It seems difficult to analyse the dependence of this error on ε when the problem is singularly perturbed.

In [Guz06] the author proves local error estimates similar to those we presented for the SDFEM; these show that the dGFEM works well in regions where the solution is smooth. ♣

Remark 3.99. (Superconvergence and a posteriori error estimates) Superconvergence phenomena for the dGFEM applied to convection-diffusion problems are discussed (for the one-dimensional case) in [CC07, XZ07].

For *a posteriori* error estimation for the dGFEM see [HSW07]. A convergence proof of a corresponding adaptive algorithm for dGFEM is in [KP07].
♣

While most papers on discontinuous Galerkin methods consider isotropic meshes, some extensions of the theory to general anisotropic meshes are given in [Geo06, GHH07a, GHH07b].

On layer-adapted meshes, a special strategy is to apply a Galerkin method on the fine mesh and a dGFEM stabilization outside the layer regions. Some details of this approach will be examined in the next section. Another combination of continuous and discontinuous methods is also possible: on a standard isotropic mesh one could use the dGFEM stabilization only in the regions near layers and the standard (unstabilized) Galerkin method away from the layers [CGJ06a].

3.5 Uniformly Convergent Methods

Most of this section is concerned with the finite element solution of convection-diffusion problems such as Example 1.15 whose solutions have only exponential boundary layers. Convection-diffusion problems like Example 1.16 (where parabolic boundary layers appear) and reaction-diffusion problems (see Remark 1.27) will also receive some attention.

Consider the convection-diffusion problem

$$Lu := -\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{on } \Omega := (0,1) \times (0,1), \tag{3.111a}$$
$$u = 0 \quad \text{on } \partial\Omega. \tag{3.111b}$$

Assume that $0 < \varepsilon \le 1$, that $b = (b_1(x,y), b_2(x,y)) > (\beta_1, \beta_2)$ on $\bar\Omega$, where β_1 and β_2 are positive constants, and that b, c and f are smooth. Recall that the solution u typically has exponential boundary layers at the sides $x = 1$ and $y = 1$ of $\bar\Omega$ and an exponential corner layer at the point $(1,1)$.

The hypothesis $b_1 > 0$ implies that without loss of generality one can assume that

$$c(x,y) \ge \gamma > 0 \quad \text{and} \quad c(x,y) - \frac{\nabla \cdot b(x,y)}{2} \ge \omega > 0 \quad \text{on } \bar\Omega \tag{3.112}$$

for some positive constants γ and ω, since these inequalities can be ensured via the change of variable $v(x,y) := e^{kx}u(x,y)$ for some suitable k.

For convenience assume also that $f \in C^{1,\alpha}(\bar\Omega)$ and that

$$f(0,0) = f(1,0) = f(0,1) = f(1,1) = 0, \tag{3.113}$$

so that $u \in C^{3,\alpha}(\bar\Omega)$, as was seen in Example 1.3. Nevertheless finite element analyses may not demand this much regularity of the solution.

The ε-weighted energy norm associated with (3.141) is

$$\|w\|_\varepsilon := \left(\varepsilon \|\nabla w\|_0^2 + \|w\|_0^2 \right)^{1/2}, \tag{3.114}$$

where $\|\cdot\|_0$ is the norm on $L_2(\Omega)$.

The finite element methods of Section 3.5 are *uniformly convergent* with respect to the norm $\|\cdot\|_\varepsilon$; that is, the computed solution u_h satisfies

$$\|u - u_h\|_\varepsilon \le C r^{-\alpha} \tag{3.115}$$

for some positive constants C and α that are independent of ε and of the number r of degrees of freedom in the finite element method. Using a power of r to quantify the error $u - u_h$ is valid for the particular families of meshes that we shall consider in Section 3.5, but such a bound would not be suitable for all possible mesh families – see [SW96] for a general discussion of such issues. Uniform convergence with respect to other norms (L_2 and L_∞) will also be examined in Section 3.5.

3.5.1 Operator-Fitted Methods

We begin with a generalization of the finite element methods of Section I.2.2.4, by taking tensor products of one-dimensional exponentially-fitted elements on rectangular axi-parallel meshes. This approach appears as a conforming method in [OS89, OS91b], then is generalized to a nonconforming method in [RAF96]. A conforming method that uses $\bar{L}$-splines slightly different from ours is analysed by Dörfler [Doe99b] in a more powerful framework that provides error bounds in various L_p generalizations of $\| \cdot \|_\varepsilon$ and is applicable to other problems (including convection-diffusion with parabolic boundary layers and reaction-diffusion), but our attention here is confined ourselves to error estimates in $\| \cdot \|_\varepsilon$ for the solution of (3.111).

Consider tensor-product meshes with nodes (x_i, y_j), where $x_i = ih$ and $y_j = jh$ for $i, j = 0, \ldots, M$ and $h = 1/M$. This mesh is square, but the results below remain valid on equidistant tensor-product meshes provided that the aspect ratio of each mesh rectangle in the family of meshes is bounded away from zero, uniformly in ε.

The conforming and nonconforming Galerkin methods both use a space S^h of exponentially-fitted $\bar{L}$-splines (cf. Section I.2.2.4) as trial and test space.

Set $\Omega^{ij} = (x_{i-1}, x_i) \times (y_{j-1}, y_j)$ for each i and j. Define the piecewise constant function $\bar{b}_k$ by $\bar{b}_k|_{\Omega^{ij}} := b_k^{ij}$, where

$$b_k^{ij} := ((b_k)_{i-1,j-1} + (b_k)_{i,j-1} + (b_k)_{i-1,j} + (b_k)_{ij})/4 \quad \text{for } k = 1, 2.$$

Analogously define $\bar{c}$ and $\bar{f}$. A typical basis function $\phi^{ij}(x, y) \in S^h$ satisfies $\phi^{ij}(x, y) = \phi^i(x)\phi_j(y)$; on each Ω^{mn}, ϕ^i and ϕ_j are defined respectively by

$$- \varepsilon(\phi^i)''(x) + b_1^{mn}(\phi^i)'(x) = 0, \quad \phi^i(x_{m-1}) = \delta_{i,m-1}, \quad \phi^i(x_m) = \delta_{i,m},$$

and

$$- \varepsilon(\phi_j)''(y) + b_2^{mn}(\phi_j)'(y) = 0, \quad \phi_j(y_{n-1}) = \delta_{j,n-1}, \quad \phi_j(y_n) = \delta_{j,n}.$$

Hence the support of each ϕ^{ij} is the union of the four mesh squares that meet at (x_i, y_j).

Suppose for the present that $b_1 = b_1(x)$ and $b_2 = b_2(y)$. Then the functions in S^h are continuous on $\bar{\Omega}$ and the finite element method is conforming. We define the bilinear form a_h associated with (3.111) by

$$a_h(w, z) = \int_\Omega [\varepsilon \nabla w \cdot \nabla z + (\bar{b} \cdot \nabla w)z + \bar{c}wz] \, dx \, dy \qquad (3.116)$$

for all w and z in $\tilde{H}_1 := \{w : w|_{\Omega^{ij}} \in H_1(\Omega_{ij}) \ \forall i, j\}$, where $\bar{b} := (\bar{b}_1, \bar{b}_2)$.

The computed solution u_h is required to satisfy

$$a_h(u_h, z) = (f, z)_h := \sum_{i,j=1}^{M} \int_{\Omega^{ij}} \bar{f}z \, dx \, dy \quad \text{for all } z \in S^h.$$

It is easy to see that

$$a_h(z, z) \geq \min\{1, \omega\} \|z\|_\varepsilon^2 \quad \text{for all } z \in S^h,$$

so u_h is well defined.

Using $b_1 = b_1(x)$, $b_2 = b_2(y)$ and the maximum principle property satisfied by the differential operator L, one can show that for all $(x, y) \in \Omega$, there exists a positive constant C such that

$$|u_x(x, y)| \leq C[1 + \varepsilon^{-1} \exp(-\beta_1(1 - x)/\varepsilon)], \tag{3.117a}$$

$$|u_y(x, y)| \leq C[1 + \varepsilon^{-1} \exp(-\beta_2(1 - y)/\varepsilon)], \tag{3.117b}$$

$$|(-\varepsilon u_{xx} + b_1 u_x)(x, y)| \leq C \quad \text{and} \quad |(-\varepsilon u_{yy} + b_2 u_y)(x, y)| \leq C. \tag{3.117c}$$

The bounds (3.117) are used in [OS89, OS91b] to prove the following two-dimensional analogue of Theorem I.2.74.

Theorem 3.100. *Assume that $b_1 = b_1(x)$ and $b_2 = b_2(y)$, that the data of (3.111) are smooth, and that (3.112) and (3.113) hold true. Then the finite element solution u_h satisfies*

$$\|u - u_h\|_\varepsilon \leq Ch^{1/2}.$$

Schieweck [Sch87] had earlier considered a conforming method that used exponentially-fitted splines near the layers and bilinear trial functions otherwise. He proved that

$$\|u - u_h\|_\varepsilon \leq C(\varepsilon^{1/2} + h^{1/2}) + C(m)(\varepsilon/h)^m,$$

where m is an arbitrary positive integer. When $\varepsilon \ll h$, this error bound is similar to the bound of Theorem 3.100.

Remark 3.101. Suppose for a moment that $b_2(x, y) > \beta_2 > 0$ and $b_1 \equiv 0$ in (3.111). Then Example 1.16 tells us that the solution has parabolic boundary layers at $x = 0$ and $x = 1$ and an exponential boundary layer at $y = 1$. To solve this problem on our square tensor-product mesh, use a Galerkin finite element method with piecewise bilinear trial functions, except in the exponential boundary layer, where the trial functions in the y-direction have the form of $\phi_j(y)$ above. Then Schieweck [Sch86] proves that

$$\|u - u_h\|_\varepsilon \leq C[h^{1/2} + (\varepsilon/h)^{1/2}],$$

where u_h is the computed solution.   ♣

We now return to the general case of (3.111) where $b_1 = b_1(x, y)$ and $b_2 = b_2(x, y)$. Then each finite element basis function ϕ^{ij} defined above may be discontinuous as one moves between the four mesh squares that constitute its support. Thus the method is nonconforming, which complicates the analysis in [RAF96]. The bilinear form $a_h(\cdot, \cdot)$ of (3.116) is now modified to

$$a_h(w, z) := \sum_{i,j=1}^{M} \int_{\Omega^{ij}} \left[\varepsilon \nabla w \cdot \nabla z + \frac{1}{2}(\bar{b} \cdot \nabla w)z - \frac{1}{2}(\bar{b} \cdot \nabla z)w \right.$$

$$\left. + \left(\bar{c} - \frac{1}{2}\overline{\nabla \cdot b} \right)wz \right] dx\, dy,$$

where the piecewise constant function $\overline{\nabla \cdot b}$ is defined analogously to $\bar{b}$. This type of bilinear form is frequently used when dealing with nonconforming methods, as it makes coercivity of the form easy to prove. It is obtained by integrating half the convective term by parts, then applying piecewise constant approximations.

As before, the computed solution u_h is required to satisfy

$$a_h(u_h, z) = (f, z)_h \quad \text{for all } z \in S^h.$$

Clearly

$$a_h(w, w) \geq \min\{1, \omega\}\|w\|_\varepsilon^2 \quad \text{for all } w \in S^h; \tag{3.118}$$

here we have abused the notation $\| \cdot \|_\varepsilon$ slightly by using it to mean the piecewise ε-weighted norm defined by

$$\|w\|_\varepsilon := \left[\sum_{i,j=1}^{M} \int_{\Omega^{ij}} \left(\varepsilon|\nabla w|^2 + w^2 \right) dx\, dy \right]^{1/2}.$$

By (3.118) the solution u_h is well defined.

As $b_1 = b_1(x, y)$ and $b_2 = b_2(x, y)$, it is now more difficult to prove bounds on the derivatives of u. A maximum principle for systems of operators is invoked in [RAF96] to get (3.117) again, provided that

$$(b_1)_y = (b_2)_x \leq 0, \quad (b_1)_x + c \geq 0, \quad (b_2)_y + c \geq 0, \tag{3.119a}$$
$$(b_1)_x + c > (b_2)_x \quad \text{and} \quad (b_2)_y + c > (b_1)_y. \tag{3.119b}$$

Theorem 3.102. *Assume that the data of (3.111) are smooth and that the conditions (3.112), (3.113) and (3.119) are satisfied. Then the finite element solution u_h satisfies*

$$\|u - u_h\|_\varepsilon \leq Ch^{1/2}.$$

Proof. We sketch the argument. Define an interpolant u^I of u in S^h by setting $u^I = \sum_{ij} u(x_i, y_j)\phi^{ij}$. Using the estimates (3.117) and the maximum principle satisfied by L on each Ω^{ij}, one can show that

$$\|u - u^I\|_{L_\infty(\Omega)} \leq Ch. \tag{3.120}$$

Now (3.118) implies that

$$\min\{1, \omega\} \|u - u^I\|_\varepsilon^2 \leq a_h(u - u^I, u - u^I)$$

and the right-hand side can be estimated using (3.111a) and (3.120). This yields

$$\|u - u^I\|_\varepsilon \le Ch^{1/2}. \tag{3.121}$$

Next, consider $\|u^I - u_h\|_\varepsilon$. By (3.118),

$$\min\{1, \omega\} \|u^I - u_h\|_\varepsilon^2 \le a_h(u^I - u_h, u^I - u_h)$$
$$= a_h(u^I - u, u^I - u_h) + a_h(u - u_h, u^I - u_h).$$

Hence

$$\|u^I - u_h\|_\varepsilon \le C \sup_{w \in S^h} \frac{|a_h(u^I - u, w)|}{\|w\|_\varepsilon} + C \sup_{w \in S^h} \frac{|a_h(u - u_h, w)|}{\|w\|_\varepsilon}.$$

Using detailed and careful estimates, one can deduce that

$$\|u^I - u_h\|_\varepsilon \le Ch^{1/2}.$$

Combine this inequality with (3.121) to finish.   □

Remark 3.103. Remark I.2.73 shows that the power of h in (3.121) is best possible. Thus Theorem 3.102 proves that our finite element method has the optimal rate of convergence with respect to $\|\cdot\|_\varepsilon$.   ♣

Theorem 3.102 implies that $\|u - u_h\|_0 \le Ch^{1/2}$. Numerical results exhibit a uniform rate of convergence in $L_2(\Omega)$ exceeding 0.6, even when the hypotheses of Theorem 3.102 are not all satisfied. Experimental data in [HOS93] also show that, if one replaces the test space S^h by piecewise bilinears while retaining the same exponentially fitted trial space, this yields a Petrov-Galerkin method that seems to have significantly better uniform convergence properties than both the Galerkin scheme above and all other schemes considered in that paper: it exhibited a uniform convergence rate greater than 1 in $L_2(\Omega)$ in several test problems. When $\varepsilon = 0$, it reduces to the box scheme of Example II.3.8.

Remark 3.104. Finite element methods that use *exponentials on triangles* are considered in [MW94, OS91a, RF08, SGG99, SS98]. Miller and Wang's method [MW94] was discussed in Section 3.1. In the *exponential streamline diffusion* method of [OS91a], exponential test functions upwind along the subcharacteristics of (3.111a). Convergence of order $h^{11/8} \ln(1/h)$ is obtained pointwise on regions where u is smooth (cf. (3.72)).

The method of [RF08] can be applied on triangles or quadrilaterals; it uses test functions that are approximate solutions of a dual problem, i.e., that approximate a Green's function. These test functions are combinations of exponentials and polynomials and the method aims to achieve higher orders of convergence in $L_2(\Omega)$. Numerical results illustrate its implementation and accuracy.

Sacco et al. [SGG99] give an elegant two-dimensional generalization of the exponentially-fitted trial space used in one dimension (see also [Wan99]). They

approximate b by a piecewise constant $\bar{b}$, and on each triangle consider trial functions of the form

$$\phi(x) = k_1 + k_2 e^{\bar{b}\cdot x} + k_3(\bar{b}\times x)\cdot e_z,$$

where x is any point in the triangle, e_z is a unit vector perpendicular to the (x,y)-plane, and the k_i are arbitrary constants. An error bound (which depends on ε) is proved for the case when this trial space is used in a Galerkin method. These trial functions were also proposed independently by Babuška et al. [BCO94]. Numerical results are given in [SGG99] for this method and for a Petrov-Galerkin method with the same trial space and piecewise linear test functions. In [SS98] it is shown that this Petrov-Galerkin method is essentially equivalent to the nonstandard upwinded scheme used in the well-known PLTMG package [Ban98, BBFS90]; the method is also generalized by imbedding it in a family of such methods, which enhances its ability to compute layers accurately on coarse meshes. ♣

Exponentially-fitted finite elements are closely related (see [FNS98]) to the residual-free bubble finite element method of Section 3.2.3.

Operator-fitted methods have two drawbacks. First, only low-order results are known for problems posed in more than one dimension. Second, when a parabolic boundary layer is present, Shishkin's obstacle result (see Remark II.3.22 and page 267) applies: no operator-fitted method on a rectangular equidistant mesh can achieve a positive order of convergence in the discrete maximum norm, uniformly in ε. Consequently we now switch our attention to methods based on an *a priori* choice of a special mesh, which can circumvent both of these failings.

3.5.2 Layer-Adapted Meshes

The rectangular meshes considered in this section are tensor products of the layer-adapted one-dimensional meshes (such as Shishkin-type meshes) that we discussed in Section I.2.4.2. The use of such meshes is reasonable when the domain is rectangular and the hypotheses of the problem ensure that the only layers appearing in its solution are boundary layers along certain sides of the domain together with corner layers, so interior layers are excluded. This is the case for instance in the problem that is analysed in Theorem 1.26.

Triangular meshes are constructed from these rectangular meshes by bisecting each mesh rectangle into two triangles.

For non-rectangular domains, see Remarks 3.121 and 3.123 below.

Exponential Boundary Layers

We describe a Shishkin mesh for problem (3.111). Let N be an even integer. For brevity we assume that one has N mesh intervals in each coordinate

direction. One could more generally obtain the same results with Shishkin meshes having M mesh intervals in the x-direction and N mesh intervals in the y-direction, where $\max\{M/N, N/M\} \le C$ for some C; and one could work with Shishkin-type meshes in each coordinate direction, with a similar restriction on the numbers M and N of mesh intervals.

Set

$$\sigma_x = \min\{1/2, (k\varepsilon/\beta_1)\ln N\} \quad \text{and} \quad \sigma_y = \min\{1/2, (k\varepsilon/\beta_2)\ln N\}, \quad (3.122)$$

where the positive constant k will be specified later for particular results (usually k is greater than or equal to the rate of convergence attained, so typically $1 \le k \le 3$ for low-order methods; recall Remarks I.2.99 and I.2.104). In fact we assume that $\sigma_x = (k\varepsilon/\beta_1)\ln N$ and $\sigma_y = (k\varepsilon/\beta_2)\ln N$, as otherwise one has $N \ge \min\{e^{\beta_1/(5\varepsilon)}, e^{\beta_2/(5\varepsilon)}\}$, which is very unlikely in practice and would enable one to analyse the method by means of classical techniques.

Define the mesh transition points on the x- and y-axes to be $1 - \sigma_x$ and $1 - \sigma_y$ respectively. Divide each of $[0, 1 - \sigma_x]$ and $[1 - \sigma_x, 1]$ into $N/2$ equal subintervals; similarly divide the y-interval $[0,1]$ using σ_y; then take the tensor product of these 1-dimensional meshes to get a 2-dimensional rectangular Shishkin mesh. See Figure 2.1. Set

$$\Omega_0 = [0, 1 - \sigma_x] \times [0, 1 - \sigma_y], \qquad \Omega_1 = [1 - \sigma_x, 1] \times [0, 1 - \sigma_y],$$
$$\Omega_2 = [0, 1 - \sigma_x] \times [1 - \sigma_y, 1], \qquad \Omega_{12} = [1 - \sigma_x, 1] \times [1 - \sigma_y, 1].$$

The mesh is coarse on Ω_0, fine on Ω_{12}, and highly anisotropic ("coarse/fine") on $\Omega_1 \cup \Omega_2$. Here all coarse mesh widths are $\mathcal{O}(N^{-1})$ while the fine mesh widths are $\mathcal{O}(\varepsilon N^{-1}\ln N)$.

Let $\tilde{\Omega}$ be any measurable subset of Ω. Write $\|\cdot\|_{m,p,\tilde{\Omega}}$ and $|\cdot|_{m,p,\tilde{\Omega}}$ for the norm and strongest seminorm in the Sobolev space $W^{m,p}(\tilde{\Omega})$. When $\tilde{\Omega} = \Omega$ and/or $p = 2$ they can be omitted from the notation.

Remark 3.105. (Anisotropic interpolation estimates) In every finite element analysis an interpolation error estimate is needed. Let K be a mesh element (rectangle or triangle). Let $v \in W^{m,p}(K)$, where $m \ge 1$ and $1 \le p \le \infty$. Let v^I denote the nodal interpolant (linear or bilinear) of v, where we assume that $mp > 2$ so that this interpolant is well defined. Then the classical interpolation error bound [Cia02, Theorem 3.1.6] is

$$\|v - v^I\|_{0,p,K} \le Ch^2 \sum_{|\alpha|=2} \|D^\alpha v\|_{0,p,K}, \qquad (3.123)$$

where h is the diameter of the element. Here α is the multi-index (α_1, α_2), $|\alpha| := \alpha_1 + \alpha_2$, and

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

For classical problems it follows that $\|v - v^I\|_{0,p,K} = \mathcal{O}(h^2\|u\|_{2,p,K})$ for each K, and summing over all K produces the familiar bound $\|v - v^I\|_{0,p} = \mathcal{O}(h^2)$. But with the 2-dimensional Shishkin mesh, on a highly anisotropic mesh rectangle K (in Ω_1 say, so the element is coarse in the y direction and fine in the x direction) the classical bound (3.123) and the derivative bounds of Theorem 1.26 yield

$$\|v - v^I\|_{0,\infty,K} \leq Ch^2 \max_{(x,y)\in K} \left[1 + \varepsilon^{-1}e^{-\beta_1(1-x)/\varepsilon}\right] = \mathcal{O}(h^2\varepsilon^{-1})$$

which is unsatisfactory because ε can be very small. This failure occurs because the multiplier h^2 in (3.123) takes no account of the small mesh width in the direction in which the layer is decaying, even though one would expect this feature to improve the accuracy of the interpolation.

Thus on Shishkin meshes one must replace (3.123) by the sharp anisotropic interpolation estimates of [AD92, Ape99] for general meshes, which we now describe.

Suppose that each element K (triangle or rectangle) of a mesh is contained in a rectangle with side lengths (h_x, h_y) and contains a rectangle with side lengths $(C_2 h_x, C_2 h_y)$ for some fixed constant $C_2 > 0$. In the case of triangles, assume also a maximum angle condition: the interior angles of every mesh triangle are bounded away from π. (Triangular Shishkin meshes have maximum angle $\pi/2$ and consequently satisfy this condition.) Then there exists a constant C such that

$$\|v - v^I\|_{0,p,K} \leq C \sum_{|\alpha|=m} h^\alpha \|D^\alpha v\|_{0,p,K} \quad \text{for } m = 1, 2, \qquad (3.124a)$$

$$\|\partial_x(v - v^I)\|_{0,p,K} \leq C \sum_{|\alpha|=1} h^\alpha \|D^\alpha \partial_x v\|_{0,p,K}, \qquad (3.124b)$$

$$\|\partial_y(v - v^I)\|_{0,p,K} \leq C \sum_{|\alpha|=1} h^\alpha \|D^\alpha \partial_y v\|_{0,p,K}, \qquad (3.124c)$$

where we set $h^\alpha = h_x^{\alpha_1} h_y^{\alpha_2}$.

In the case of a Shishkin mesh, these bounds have a small multiplier h^α precisely when the corresponding derivative is large. This is a great improvement on (3.123). Furthermore, the right-hand side of the bound (3.124b) involves the derivatives v_{xx} and v_{xy} but v_{yy} does not appear, which is crucial in certain calculations.

More sophisticated interpolants (Clément, Scott-Zhang) are sometimes used, e.g., when the interpolated function is not defined pointwise; these are considered in Apel's monograph [Ape99]. See also (III.3.157) and Remark IV.3.1. ♣

To derive satisfactory interpolation error estimates from (3.124), it is helpful to have a decomposition of the solution u of (3.111) into smooth and layer

components that includes bounds on the derivatives of these components. We now give such a decomposition in a form used by several authors.

Assume that

$$u = S + E_1 + E_2 + E_{12}, \tag{3.125a}$$

and that there exists a constant C such that

$$\left| \frac{\partial^{i+j} S}{\partial x^i \partial y^j}(x, y) \right| \leq C, \tag{3.125b}$$

$$\left| \frac{\partial^{i+j} E_1}{\partial x^i \partial y^j}(x, y) \right| \leq C \varepsilon^{-i} e^{-\beta_1 (1-x)/\varepsilon}, \tag{3.125c}$$

$$\left| \frac{\partial^{i+j} E_2}{\partial x^i \partial y^j}(x, y) \right| \leq C \varepsilon^{-j} e^{-\beta_2 (1-y)/\varepsilon}, \tag{3.125d}$$

$$\left| \frac{\partial^{i+j} E_{12}}{\partial x^i \partial y^j}(x, y) \right| \leq C \varepsilon^{-(i+j)} e^{-(\beta_1 (1-x)+\beta_2 (1-y))/\varepsilon} \tag{3.125e}$$

for all $(x, y) \in \Omega$ and $0 \leq i + j \leq m$, where m is a non-negative integer that will be specified each time we invoke (3.125).

When $m = 2$ these bounds are weaker than the derivative bounds of Theorem 1.26 with $n = 2$. For $m = 3$ a comparison is more difficult because the bounds of (3.125) are symmetric in x and y, unlike those of Theorem 1.26; for example the derivative $(E_1)_{xxx}$ is then bounded in (3.125) and $(E_1)_{xxyy}$ is not, while the opposite is true with regard to Theorem 1.26 when $n = 2$.

The following pair of lemmas, which are based on [DR97, SO97], illustrate the different types of argument needed on different parts of the Shishkin mesh when estimating the interpolation error in various norms. Both lemmas hold true under the assumption that (3.125) is valid with $m = 2$, but we prove the first under weaker hypotheses.

Lemma 3.106. *Choose $k \geq 2$ in (3.122). Assume that (3.125) hold true with $m = 0$, and that there exists a constant C such that for all i and j with $i + j = 2$ and all $(x, y) \in \Omega$ one has*

$$\left| \frac{\partial^{i+j} S}{\partial x^i \partial y^j}(x, y) \right| \leq C, \tag{3.126a}$$

$$\left| \frac{\partial^{i+j} E_1}{\partial x^i \partial y^j}(x, y) \right| \leq C \varepsilon^{-i}, \tag{3.126b}$$

$$\left| \frac{\partial^{i+j} E_2}{\partial x^i \partial y^j}(x, y) \right| \leq C \varepsilon^{-j}, \tag{3.126c}$$

$$\left| \frac{\partial^{i+j} E_{12}}{\partial x^i \partial y^j}(x, y) \right| \leq C \varepsilon^{-(i+j)}. \tag{3.126d}$$

Then there exists a constant C such that

$$|(u - u^I)(x, y)| \leq \begin{cases} CN^{-2} & \text{if } (x, y) \in \Omega_0, \\ CN^{-2} \ln^2 N & \text{otherwise.} \end{cases} \tag{3.127}$$

Proof. First, by (3.123), (3.126a) and (3.126b) one gets $\|S - S^I\|_{0,\infty} \le CN^{-2}$ and

$$\|E_1 - E_1^I\|_{0,\infty,\Omega_{12}} \le C(\varepsilon N^{-1} \ln N)^2 \varepsilon^{-2} = CN^{-2} \ln^2 N.$$

Next, invoking (3.124a) and (3.126b) gives

$$
\begin{aligned}
\|E_1 - E_1^I\|_{0,\infty,\Omega_1} &\le C\big[(\varepsilon N^{-1} \ln N)^2 \|(E_1)_{xx}\|_{0,\infty,\Omega_1} \\
&\quad + (\varepsilon N^{-1} \ln N)N^{-1} \|(E_1)_{xy}\|_{0,\infty,\Omega_1} + N^{-2} \|(E_1)_{yy}\|_{0,\infty,\Omega_1}\big] \\
&\le C[(\varepsilon N^{-1} \ln N)^2 \varepsilon^{-2} + (\varepsilon N^{-1} \ln N)N^{-1}\varepsilon^{-1} + N^{-2}] \\
&\le CN^{-2} \ln^2 N.
\end{aligned}
$$

A direct application of (3.124a) on $\Omega_0 \cup \Omega_2$ fails because $(E_1)_{xx}$ is still large on part of this region (recall Remark I.2.93). Instead we call upon the decay guaranteed by (3.125c) and the choice of k in the mesh: $|E_1(x,y)| \le CN^{-2}$ for $(x,y) \in \Omega_0 \cup \Omega_2$. It follows that $|E_1^I(x,y)| \le CN^{-2}$ on the same domain. Combining all these results for E_1, one gets

$$|(E_1 - E_1^I)(x,y)| \le \begin{cases} CN^{-2} & \text{if } (x,y) \in \Omega_0 \cup \Omega_2, \\ CN^{-2} \ln^2 N & \text{otherwise.} \end{cases}$$

One can prove analogous results for $|(E_2 - E_2^I)(x,y)|$ and $|(E_{12} - E_{12}^I)(x,y)|$. From the decomposition (3.125a) and $u^I = S^I + E_1^I + E_2^I + E_{12}^I$ it then follows via a triangle inequality that (3.127) holds true. □

Lemma 3.107. *Choose $k \ge 2$ in (3.122). Assume that (3.125) hold true with $m = 2$. Then there exists a constant C such that*

$$\|u - u^I\|_0 \le CN^{-2} + C\varepsilon^{1/2}N^{-2} \ln^2 N; \tag{3.128a}$$

if $\varepsilon^{1/2} \le (\ln N)^{-2}$, then

$$\|u - u^I\|_0 \le CN^{-2} \tag{3.128b}$$

$$and \quad \|u - u^I\|_\varepsilon \le CN^{-1} \ln N. \tag{3.128c}$$

Proof. The local L_∞ bounds in the proof of Lemma 3.106 immediately yield

$$\|S - S^I\|_0 + \|E_1 - E_1^I\|_{0,\Omega_0 \cup \Omega_2} \le CN^{-2}.$$

Squaring (3.124a) and adding over all $K \subset \Omega_1$, then substituting the bounds from (3.125c) into each $\|\cdot\|_{0,\Omega_1}$ and evaluating the resulting integrals, one sees that

$$
\begin{aligned}
\|E_1 - E_1^I\|_{0,\Omega_1}^2 &\le C\big[(\varepsilon N^{-1} \ln N)^4 \|(E_1)_{xx}\|_{0,\Omega_1}^2 \\
&\quad + (N^{-1}\varepsilon N^{-1} \ln N)^2 \|(E_1)_{xy}\|_{0,\Omega_1}^2 + N^{-4} \|(E_1)_{yy}\|_{0,\Omega_1}^2\big] \\
&\le C\big[(\varepsilon N^{-1} \ln N)^4 \varepsilon^{-3} \\
&\quad + (\varepsilon N^{-2} \ln N)^2 \varepsilon^{-1} + N^{-4}N^{-1}\varepsilon \ln N\big] \\
&\le C\big[\varepsilon N^{-4} \ln^4 N + \varepsilon N^{-4} \ln^2 N + \varepsilon N^{-5} \ln N\big],
\end{aligned}
$$

so

$$\|E_1 - E_1^I\|_{0,\Omega_1} \leq C\varepsilon^{1/2} N^{-2} \ln^2 N.$$

Essentially the same calculation gives $\|E_1 - E_1^I\|_{0,\Omega_{12}} \leq C\varepsilon^{1/2} N^{-2} \ln^2 N$. Combining all these bounds yields

$$\|E_1 - E_1^I\|_0 \leq CN^{-2} + C\varepsilon^{1/2} N^{-2} \ln^2 N.$$

By kindred arguments we obtain

$$\|E_2 - E_2^I\|_0 + \|E_{12} - E_{12}^I\|_0 \leq CN^{-2} + C\varepsilon^{1/2} N^{-2} \ln^2 N.$$

The decompositions of u and u^I and a triangle inequality now produce (3.128a), from which (3.128b) is an immediate consequence.

For the final bound (3.128c), $\|\nabla(S - S^I)\|_0 \leq CN^{-1}$ follows from (3.124b), (3.124c) and (3.125b). The estimate (3.124b) also gives

$$\begin{aligned}
\|(E_1)_x - (E_1^I)_x\|_{0,\Omega_1} &\leq C\Big[(\varepsilon N^{-1} \ln N)\|(E_1)_{xx}\|_{0,\Omega_1} + N^{-1}\|(E_1)_{xy}\|_{0,\Omega_1}\Big] \\
&\leq C\Big[(\varepsilon N^{-1} \ln N)\varepsilon^{-3/2} + N^{-1}\varepsilon^{-1/2}\Big] \\
&\leq C\varepsilon^{-1/2} N^{-1} \ln N
\end{aligned} \qquad (3.129)$$

on substituting the bounds from (3.125c) and evaluating the integrals. A similar calculation yields

$$\|(E_1 - E_1^I)_x\|_{0,\Omega_{12}} \leq C\varepsilon^{-1/2} N^{-1} \ln N.$$

Next,

$$\begin{aligned}
\|(E_1)_x - (E_1^I)_x\|_{0,\Omega_0 \cup \Omega_2} &\leq \|(E_1)_x\|_{0,\Omega_0 \cup \Omega_2} + \|(E_1^I)_x\|_{0,\Omega_0 \cup \Omega_2} \\
&\leq \|(E_1)_x\|_{0,\Omega_0 \cup \Omega_2} + CN\|E_1^I\|_{0,\Omega_0 \cup \Omega_2},
\end{aligned}$$

where we used an inverse estimate that follows easily from transforming the classical inverse estimate [Cia02, Theorem 3.2.6] to our anisotropic elements. The first integral here is bounded, similarly to our previous calculations, by $C\varepsilon^{-1/2} N^{-2}$, while

$$N\|E_1^I\|_{0,\Omega_0 \cup \Omega_2} \leq N\|E_1^I\|_{0,\infty,\Omega_0 \cup \Omega_2} \leq N\|E_1\|_{0,\infty,\Omega_0 \cup \Omega_2} \leq CN^{-1},$$

since $k \geq 2$. Putting all these bounds together, one obtains

$$\|(E_1)_x - (E_1^I)_x\|_0 \leq C\varepsilon^{-1/2} N^{-1} \ln N.$$

The estimate $\|(E_1)_y - (E_1^I)_y\|_0 \leq C\varepsilon^{-1/2} N^{-1} \ln N$ is immediate from (3.124c) since $(E_1)_y$ is better behaved than $(E_1)_x$. One can similarly bound $\|\nabla(E_2 - E_2^I)\|_0$ and $\|\nabla(E_{12} - E_{12}^I)\|_0$; the only extra complication is that an inverse inequality and (3.125e) are invoked to give

$$\|(E_{12})_x - (E_{12}^I)_x\|_{0,\Omega_1} \le \|(E_{12})_x\|_{0,\Omega_1} + \|(E_{12}^I)_x\|_{0,\Omega_1}$$
$$\le \|(E_{12})_x\|_{0,\Omega_1} + C\varepsilon^{-1}N(\ln N)^{-1}\|E_{12}^I\|_{0,\Omega_1}$$
$$\le \|(E_{12})_x\|_{0,\Omega_1} + C\varepsilon^{-1}N(\ln N)^{-1}\left(\int_{\Omega_1} N^{-4}\right)^{1/2}$$
$$\le C\varepsilon^{-1/2}N^{-1}(\ln N)^{-1/2}.$$

Hence $\varepsilon^{1/2}\|\nabla(u - u^I)\|_0 \le CN^{-1}\ln N$. This inequality and (3.128b) yield (3.128c). □

Remark 3.108. An inspection of the proof of Lemma 3.107 shows that the pointwise derivative bounds of (3.125) are stronger than needed: the argument works under the weaker hypotheses that (3.125) holds true with $m = 0$ and that for $i + j = 2$ one has

$$\left\|\frac{\partial^{i+j}S}{\partial x^i\partial y^j}\right\|_0 \le C, \qquad \left\|\frac{\partial^{i+j}E_1}{\partial x^i\partial y^j}\right\|_0 \le C\varepsilon^{-i+1/2},$$
$$\left\|\frac{\partial^{i+j}E_2}{\partial x^i\partial y^j}\right\|_0 \le C\varepsilon^{-j+1/2}, \qquad \left\|\frac{\partial^{i+j}E_{12}}{\partial x^i\partial y^j}\right\|_0 \le C\varepsilon^{-i-j+1/2},$$

with moreover

$$\|(E_1)_x\|_{0,\Omega_0\cup\Omega_2} + \|(E_2)_y\|_{0,\Omega_0\cup\Omega_1} \le C\varepsilon^{-1/2}N^{-1},$$
$$\|(E_{12})_x\|_{0,\Omega\setminus\Omega_0} + \|(E_{12})_y\|_{0,\Omega\setminus\Omega_0} \le C\varepsilon^{-1/2}N^{-1}.$$

These are all L_2 bounds on derivatives.   ♣

The bounds of Lemma 3.107 are sharp, as can be seen by considering simple examples. The mild condition $\varepsilon^{1/2} \le (\ln N)^{-2}$ that is used to prove (3.128b) and (3.128c) is satisfied in the numerical solution of all typical convection-diffusion problems, and we shall ignore it when we invoke Lemma 3.107 in the future. Roos and Linß [RL99] generalize these lemmas to other layer-adapted meshes, writing the results in terms of the mesh-characterizing function ψ of Section I.2.4.2.

Define the bilinear form

$$B_{GAL}(v, w) = (\varepsilon\nabla v, \nabla w) + (b \cdot \nabla v, w) + (cv, w) \quad \forall v, w \in H^1(\Omega), \quad (3.130)$$

where $(\cdot, \cdot)$ denotes the $L_2(\Omega)$ inner product. Then (3.112) implies that

$$B_{GAL}(v, v) \ge \min\{1, \omega\}\|v\|_\varepsilon^2 \quad \forall v \in H_0^1(\Omega). \quad (3.131)$$

The following error bound from [SO97] (where only bilinears were considered) was the first uniform convergence result (in the sense of (3.115)) for a finite element method on a Shishkin mesh. A Galerkin method based on the bilinear form (3.130) was used.

Theorem 3.109. *Consider a Shishkin mesh with $k \geq 2$ in (3.122). Let V^N be the space of continuous piecewise linears or bilinears on this mesh that vanish on $\partial\Omega$. Define $u^N \in V^N$ by $B_{GAL}(u^N, v^N) = (f, v^N)$ for all $v^N \in V^N$. Assume that (3.125) holds true with $m = 2$. Then there exists a constant C such that*

$$\|u - u^N\|_\varepsilon \leq CN^{-1} \ln N. \tag{3.132}$$

Proof. We show how (3.132) depends on the interpolation properties of V^N. Write u^I for the nodal interpolant to u from V^N. By (3.131) and the Galerkin orthogonality property $B_{GAL}(u - u^N, v^N) = 0$ for all $v^N \in V^N$, one has

$$
\begin{aligned}
\min\{1, \omega\}\|u^I - u^N\|_\varepsilon^2 &\leq B_{GAL}(u^I - u^N, u^I - u^N) \\
&= B_{GAL}(u^I - u, u^I - u^N) \\
&\leq \varepsilon |u^I - u|_1 |u^I - u^N|_1 + |(b \cdot \nabla(u^I - u), u^I - u^N)| \\
&\quad + C\|u^I - u\|_0 \|u^I - u^N\|_0, \tag{3.133}
\end{aligned}
$$

where we recall that $|\cdot|_1$ is the $H^1(\Omega)$ seminorm. Now

$$
\begin{aligned}
|(b \cdot \nabla(u^I - u), u^I - u^N)| &= |-(b(u^I - u), \nabla \cdot (u^I - u^N)) \\
&\quad - (\nabla \cdot b(u^I - u), u^I - u^N)| \\
&\leq C\big[|(b(u^I - u), \nabla \cdot (u^I - u^N))| \\
&\quad + \|u^I - u\|_0 \|u^I - u^N\|_0\big].
\end{aligned}
$$

Using a standard inverse inequality,

$$
\begin{aligned}
|(b(u^I - u), \nabla \cdot (u^I - u^N))| &\leq C\big[\|u^I - u\|_{0,\Omega_0} \|\nabla \cdot (u^I - u^N)\|_{0,\Omega_0} \\
&\quad + \|u^I - u\|_{0,\Omega\setminus\Omega_0}\big]\|\nabla \cdot (u^I - u^N)\|_{0,\Omega\setminus\Omega_0} \\
&\leq C\big[N\|u^I - u\|_{0,\Omega_0} \|u^I - u^N\|_{0,\Omega_0} \\
&\quad + \|u^I - u\|_{0,\Omega\setminus\Omega_0} \|\nabla \cdot (u^I - u^N)\|_{0,\Omega\setminus\Omega_0}\big]
\end{aligned}
$$

Combining these inequalities with (3.133) leads after some typical finite element analysis manipulation to

$$\|u^I - u^N\|_\varepsilon \leq C\big[\varepsilon^{1/2}|u^I - u|_1 + \varepsilon^{-1/2}\|u^I - u\|_{0,\Omega\setminus\Omega_0} + N\|u^I - u\|_0\big]. \tag{3.134}$$

Lemma 3.107 furnishes the bounds $\varepsilon^{1/2}|u^I - u|_1 \leq CN^{-1} \ln N$ and $\|u^I - u\|_0 \leq CN^{-2}$; also, (3.127) implies that

$$\varepsilon^{-1/2}\|u^I - u\|_{0,\Omega\setminus\Omega_0} \leq C\varepsilon^{-1/2}\left[\int_{\Omega\setminus\Omega_0} (N^{-2} \ln^2 N)^2\right]^{1/2} \leq CN^{-2} \ln^{5/2} N.$$

Substituting these inequalities into (3.134), one obtains

$$\|u^I - u^N\|_\varepsilon \leq CN^{-1} \ln N \tag{3.135}$$

and (3.132) then follows from (3.128c) and a triangle inequality. $\square$

Figure 3.11 shows a computed solution that is typical of those obtained when the Galerkin method, with bilinear trial functions, is used to solve (3.111) on a Shishkin mesh.
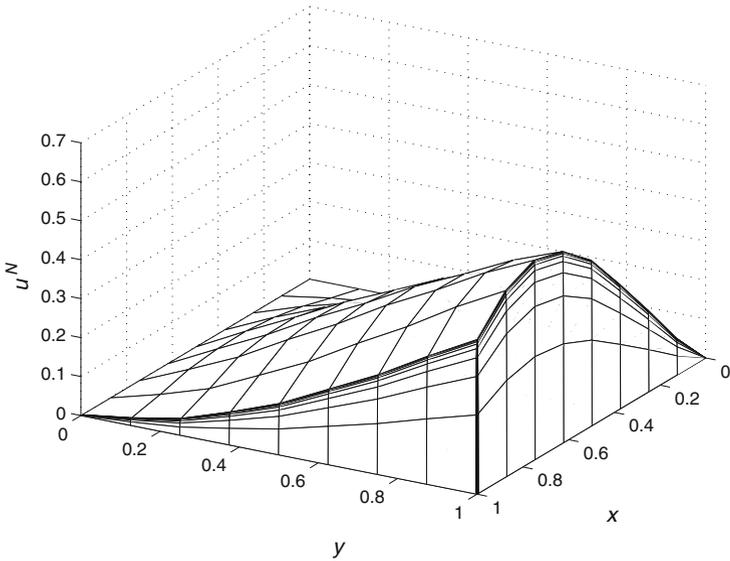
Fig. 3.11. Problem with two exponential layers; solution computed by Galerkin finite element method using bilinears on a Shishkin mesh

This figure is in stark contrast to the solution to the same problem, shown in Figure 3.12, computed using the Galerkin method with bilinears on an equidistant mesh.

Remark 3.110. In Theorem 3.109, the hypothesis that (3.125) holds true with $m = 2$ can be weakened. Instead of this hypothesis, assume that the bounds listed in Remark 3.108 are valid and that for $i + j = 1$ one has

$$\left| \frac{\partial^{i+j} E_1}{\partial x^i \partial y^j}(x, y) \right| \leq C\varepsilon^{-i} \text{ on } \Omega_1 \cup \Omega_{12},$$

$$\left| \frac{\partial^{i+j} E_2}{\partial x^i \partial y^j}(x, y) \right| \leq C\varepsilon^{-j} \text{ on } \Omega_2 \cup \Omega_{12},$$

$$\left| \frac{\partial^{i+j} E_{12}}{\partial x^i \partial y^j}(x, y) \right| \leq C\varepsilon^{-1} \text{ on } \Omega_{12}.$$

Then the proof of Lemma 3.106 (modified by invoking (3.124a) with $m = 1$ instead of $m = 2$) yields $|(u - u^I)(x, y)| \leq CN^{-1} \ln N$ on $\Omega \setminus \Omega_0$, which leads
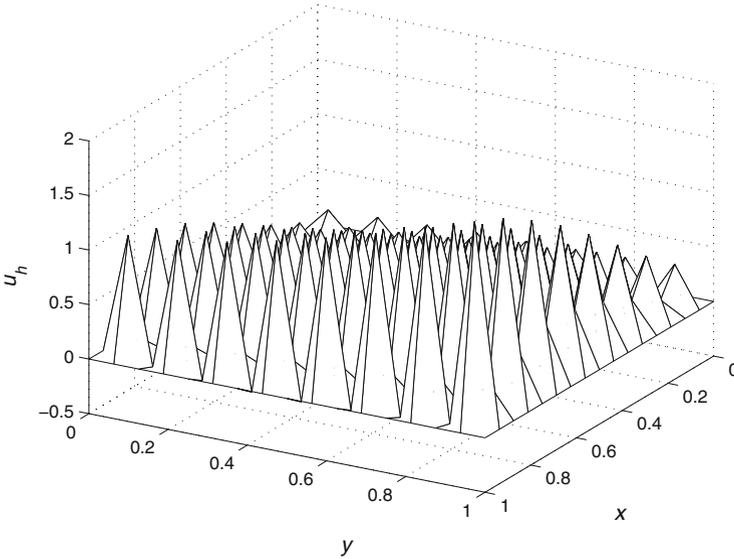
Fig. 3.12. Problem with two exponential layers; solution computed by Galerkin finite element method using bilinears on an equidistant mesh

to a slightly weaker result in Theorem 3.109: $\|u - u^N\|_\varepsilon \le CN^{-1}\ln^{3/2} N$. (In fact a more detailed argument will again yield (3.132).)   ♣

   The bound (3.132) could be written as $\|u - u^N\|_\varepsilon \le Ch|\ln h|$, but for layer-adapted meshes with N intervals in each coordinate direction one usually states convergence results in terms of N. This theorem is generalized in [RL99, Lin00a] from Shishkin meshes to more general layer-adapted meshes; in particular when certain features of the Shishkin and Bakhvalov meshes are combined, it is shown that $\|u - u^N\|_\varepsilon \le CN^{-1}$. In [DL06] the same Galerkin method with bilinears is used on a tensor-product recursively-defined graded mesh but the convergence result ($\|u - u^N\|_\varepsilon \le CN^{-1}|\ln\varepsilon|^2$) is not quite uniform in ε.

Remark 3.111. It is noteworthy that uniform convergence is achieved in this theorem without stabilizing the method by means of the bilinear form or the choice of finite element spaces: the only mechanism preventing excessive oscillations in the solution is the Shishkin mesh.

   In fact the bound (3.135) is not sharp for bilinears: if $k \ge 2.5$ in (3.122), so that the layer components have decayed sufficiently when they reach the coarse mesh, then [Lin00b, Zha03] one has

$$\|u^I - u^N\|_\varepsilon \le CN^{-2}\ln^2 N \tag{3.136}$$

and hence $\|u - u^N\|_0 \leq CN^{-2} \ln^2 N$. Superclose properties such as (3.136) will be discussed later in this section. A generalization of (3.136) to general layer-adapted meshes is proved in [Lin00b], where it is also shown that

$$|(u - u^N)(x,y)| \leq \begin{cases} C \min\{1, \varepsilon^{-1/2} N^{-1} \ln N\} N^{-1} \ln^{5/2} N & \text{if } (x,y) \in \Omega_0, \\ CN^{-3/2} \ln^3 N & \text{otherwise,} \end{cases}$$

provided that $\varepsilon \leq CN^{-1}$ for some C.

Despite the accuracy guaranteed by these results, the computed solution exhibits small oscillations, as can be inferred from [LS01b, Figure 8]. More seriously, the stiffness matrix has eigenvalues with large imaginary parts, so standard iterative methods do not solve the discrete linear system in an efficient manner and one must resort to direct factorization, as was done in [LS01b]; see also [ESW05, p.197].

While Theorem 3.109 holds true for both bilinears and linears, numerical results [LS01b] reveal that bilinears give a much better pointwise convergence rate than linears inside the boundary layers. See Remark 3.118.   ♣

To obtain a stabler method that nevertheless generates an accurate solution, we turn to the SDFEM of Section 3.2.1, using piecewise bilinears. Let the user-chosen piecewise constant SDFEM parameter δ_K be specified on the Shishkin mesh by

$$\delta_K = \begin{cases} N^{-1} & \text{if } K \subset \Omega_0 \text{ and } \varepsilon \leq N^{-1}, \\ \varepsilon^{-1} N^{-2} & \text{if } K \subset \Omega_0 \text{ and } \varepsilon > N^{-1}, \\ 0 & \text{otherwise.} \end{cases}$$

The choices of δ_K made here on Ω_0, where the mesh is shape-regular, are the values recommended in (3.38). On the highly anisotropic rectangles or triangles in $\Omega \setminus \Omega_0$, we cannot invoke (3.38); here one takes $\delta_K = 0$ as heuristic analysis and numerical experience show that this gives satisfactory results.

Let u^N be the piecewise bilinear SDFEM solution on the Shishkin mesh, where $k \geq 2.5$. Then the following result is proved in [ST03].

Theorem 3.112. *Let assumptions (3.125c)–(3.125e) hold true for $m = 3$. Assume also that S lies in the Sobolev space $H^3(\Omega)$ with $\|S\|_{H^3(\Omega)} \leq C$. Write u^I for the nodal bilinear interpolant to u. Then there exists a constant C such that*

$$|||u^I - u^N|||_{SD} \leq C(\varepsilon N^{-3/2} + N^{-2} \ln^2 N), \tag{3.137}$$

where the norm $||| \cdot |||_{SD}$ was defined in Section 3.2.1.

Proof. To prove this result, one uses the coercivity inequality of Lemma 3.25:

$$\frac{1}{2} |||u^N - u^I|||^2_{SD}$$
$$\leq B_{SD}(u^N - u^I, u^N - u^I) \quad = B_{SD}(u - u^I, u^N - u^I)$$
$$= B_{GAL}(u - u^I, u^N - u^I) + B_{STAB}(u - u^I, u^N - u^I), \tag{3.138}$$

where, writing $(\cdot,\cdot)_K$ for the $L_2(K)$ inner product over each mesh element K,

$$B_{STAB}(w,v) = \sum_{K \subset \Omega_0} \delta_K(-\varepsilon \Delta w + b \cdot \nabla w + cw, b \cdot \nabla v)_K$$

for all $(w,v) \in \tilde{H}^1(\Omega) \times H^1(\Omega)$ and $\tilde{H}^1(\Omega)$ denotes the set of functions in $H^1(\Omega)$ that lie in $H^2(K)$ for each K.

The two terms in (3.138) are bounded by a calculation that is too long to reproduce in full here, but we shall give a sample to impart some of the flavour of the argument.

After splitting $u = S + E_1 + E_2 + E_{12}$ as in (3.125a), when bounding $B_{STAB}(u - u^I, u^N - u^I)$ one must deal with $B_{STAB}(E - E^I, v^N)$, where E can be E_1, E_2 or E_{12} and $v^N = u^N - u^I$ (in fact v^N can be any piecewise bilinear function in what follows). In the case $\varepsilon \leq N^{-1}$ one gets

$$\begin{aligned}
&\left|B_{STAB}(E - E^I, v^N)\right| \\
&\leq CN^{-1}\left[\varepsilon\|\Delta E\|_{1,\Omega_0} + \|\nabla E\|_{1,\Omega_0}\right]\|b \cdot \nabla v^N\|_{0,\infty,\Omega_0} \\
&\quad + C\,N^{-1/2}\left(\|\nabla E^I\|_{0,\Omega_0} + \|E - E^I\|_{0,\Omega_0}\right)|||v^N|||_{SD}. \qquad (3.139)
\end{aligned}$$

Clearly $\|E\|_{0,\infty,\Omega_0} \leq CN^{-5/2}$ follows from the decay properties (3.125c)–(3.125d) and the choice of k. Consequently $\|E^I\|_{0,\infty,\Omega_0} \leq CN^{-5/2}$ and hence $\|E - E^I\|_{0,\Omega_0} \leq CN^{-5/2}$. By a standard inverse inequality, $\|\nabla E^I\|_{0,\Omega_0} \leq CN\|E^I\|_{0,\Omega_0} \leq CN^{-3/2}$. Thus in (3.139) one obtains

$$CN^{-1/2}(\|\nabla E^I\|_{0,\Omega_0} + \|E - E^I\|_{0,\Omega_0}) \leq CN^{-2}.$$

For $E = E_1$ (the proof for the other layer functions is similar), inequality (3.125c) yields

$$\begin{aligned}
\varepsilon\|\Delta E_1\|_{1,\Omega_0} + \|\nabla E_1\|_{1,\Omega_0} &\leq C\varepsilon^{-1}\int_0^{1-\sigma_y}\int_0^{1-\sigma_x} e^{-\beta_1(1-x)/\varepsilon}\,dx\,dy \\
&\leq Ce^{-\beta_1\sigma_x/\varepsilon} \\
&\leq C\,N^{-5/2}.
\end{aligned}$$

Invoking the inverse inequality $\|b \cdot \nabla v^N\|_{0,\infty,\Omega_0} \leq CN\|b \cdot \nabla v^N\|_{0,\Omega_0}$ and the bounds just proved in (3.139), we finally arrive at

$$\begin{aligned}
\left|B_{STAB}(E - E^I, v^N)\right| &\leq CN^{-5/2}\|b \cdot \nabla v^N\|_{0,\Omega_0} + CN^{-2}|||v^N|||_{SD} \\
&\leq CN^{-2}|||v^N|||_{SD},
\end{aligned}$$

which suffices when proving (3.137).   □

The other main ingredients in the proof of (3.137) are the useful identities of Lin [Lin91], which are used in [Lin00b, Zha03] – see also [GRS07] – to

sharpen the Galerkin bound (3.135) to (3.136). For each mesh rectangle K, set

$$G_K(x) = \frac{1}{2}\left[(x - x_K)^2 - \left(\frac{h_{x,K}}{2}\right)^2\right], \quad F_K(y) = \frac{1}{2}\left[(y - y_K)^2 - \left(\frac{h_{y,K}}{2}\right)^2\right].$$

Denote the east, north, west and south edges of K by $l_{i,K}$ for $i = 1, \ldots, 4$ respectively.

Lemma 3.113. (Lin identities) *Let K be a mesh rectangle with sides parallel to the coordinate axes. Let $w \in H^3(K)$ and let w^I be its bilinear nodal interpolant on K. Then for each bilinear function v^N defined on K one has*

$$\int_K (w - w^I)_x v_x^N \, dx\, dy = \int_K w_{xyy}\left(F_K v_x^N - \frac{1}{3}\left(F_K^2\right)' v_{xy}^N\right) dx\, dy,$$

$$\int_K (w - w^I)_x v_y^N \, dx\, dy = \int_K \left(F_K w_{xyy}(v_y^N - G_K' v_{xy}^N) + G_K w_{xxy} v_x^N\right) dx\, dy$$
$$- \int_{l_{2,K}} G_K w_{xx} v_x^N \, dx + \int_{l_{4,K}} G_K w_{xx} v_x^N \, dx,$$

$$\int_K (w - w^I)_y v_x^N \, dx\, dy = \int_K \left(G_K w_{xxy}(v_x^N - F_K' v_{xy}^N) + F_K w_{xyy} v_y^N\right) dx\, dy$$
$$- \int_{l_{1,K}} F_K w_{yy} v_y^N \, dy + \int_{l_{3,K}} F_K w_{yy} v_y^N \, dy,$$

$$\int_K (w - w^I)_y v_y^N \, dxdy = \int_K w_{xxy}\left(G_K v_y^N - \frac{1}{3}\left(G_K^2\right)' v_{xy}^N\right) dx\, dy.$$

Proof. Start from the right-hand side of each identity. Since $(w^I)_{xx}$, $(w^I)_{yy}$ and all third-order derivatives of w^I vanish, these terms can be introduced at appropriate places in the right-hand side; then one integrates by parts and takes into consideration the definitions of F_K and G_K. For more details see [Lin91] or [Zha03]. $\square$

Zhang [Zha03] gives further identities of this type, originally from [Lin91, LY96], for the integrals $\int_K (w - w^I)_x v^N \, dx\, dy$ and $\int_K (w - w^I)_y v^N \, dx\, dy$.

As an illustration of the power of Lemma 3.113, consider the estimate

$$|\varepsilon(\nabla(u^I - u), \nabla w^N)| \leq \varepsilon |u^I - u|_1 |w^N|_1 \leq C(N^{-1} \ln N)\varepsilon^{1/2}|w^N|_1,$$

which (with $w^N = u^I - u^N$) was used in the derivation of (3.135). If u is replaced by E_1, this still produces the same final bound since (3.129) delivers only

$$\left|\varepsilon \int_{\Omega_1} (E_1^I - E_1)_x w_x^N\right| \leq \varepsilon \|(E_1^I)_x - (E_1)_x\|_{0,\Omega_1} \|w_x^N\|_{0,\Omega_1}$$

$$\leq C(N^{-1} \ln N)\varepsilon^{1/2}\|w_x^N\|_{0,\Omega_1}. \qquad (3.140)$$

On the other hand Lemma 3.113 gives

$$\left| \varepsilon \int_{\Omega_1} (E_1^I - E_1)_x w_x^N \right|$$

$$= \varepsilon \left| \sum_{K \subset \Omega_1} \left(-F_K w_x^N + \frac{1}{3}(F_K^2)' w_{xy}^N, (E_1)_{xyy} \right)_K \right|$$

$$\leq C\varepsilon \sum_{K \subset \Omega_1} \left(N^{-2}|w_x^N| + N^{-3}|w_{xy}^N|, |(E_1)_{xyy}| \right)_K$$

$$\leq C\varepsilon \sum_{K \subset \Omega_1} N^{-2} \left(\|w_x^N\|_{0,K} + N^{-1}\|w_{xy}^N\|_{0,K} \right) \|(E_1)_{xyy}\|_{0,K}$$

$$\leq C\varepsilon N^{-2} \sum_{K \subset \Omega_1} \|w_x^N\|_{0,K} \|(E_1)_{xyy}\|_{0,K}$$

$$\leq C\varepsilon N^{-2} \|w_x^N\|_{0,\Omega_1} \|(E_1)_{xyy}\|_{0,\Omega_1},$$

where we used the inverse inequality $\|w_{xy}^N\|_{0,K} \leq CN\|w_x^N\|_{0,K}$, which follows from a transformation of the classical inverse estimate to K. Hence

$$\left| \varepsilon \int_{\Omega_1} (E_1^I - E_1)_x w_x^N \right| \leq CN^{-2} \varepsilon^{1/2} \|w_x^N\|_{0,\Omega_1}$$

by (3.125c). This is a gain of a full order of convergence over (3.140).

Remark 3.114. When Lemma 3.113 is applied in the error analysis of any finite element method, one replaces w by various components of u. This requires a knowledge of the third-order derivatives of the components of u, unlike the analysis leading for instance to (3.135), which uses only second-order derivatives. Thus sharper bounds such as (3.136) can be proved only when one has extra regularity of the components of the solution u.   ♣

Combining (3.137) with (3.128c) gives immediately

$$\|u - u^N\|_\varepsilon \leq CN^{-1} \ln N \qquad (3.141)$$

for the SDFEM solution u^N. While this rate is optimal, it is a lower rate of convergence than in (3.137) even though the norm $\|\cdot\|_\varepsilon$ is weaker than $\||\cdot\||_{SD}$. This discrepancy will be exploited later when a postprocessing of the computed solution u^N is shown to yield a more accurate approximation of u with respect to $\|\cdot\|_\varepsilon$.

Remark 3.115. For the L_2 norm, (3.137) and (3.128b) yield

$$\|u - u^N\|_0 \leq C\left(\varepsilon N^{-3/2} + N^{-2} \ln^2 N\right),$$

so

$$\|u - u^N\|_0 \leq CN^{-2} \ln^2 N \quad \text{if} \quad \varepsilon \leq N^{-1/2} \ln^2 N,$$

which by (3.128b) is optimal up to the factor $\ln^2 N$. It is likely that for more general problems one could use cut-off functions to prove a similar result for rectangular locally uniform meshes in regions that extend downstream from an inflow boundary and on which the solution u is smooth. Thus in the SDFEM piecewise bilinears attain the same order of convergence in L_2 as the interpolation error, unlike piecewise linears (see Remark 3.118 below).   ♣

Superconvergence

The term "superconvergence" is used in different ways by different authors, so we begin with some definitions to establish our terminology.

Let u be the solution of a boundary value problem and u^N its computed solution in some finite-dimensional finite element space S^N. Suppose that the error of the finite element method, measured in some norm or seminorm $\|\cdot\|$, satisfies $\|u - u^N\| \leq CN^{-\alpha}$ for some constant $\alpha > 0$.

- If there exists $u^I \in S^N$ that is an interpolant (in some sense) or projection of u and for which $\|u^I - u^N\| \leq CN^{-\beta}$ for some constant $\beta > \alpha$, we say that the finite element method has the *superclose property*.
- If at special known points in elements (e.g., barycentres) the rate of convergence, measured in some discrete norm $\|\cdot\|_d$, is greater than is implied by the bound $\|u - u^N\| \leq CN^{-\alpha}$, we say that u^N is $\|\cdot\|_d$ *superconvergent*.
- If there is a higher-order finite element space $\tilde{S}^N$ and an interpolant or projection $\tilde{u}^N \in \tilde{S}^N$ of u^N such that $\|u - \tilde{u}^N\| \leq CN^{-\gamma}$ for some constant $\gamma > \alpha$, we say that $\tilde{u}^N$ is *interpolantwise superconvergent*.

For a general introduction to superconvergence in its various forms and a summary of known results for classical problems, see [BK01, LL06, Wah95] and their references.

Let us examine again the convection-diffusion problem (3.111), whose solution has two exponential boundary layers at the sides $x = 1$ and $y = 1$ of the unit square Ω and an exponential corner layer at the point $(1,1)$. Consider the Galerkin finite element method based on (3.130) and using piecewise bilinears on a tensor-product Shishkin mesh. In (3.136) we quoted the result from [Lin00b, Zha03] that $\|u^I - u^N\|_\varepsilon \leq CN^{-2} \ln^2 N$, where u^I is the nodal interpolant of u and the energy norm $\|\cdot\|_\varepsilon$ is defined in (3.114). A triangle inequality and (3.128c) – which is in general sharp – then yield $\|u - u^N\|_\varepsilon \leq CN^{-1} \ln N$. Thus the computed solution has the superclose property with respect to the nodal interpolant and $\|\cdot\|_\varepsilon$.

These bounds are proved under the assumption that (3.125) holds true with $m = 2$. Under the additional hypothesis that (3.125b) holds true for $i + j = 3$, Zhang [Zha03] also proves that

$$\|u - u^N\|_{\varepsilon,d} \leq C(\varepsilon N^{-3/2} + N^{-2} \ln^2 N). \tag{3.142}$$

Here $\|\cdot\|_{\varepsilon,d}$ is a discrete but weaker analogue of $\|\cdot\|_\varepsilon$ defined by

$$\|w\|_{\varepsilon,d} := \left[\varepsilon \sum_K (\text{area } K)|\nabla w(x_K, y_K)|^2 + \|w\|_0^2\right]^{1/2}, \tag{3.143}$$

where the sum is over all mesh rectangles K and (x_K, y_K) is the barycentre of K. One can show that $\|u - u^I\|_{\varepsilon,d} \le CN^{-2}\ln^2 N$. The improvement in the rate of convergence of (3.142) over $\|u - u^N\|_\varepsilon \le CN^{-1}\ln N$ is due to discrete-L_2 superconvergence of ∇u^N at the barycentres of the mesh rectangles.

In [ST03] some related results are established for the SDFEM solution u^N, computed on a Shishkin mesh. Inequality (3.137), when compared with (3.128c), shows that the computed solution has the superclose property. When $\varepsilon \le N^{-1}$, so $\delta_K = N^{-1}$ for $K \subset \Omega_0$, one has in particular

$$\left(\sum_{K \subset \Omega_0} \|b \cdot \nabla(u^I - u^N)\|_{0,K}^2\right)^{1/2} \le CN^{-3/2}\ln^2 N,$$

although $\left[\sum_{K \subset \Omega_0} \|b \cdot \nabla(u - u^I)\|_{0,K}^2\right]^{1/2} \le CN^{-1}$ is the best possible general result predicted by approximation theory. If, imitating (3.143), the error is measured in the discrete analogue of $||| \cdot |||_{SD}$ defined by

$$|||v|||_{SD,d} = \left[\sum_{K \subset \Omega} \varepsilon (\text{area } K)|\nabla v(x_K, y_K)|^2 \right.$$
$$\left. + \sum_{K \subset \Omega_0} \delta_K |(\text{area } K)(b \cdot \nabla v)(x_K, y_K)|^2 + \|v\|_0^2\right]^{1/2},$$

then [ST03, Theorem 5.3]

$$|||u - u^N|||_{SD,d} \le C\left(\varepsilon N^{-3/2} + N^{-2}\ln^2 N\right)$$

although (3.137) and (3.128c) yield only $\|u - u^N\|_\varepsilon \le CN^{-1}\ln N$; the SD-FEM attains discrete-L_2 superconvergence of the gradient of the error at the barycentres of the mesh rectangles. This increased order of convergence occurs because $||| \cdot |||_{SD,d}$ is a weaker norm than $||| \cdot |||_{SD}$ and one can prove that $|||u - u^I|||_{SD,d} \le CN^{-2}\ln^2 N$ – compare (3.128c).

Postprocessing and Recovery

Ainsworth and Oden [AO00] present a general theory of recovery operators that encompasses some of the special cases to be discussed below.

We shall attain a higher order of convergence by applying a local postprocessing technique to a computed solution u^N to construct a new discrete solution Pu^N in a higher-order space for which $\|u - Pu^N\|_\varepsilon \ll \|u - u^N\|_\varepsilon$, i.e., Pu^N is interpolantwise superconvergent. Our analysis of this approach requires u^N to possess a superclose property.

Postprocessing is applied in the following way in [ST03] to the solution $u^N \in V^N$ computed by the SDFEM, where V^N is the space of piecewise bilinears on a Shishkin mesh. Consider a family of Shishkin meshes $\mathcal{T}_N$ with meshpoints (x_i, y_j) for $i, j = 0, \ldots, N$, where we require $N/2$ to be even.
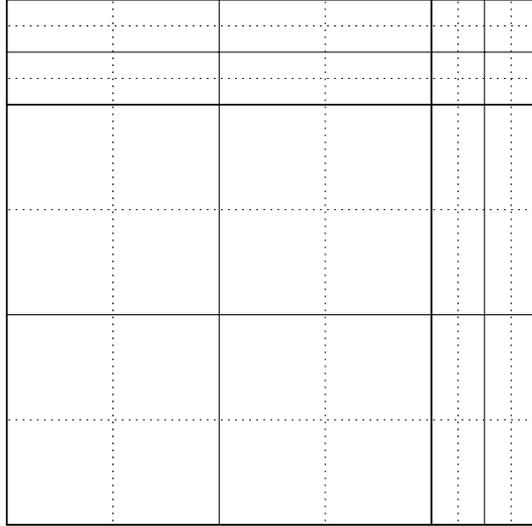
Fig. 3.13. Macroelements formed from four mesh rectangles

Form a coarser mesh composed of disjoint macrorectangles M, each comprising 4 mesh rectangles from $\mathcal{T}_N$, where M belongs to only one of the four domains Ω_0, Ω_1, Ω_2, and Ω_{12}. See Figure 3.13. Associate with each macrorectangle M an interpolation operator $P_M : C(\bar{M}) \to Q_2(M)$ defined by the standard biquadratic interpolation at the barycentre, nodes, and midpoints of edges of the macrorectangle. Then P_M can be extended to a continuous global interpolation operator $P : C(\bar{\Omega}) \to W^N$, where W^N is the space of piecewise biquadratic finite elements, by setting

$$(Pv)\Big|_{M} := P_M(v|_M) \quad \forall M.$$

It is straightforward to derive the consistency property

$$P(v^I) = P(v) \quad \forall v \in C(\bar{\Omega}) \tag{3.144a}$$

and the stability bound

$$\|Pv^N\|_\varepsilon \leq C \|v^N\|_\varepsilon \quad \forall v^N \in V^N. \tag{3.144b}$$

A detailed investigation similar to the proof of Lemma 3.107 yields

$$\|u - Pu\|_\varepsilon \leq C \left(\varepsilon N^{-3/2} + N^{-2} \ln^2 N \right). \qquad (3.145)$$

Theorem 3.116. *Let the* a priori *bounds (3.125) hold true with $m = 3$. Then after postprocessing by P, the numerical solution u^N generated by the SDFEM satisfies*

$$\|u - Pu^N\|_\varepsilon \leq C \left(\varepsilon N^{-3/2} + N^{-2} \ln^2 N \right).$$

Proof. The triangle inequality and (3.144)–(3.145) yield

$$\begin{aligned}
\|u - Pu^N\|_\varepsilon &= \|u - Pu + P(u^I) - Pu^N\|_\varepsilon \\
&\leq \|u - Pu\|_\varepsilon + \|P(u^I - u^N)\|_\varepsilon \\
&\leq C \left(\varepsilon N^{-3/2} + N^{-2} \ln^2 N + \|u^I - u^N\|_{1,\varepsilon} \right) \\
&\leq C(\varepsilon N^{-3/2} + N^{-2} \ln^2 N),
\end{aligned}$$

where we used the superclose property (3.137).   □

Theorem 3.116 exhibits a higher rate of convergence than (3.141) at only a minor additional computational cost since Pu^N can be computed very cheaply from u^N.

A related analysis appears in [RL01b], which is concerned with gradient recovery from the solution u^N computed using the Galerkin method $B_{GAL}(u^N, v^N) = (f, v^N)$ for all $v^N \in S^N$, the space of piecewise bilinears on a general layer-adapted tensor-product mesh with nodes (x_i, y_j). The construction is as follows: compute the gradient of u^N at the barycentre of each mesh rectangle K, then by means of bilinear interpolation between these barycentric values generate a recovered gradient at the (x_i, y_j); finally, again apply bilinear interpolation – now with the (x_i, y_j) as nodes – to extend the recovered gradient Ru^N to all of Ω. (Near the boundary the method is modified slightly so that values are defined by extrapolation.) In the case of a tensor-product Shishkin mesh where $k = 3$ in (3.122), if we assume that (3.125) is valid with $m = 3$, then by an argument like that of Theorem 3.116 one can improve the estimate $\varepsilon^{1/2}\|\nabla(u - u^N)\|_0 \leq CN^{-1} \ln N$, which follows from (3.128c) and (3.136), to

$$\varepsilon^{1/2}\|\nabla u - Ru^N\|_0 \leq CN^{-2} \ln^2 N.$$

Under the mild extra hypothesis that $\varepsilon \leq CN^{-1/2} \ln^2 N$ we get an identical bound on $\varepsilon^{1/2}\|\nabla(u - Pu^N)\|_0$ from Theorem 3.116, but $\nabla(Pu^N)$ is only piecewise continuous on Ω while $Ru^N \in C(\bar{\Omega})$.

To conclude this section, we temporarily leap forward to the reaction-diffusion problem (3.147). When a piecewise biquadratic postprocessing P (similar to that described above for convection-diffusion) is applied to the computed solution u^N of (3.150), one obtains [Li01] the improved convergence rate $\varepsilon^{1/2}|u - Pu^N|_1 \leq CN^{-2}$. A local postprocessing S can also be applied to the computed solution v^N of (3.151) to yield [LW00]

$$\varepsilon^{1/2}|u - Sv^N|_1 + \|u - Sv^N\|_0 \leq CN^{-2}.$$

Remark 3.117. Gradient recovery is an important technique in adaptive methods, as will be seen in Section 3.6.1. ♣

L_∞ Error Bounds

Now we move on to error bounds in L_∞. From (3.137) and (3.127) one can deduce bounds on the pointwise error $(u - u^N)(x, y)$ for the SDFEM solution u^N when (x, y) lies in each of the regions Ω_i, as will now be demonstrated.

Consider first Ω_1. Let (x_i, y_j) be any mesh node in Ω_1 with $j \geq 1$, as the case $j = 0$ is trivial. Then

$$|(u^I - u^N)(x_i, y_j)| = \left| \sum_{k=i+1}^{N} \int_{x=x_{k-1}}^{x_k} (u^I - u^N)_x(x, y_j) \, dx \right|$$

$$\leq CN \sum_{k=i+1}^{N} \int_{x=x_{k-1}}^{x_k} \int_{y=y_{j-1}}^{y_j} |(u^I - u^N)_x(x, y)| \, dy \, dx,$$

since $(u^I - u^N)_x$ is a linear function of only y on $[x_{k-1}, x_k] \times [y_{j-1}, y_j]$. Set $\Omega^{ij} = [x_i, 1] \times [y_{j-1}, y_j]$. Hence, by the Cauchy-Schwarz inequality one has

$$|(u^I - u^N)(x_i, y_j)| \leq CN \int_{\Omega_{ij}} |(u^I - u^N)_x|$$

$$\leq CN(\text{meas } \Omega^{ij})^{1/2} \|(u^I - u^N)_x\|_{0, \Omega^{ij}}$$

$$\leq CN(\varepsilon N^{-1} \ln N)^{1/2} \|\nabla(u^I - u^N)\|_0$$

$$\leq C(\varepsilon N^{-1} \ln^{1/2} N + N^{-3/2} \ln^{5/2} N)$$

where we used (3.137). Now (3.127) and a triangle inequality give

$$\|u - u^N\|_{0, \infty, \Omega_1} \leq C(\varepsilon N^{-1} \ln^{1/2} N + N^{-3/2} \ln^{5/2} N).$$

A similar argument yields the same bound for $\|u - u^N\|_{0, \infty, \Omega_2}$.

In particular these bounds are valid for all (x, y) on the boundary of Ω_{12}. Now consider the restriction of our SDFEM to Ω_{12}. Here the fineness of the mesh in both coordinate directions ensures that the associated difference operator is inverse-monotone. If we assume more regularity of the solution u, viz., that the bounds (3.125c)–(3.125e) are valid for $0 \leq i + j \leq 4$ and that $\varepsilon |\partial^3 S(x, y)/(\partial^i x \partial^j y)| \leq C$ for $i + j = 3$, then a barrier function argument on Ω_{12} (cf. [LS99]) shows that

$$\|u - u^N\|_{0, \infty, \Omega_{12}} \leq C(\varepsilon N^{-1} \ln^{1/2} N + N^{-3/2} \ln^{5/2} N).$$

Finally, consider the coarse mesh Ω_0. Apply a standard inverse estimate to the L_2 error component of (3.137), then invoke a triangle inequality and (3.127); we obtain

$$\|u - u^N\|_{0,\infty,\Omega_0} \leq C(\varepsilon N^{-1/2} + N^{-1} \ln^2 N).$$

Piecewise linears on a triangular Shishkin mesh are used in the SDFEM in [LS01c]. Take $k \geq 2$. For theoretical reasons $\mathcal{O}(N^{-3/2})$ artificial crosswind diffusion is added to the method on Ω_0, imitating [JSW87]. Pointwise error bounds are derived via weighted estimates for discrete Green's functions; these bounds are quite detailed and we give only some of the results here.

Assume that $\varepsilon \leq N^{-3/2}$ and u satisfies the assumptions (3.125) with $m = 2$. Then for any $(x,y) \in \Omega$,

$$|(u - u^N)(x,y)| \leq \begin{cases} CN^{-1/2} \ln^{3/2} N & \text{if } (x,y) \in \Omega_{12}, \\ CN^{-3/4} \ln^{(2+\delta)/2} N & \text{if } (x,y) \in \Omega \setminus \Omega_{12}, \end{cases}$$

where

$$\delta = \begin{cases} 0 & \text{if } (x,y) \in \Omega_0, \\ 1 & \text{if } (x,y) \in \Omega_2 \cup \Omega_1. \end{cases}$$

Furthermore, on subregions "away from" layers ([LS01c] gives a precise definition) one has

$$|(u - u^N)(x,y)| \leq CN^{-11/8} \ln^{1/2} N.$$

Remark 3.118. (Bilinears versus linears) Numerical pointwise convergence results on Shishkin meshes in [LS01b] (see also [TMS00]) show that both the bilinear and linear SDFEM are second-order accurate on Ω_0, but on $\Omega \setminus \Omega_0$ the linear SDFEM is much less accurate than the bilinear method, which achieves better than first-order accuracy. Should we be surprised that bilinears are superior to linears on Shishkin meshes? A clue is given by the fact that for linears on triangles the analogue of the Lin identities of Lemma 3.113 (see Lemma 3.35 or [BX03]) includes line integrals over the boundary ∂K and these can be controlled in the subsequent error analysis only when the union of each pair of neighbouring triangles forms an approximate parallelogram – a condition that is not satisfied on a Shishkin mesh where the coarse and fine meshes meet. For bilinears, the analysis on each mesh rectangle is independent of all other mesh rectangles so abrupt changes in mesh size are not an obstacle. Thus while the analysis leading to (3.135) can easily be replicated for piecewise linears, our proof of the stronger result (3.136), which draws on Lemma 3.113, seems to work only for bilinears.   ♣

Remark 3.119. Several numerical methods for a test problem of the form (3.111) are compared on the same Shishkin mesh in [LS01b]. The methods tested are central differencing, the simple upwind scheme of Section 2.1.1, the hybrid difference scheme of [LS99], defect correction (see, e.g., [AL90]), linear and bilinear Galerkin FEMs, and the linear and bilinear SDFEMs. Graphs of the computed solutions and errors and convergence rates are given with respect to the discrete $L_\infty(\Omega)$ norm. The authors conclude that, taking into account certain difficulties that arise in solving the discrete linear systems for some of the methods (see Remark 3.111), the methods that perform best for this problem are the defect correction method and the two SDFEMs.   ♣

Remark 3.120. (Other stabilization techniques) In [RZ03] an asymmetric interior penalty version (NIP) of the discontinuous Galerkin finite element method (see Section 3.4) is used to solve (3.111) with piecewise bilinears on the Shishkin mesh described at the start of this section. The value $k = 2$ is chosen in (3.122). The analysis of the dGFEM in Section 3.4 is inapplicable to long thin elements such as those appearing in $\Omega_1 \cup \Omega_2$, but in [RZ03] this analysis is extended to such elements; to this end a new choice of the discontinuity penalization parameter σ_e is made on part of the mesh. It is shown that the computed solution u^N satisfies

$$|||u - u^N|||_{dG} \leq C N^{-1} \ln^{3/2} N.$$

The norm $\|\cdot\|_{dG}$ is not directly comparable with $|||\cdot|||_{SD}$, but numerical results in [RZ03] indicate that when bilinears are used and the error is measured in the discrete L_∞ norm, both the Galerkin FEM and the SDFEM are more accurate than the dGFEM.

A related method is considered in [RZ07]: a Galerkin method with bilinears is used on the fine part of the Shishkin mesh and an NIP version of the dGFEM is used on the coarse mesh. The authors prove the supercloseness result $|||\pi u - u^N|||_{dG} \leq C(\varepsilon^{1/2} N^{-1} + N^{-3/2})$ for a certain interpolant πu of u.

Compared with the case of a computed continuous solution discussed on page 397, it is less clear whether one can postprocess a superclose piecewise discontinuous solution to generate a new finite element solution that achieves a higher order of convergence in a norm associated with the original finite element method. The key issue is how to specify the degrees of freedom for the postprocessed solution in a way that uniquely defines that solution and yields consistency and stability properties analogous to (3.144a) and (3.144b). In [FTZ08] the authors show how to do this for superclose discontinuous piecewise polynomial solutions generated by the dGFEM.

Local projection stabilization (see Section 3.3.1) for (3.111) on Shishkin meshes is examined in [Mat]. For arbitrary $r \geq 2$, the standard Q_r element is enriched by six additional functions, yielding an element that contains P_{r+1}. For (3.111) it is shown that

$$\|u - u^N\|_\varepsilon + |||\pi u - u^N|||_{LPS} \leq C(N^{-1} \ln N)^{r+1},$$

where πu is a certain interpolant of u and u^N is the computed solution.

In [FLRS08] the continuous interior penalty stabilization method (CIP) of Section 3.3.2 is applied to (3.111) on a Shishkin mesh, using bilinears on the fine mesh and linears elsewhere, and a proof is given of the superclose result

$$|||\pi u - u^N|||_{CIP} \leq C(\varepsilon^{1/2} N^{-1} + N^{-3/2})$$

where πu is a certain interpolant of u and u^N is the computed solution.

As is apparent to the reader by now, different discretization techniques often call upon different interpolants in their analyses.   ♣

Remark 3.121. (hp fem, curvilinear boundaries) The approximation, uniformly with respect to ε, of exponential and parabolic boundary layer functions by the p and hp versions of the finite element method using anisotropic meshes on domains with curvilinear boundaries is considered in [SSX98] and at much greater length in the monograph [Mel02].

The SDFEM is analysed in a hp fem context in [GMSS01], where Ω is a curvilinear Lipschitzian polygon. After decomposing the solution u into smooth and finitely many (exponential and parabolic) layer components, this leaves a small remainder – caused for example by corner singularities – that is ignored in the analysis. When the layers are resolved by suitable meshes, an exponential rate of convergence is proved for the computed solution u^N:

$$|||u - u^N|||_{SD} \leq C|\mathcal{T}|^{1/2} e^{-Cp},$$

where $|\mathcal{T}|$ denotes the number of mesh elements and p is the degree of piecewise polynomials used in the SDFEM trial space. If only exponential layers are present then the factor $|\mathcal{T}|^{1/2}$ can be discarded, but when parabolic layers are present then the mesh construction leads to $|\mathcal{T}| = \mathcal{O}(|\ln \varepsilon|^2)$.

For extensions of these ideas to problems posed in 3-dimensional domains, see [TS03] and its references. ♣

Parabolic Boundary Layers

Consider now the problem

$$-\varepsilon \Delta u + b_1(x, y)u_x + c(x, y)u = f \quad \text{on } \Omega := (0, 1) \times (0, 1), \qquad (3.146a)$$
$$u = 0 \quad \text{on } \partial\Omega. \qquad (3.146b)$$

This is obtained by setting $b_2 \equiv 0$ in (3.111a). As before, $b_1(x, y) \geq \beta_1 > 0$ and $c \geq 0$ on $\bar{\Omega}$. Assume that the data of the problem are smooth. The solution u usually has an exponential boundary layer at $x = 1$ and parabolic boundary layers at $y = 0$ and $y = 1$; see Section 1.4.

The quantity of published analysis on layer-adapted meshes for (3.146) is much less than for (3.111).

A typical Shishkin mesh for (3.146) is a tensor product of one-dimensional Shishkin meshes: in the x-direction the convection-diffusion mesh of Section I.2.4.2 is used, while in the y-direction, where a boundary layer appears at both ends of the interval, one applies the mesh used for the reaction-diffusion problem of Remark I.2.106. These one-dimensional meshes are fine at $x = 1$ and at $y = 0$ and $y = 1$ respectively. The corresponding mesh transition parameters are

$$\sigma_x = \min\{1/2, k_x \varepsilon \ln N\} \quad \text{and} \quad \sigma_y = \min\{1/4, k_y \sqrt{\varepsilon} \ln N\}.$$

Here N is the number of mesh intervals in each coordinate direction and k_x and k_y are user-chosen parameters. The final two-dimensional mesh is shown in the second diagram of Figure 2.2.

Provided that one has a decomposition of the solution u whose components satisfy certain reasonable bounds, it is shown in [Roo02] that when a Galerkin finite element method based on the bilinear form (3.130) with piecewise linears or bilinears is applied on this mesh (with $b_1 \equiv 1$ and $\sigma_x = \sigma_y = 2$) one obtains

$$\varepsilon^{1/2}|u - u^N|_1 \leq CN^{-1} \ln N,$$

where u^N is the computed solution. See Figure 3.14 for a typical computed solution. In [FL08] Franz and Linß prove that the solution also has a superclose property:

$$||u^I - u^N||_\varepsilon \leq CN^{-2} \ln^2 N,$$

where u^I is the bilinear nodal interpolant. Consequently a simple local postprocessing of u^N yields a piecewise quadratic solution Pu^N for which $||u - Pu^N||_\varepsilon \leq CN^{-2} \ln^2 N$; see [FL08].
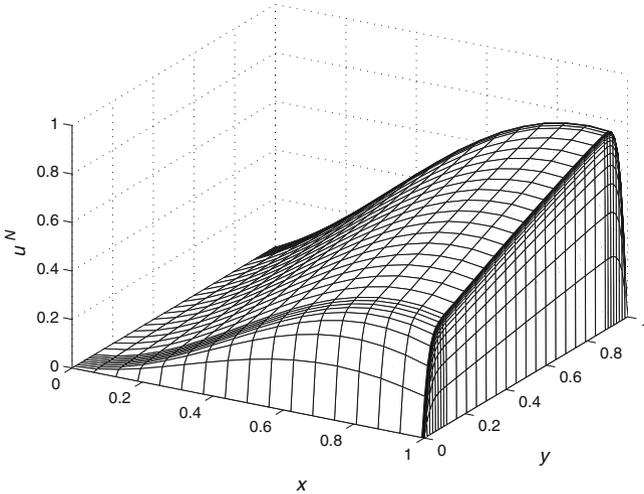
Fig. 3.14. Problem with one exponential and two parabolic layers; solution computed by Galerkin finite element method using bilinears on a Shishkin mesh

The SDFEM with piecewise bilinears on a Shishkin mesh is used to solve (3.146) in [FLR]; it is shown that one obtains the superclose bound

$$||u^I - u^N||_\varepsilon \leq CN^{-2} \ln^2 N$$

where u^I is the bilinear nodal interpolant and u^N is the computed solution. The analysis reveals that the correct choice for the SDFEM parameter on each mesh rectangle in the parabolic layer fine Shishkin mesh is $\delta_T \leq C\varepsilon^{-1/4}N^{-2}$, which is unexpected.

The local convergence results of [JSW87, Nii90] (see Section 3.2.1) for piecewise linears on a triangular mesh are of course applicable here and show for example that one can expect close to $\mathcal{O}(N^{-3/2})$ pointwise convergence away from all layers. On the other hand, Kopteva [Kop04] considers piecewise linears and bilinears and argues that inside characteristic boundary layers, when one makes the choice $\delta_T = CN^{-1}$ for the streamline-diffusion parameter – as many authors have advocated since the local mesh diameter of the Shishkin mesh in this region is $\mathcal{O}(N^{-1})$ – the SDFEM can at best give first-order pointwise convergence.

The application and analysis of Galerkin least squares finite element methods (see Section 3.2.2) on certain layer-adapted anisotropic meshes is examined in [AL96], for problems posed in 2 and 3 dimensions whose solutions exhibit exponential and parabolic boundary layers. The meshes used are uniform near the boundary of the domain then graded until they become equidistant and coarse far from the boundary, so in concept they lie between Shishkin and Bakhvalov meshes. Under various technical hypotheses, it is shown that

$$|||u - u^h|||_{GLS} \le Ch^{2k}|\ln \varepsilon|,$$

where u^h is the computed solution, h is the mesh diameter, and piecewise polynomials of degree at most k are used.

Using the NIP variant of the dGFEM with piecewise bilinears on a Shishkin mesh, in [ZR05] it is proved that the numerical solution u^N of (3.146) satisfies

$$|||u - u^N|||_{dG} \le CN^{-1}\ln^{3/2} N.$$

The theory of n-widths is used in [KS01a] to examine the approximability of the solution u of (3.146) (with $b_1 \equiv c \equiv 1$) in $L_2(\Omega)$. It is shown that, when $\varepsilon^2 n \le 1$, if we know only that $f \in L_2(\Omega)$ then $\mathcal{O}(\varepsilon^{-1/3}n^{-2/3})$ is the best rate of convergence to u in $L_2(\Omega)$ that can be attained in general by any numerical method that employs n degrees of freedom.

Reaction-diffusion Problems

The linear reaction-diffusion problem

$$-\varepsilon \Delta u + c(x,y)u = f \quad \text{on } \Omega := (0,1) \times (0,1), \tag{3.147a}$$
$$u = 0 \quad \text{on } \partial\Omega, \tag{3.147b}$$

where $c(x,y) \ge \gamma > 0$ on $\bar{\Omega}$, is considered by various authors. Its solution u typically has exponential boundary layers on all sides of Ω and corner layers at the corners of Ω; for a decomposition of u into smooth and layer components, together with bounds on its derivatives, see Remark 1.27 and its references.

For this problem a standard finite element analysis of the Galerkin method (with any trial space V^N and on any mesh) shows quickly that the computed solution u^N has the quasi-optimality property

$$\|u - u^N\|_\varepsilon \leq C \inf_{v^N \in V^N} \|u - v^N\|_\varepsilon. \tag{3.148}$$

Consequently the analysis reduces to combining a decomposition of u with some approximation theory.

When piecewise linears and bilinears are used on a Shishkin mesh, (3.148) and calculations like those of Lemma 3.107 yield

$$\|u - u^N\|_\varepsilon \leq C(\varepsilon^{1/4}N^{-1}\ln N + N^{-2}). \tag{3.149}$$

In [LMSZ08] the authors consider a sparse grid variant of this method, for which one obtains the same convergence result while reducing the number of degrees of freedom from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^{3/2})$. If instead of bilinears, piecewise polynomials of degree $k \geq 1$ are used (i.e., the space P_k on triangles or Q_k on rectangles), then it is shown in [Ape99, p.198] that on a Shishkin mesh one has

$$\|u - u^N\|_\varepsilon \leq C(\varepsilon^{1/4}N^{-k}\ln^{k+1}N + N^{-k-1}),$$

where u^N is the computed solution.

Remark 3.122. (Asymptotic mesh) In Remark 2.12 we discussed a piecewise uniform mesh where the transition points of the Shishkin mesh, which for (3.147) are $\mathcal{O}(\sqrt{\varepsilon}\ln N)$ distant from the boundary, are repositioned to $\mathcal{O}(\sqrt{\varepsilon}|\ln\varepsilon|)$ distant from the boundary. We shall refer to this mesh as the asymptotic mesh or *A-mesh*. It is satisfactory for reaction-diffusion problems, despite its deficiencies in the convection-diffusion case; in fact for reaction-diffusion the convergence analysis on this mesh is slightly simpler than on the Shishkin mesh and no $\ln N$ factor appears in the error estimates. (Such factors cannot be avoided in energy-norm estimates when Shishkin meshes are used – see [Xen03].) Consider the standard Galerkin finite element method defined by $B_{GAL}(u^N, v^N) = (f, v^N)$ for all $v^N \in S^N$, the space of piecewise bilinears on a tensor product of two one-dimensional A-meshes of this type with N mesh intervals in each coordinate direction, where we set $b = 0$ in the bilinear form (3.130). It is proved in [Li01] that the computed solution u^N satisfies

$$N^{-1}\varepsilon^{1/2}|u - u^N|_1 + \|u - u^N\|_0 \leq CN^{-2}. \tag{3.150}$$

If instead of bilinears, piecewise polynomials of degree $k \geq 1$ are used (i.e., the space P_k on triangles or Q_k on rectangles), then [Ape99, p.198] on the A-mesh one obtains

$$\|u - u^N\|_\varepsilon \leq C(\varepsilon^{1/4}N^{-k}|\ln\varepsilon|^{k+1/2} + N^{-k-1}),$$

where u^N is the computed solution. The solution v^N of a mixed finite element method using lowest-order Raviart-Thomas rectangular elements on a similar mesh is shown in [LW00] to yield

$$\varepsilon^{1/2}|u - v^N|_1 + \|u - v^N\|_0 \leq CN^{-1}. \tag{3.151}$$

In the methods associated with (3.150) and (3.151) one obtains first-order pointwise convergence away from the boundary layers. ♣

Remark 3.123. (Nonrectangular domains) All the two-dimensional domains to which we applied layer-adapted meshes to solve (3.147) have been rectangular. Xenophontos and Fulton [XF03] consider general domains with smooth boundary and use boundary-fitted coordinates to construct suitable Shishkin meshes. The standard Galerkin finite element method based on (3.130) is used. In a typical situation where $\mathcal{O}(N^2)$ mesh elements and piecewise polynomials of degree p are used, one has

$$\varepsilon^{1/2}|u - u^N|_1 + \|u - u^N\|_0 \leq CN^{-p}\ln^p N$$

for the computed solution u^N, under the hypothesis that $c(x, y)$ in (3.147a) is constant.

Melenk [Mel02] derives detailed regularity estimates for the solutions of reaction-diffusion problems on curvilinear polygonal domains. To solve such problems numerically he concentrates on hp finite element methods, for which he designs special meshes that employ a minimal number of degrees of freedom yet are able to deal with exponential layers and corner singularities; furthermore, in [Mel02, Section 2.6] he considers the h fem (as we have done) and discusses suitable boundary layer meshes for reaction-diffusion problems on, e.g., L-shaped domains, whose re-entrant corner requires a special mesh. In [Mel02, Theorem 2.6.15] a bound related to (3.149) is presented in a more general setting. ♣

Finally, we mention some results for reaction-diffusion problems on general meshes. In [KS99, Mel00] the theory of n-widths sheds light on the approximability of the solution u of (3.147) in various Sobolev norms: if $f \in H^s(\Omega)$, then any numerical method that employs n degrees of freedom can in general attain at best $\mathcal{O}(n^{-s/2}(1+\varepsilon n)^{-1})$ convergence in $L_2(\Omega)$. In [Ley08] (cf. [SW83]) the author considers (3.147a) with $c \equiv 1$ and Neumann boundary conditions where $\Omega \subset \mathbb{R}^m$ ($m \geq 2$) has a smooth boundary; he applies the Galerkin method with piecewise polynomials of any fixed degree and expresses the local pointwise error in the solution in terms of ε and the mesh diameter.

3.6 Adaptive Methods

3.6.1 Adaptive Finite Element Methods for Non-Singularly Perturbed Elliptic Problems: an Introduction

In Section I.2.5 an adaptive finite difference method for problems in one dimension was examined at length. Two adaptive techniques for one-dimensional parabolic problems were presented in Chapter II.5. In the context of adaptive techniques in several space dimensions, the theoretical basis underpinning the finite element method is much more secure than for any other class of methods. We shall therefore concentrate here on the finite element method and related techniques such as the finite volume and discontinuous Galerkin methods.

In adaptive finite element methods, the mesh is refined wherever an *a posteriori* error estimator indicates the presence of large local errors in the computed solution. In this way one hopes to place fine meshes in those regions affected by local singularities, shocks, or interior or boundary layers, and to achieve a balance between refined and unrefined regions so that satisfactory global accuracy is attained without the introduction of too many mesh points.

Given an *a posteriori* error estimator, each adaptive mesh-refinement algorithm has the following general structure:

1. Construct an initial coarse mesh $\mathcal{T}_0$ that is a sufficiently good approximation of the geometry of the problem. Put $k = 0$.
2. Solve the discrete problem on $\mathcal{T}_k$.
3. Compute an *a posteriori* error estimate for each element T in $\mathcal{T}_k$.
4. If the estimated global error is sufficiently small, then stop. Otherwise decide which elements have to be refined and hence construct a mesh $\mathcal{T}_{k+1}$. Replace k by $k + 1$ and return to Step 2.

Let $\mathcal{T}_h$ be the triangulation of the given domain. Write η_T for the estimator associated with each element $T \in \mathcal{T}_h$. Then η_T is called an *a posteriori* estimator if its evaluation depends on the computed numerical solution as well as on the given data of the problem. Set

$$\eta^2 = \sum_{T \in \mathcal{T}_h} \eta_T^2.$$

An estimator is *equivalent* to an error norm $|| \cdot ||_{err}$ if one has

$$d_l \eta \leq ||u - u_h||_{err} \leq d_u \eta \tag{3.152}$$

for some positive constants d_l and d_u. (The H^1 norm is often used for second-order elliptic problems, but other norms such as the L_2 norm also play an important rôle.) To obtain a numerical solution whose accuracy is less than a prescribed tolerance, it is sufficient in practice to use *upper error estimators* or *refinement indicators* that satisfy only $||u - u_h||_{err} \leq d_u \eta$. But there is then the risk of *over-refinement* of the mesh. To avoid this, the estimator used should satisfy local lower bounds of the form

$$d_l^* \eta_T \leq ||u - u_h||_{err,\omega(T)}, \tag{3.153}$$

where $\omega(T)$ is some small neighborhood of the element T. If the number of elements in $\omega(T)$ is bounded independently of T and h, then we can sum (3.153) over all T and obtain the left-hand inequality of (3.152).

The quality of an *a posteriori* error estimator is often measured by its *efficiency index*, i.e., the ratio of the estimated error to the true error. An error estimator is said to be *efficient* if the efficiency index and its inverse remain bounded for all meshsizes. The inequalities (3.152) guarantee efficiency. An error estimator is *asymptotically exact* if its efficiency index approaches unity as the meshsize converges to zero. In fact, asymptotical exactness is too strong a requirement: it holds true only for special uniform meshes and is certainly not true of the meshes that are used in practical computations – see [DMR91, DR92, DMR92].

We shall describe briefly four popular types of estimators:

- residual estimators
- estimators based on the solution of local problems
- estimators based on higher recovery of the gradient
- goal-oriented estimators (or the DWR method: dual weighted residuals).

For other estimators and more detailed investigations see [Ver96, AO00]; for the DWR method see in particular [BR03].

Let us study the model problem

$$-\triangle u = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \Gamma, \tag{3.154}$$

whose solution u lies in $V := H_0^1(\Omega)$. It is discretized using V_h, the space of piecewise linear conforming finite elements on a shape-regular triangulation $\mathcal{T}_h$. The discrete solution $u_h \in V_h \subset V$ satisfies

$$(\nabla(u - u_h), \nabla v) = (f, v) - (\nabla u_h, \nabla v) \quad \forall v \in V. \tag{3.155}$$

This *residual equation* is the starting point for the first two classes of estimators presented here.

From the Poincaré-Friedrichs inequality

$$||v||_0 \leq c_{PF}||\nabla v||_0 \quad \forall v \in V,$$

one has

$$\frac{1}{1 + c_{PF}^2}||v||_1 \leq \sup_{w \in V, ||w||_1=1} (\nabla v, \nabla w) \leq ||v||_1.$$

Combining this estimate with (3.155) yields

$$Q \leq ||u - u_h||_1 \leq (1 + c_{PF}^2)Q, \tag{3.156}$$

where

$$Q = \sup_{w \in V, \|w\|_1 = 1} \{(f, w) - (\nabla u_h, \nabla w)\}.$$

(Alternatively, start from

$$|u - u_h|_1 = \sup_{w \in V, |w|_1 = 1} \{(f, w) - (\nabla u_h, \nabla w)\},$$

and invoke Poincaré-Friedrichs later.) Now we try to find an upper bound for Q and hence for the right-hand side of (3.156). The projection property of the Galerkin method gives

$$(\nabla(u - u_h), \nabla v_h) = 0 \quad \forall v_h \in V_h.$$

Let $\Pi_h : V \to V_h$ be an operator that is specified later. Define

$$Q^*(w) := (f, w) - (\nabla u_h, \nabla w) = (f, w - \Pi_h w) - (\nabla u_h, \nabla(w - \Pi_h w))$$

for each $w \in V$. Integrate by parts on each triangle then sum the terms:

$$Q^*(w) = \sum_{T \in \mathcal{T}_h} (f + \triangle u_h, w - \Pi_h w)_T - \sum_{T \in \mathcal{T}_h} \int_{\partial T} (n_T \cdot \nabla u_h)(w - \Pi_h w).$$

Here n_T is the outward-pointing unit normal vector on ∂T. The term $f + \triangle u_h$ reduces to f for our piecewise linear elements, but for more complex elements this would not be the case. Next, write the ∂T terms as a sum over the edges $E \in \mathcal{E}_h$ of the triangulation and introduce the jump $[n_E \cdot \nabla u_h]_E$ of the normal derivative of u_h along each edge E. This leads to

$$Q^*(w) = \sum_{T \in \mathcal{T}_h} (f, w - \Pi_h w)_T - \sum_{E \in \mathcal{E}_h} \int_E [n_E \cdot \nabla u_h]_E (w - \Pi_h w). \qquad (3.157)$$

Choose Π_h to be the interpolation operator of Clément [Clé75]. (Recall that w lies only in H^1, so the simpler pointwise interpolant may not be defined.) The Clément operator is defined similarly to the Scott-Zhang operator that is discussed in Remark IV.3.1; see [EG04] for a general introduction to the interpolation of non-smooth functions. For the Clément interpolant, the following estimates are valid:

$$\|w - \Pi_h w\|_{0,T} \le C_1 h_T \|w\|_{1, \tilde{\omega}_T}, \qquad (3.158a)$$

$$\|w - \Pi_h w\|_{0,E} \le C_2 h_E^{1/2} \|w\|_{1, \tilde{\omega}_E}. \qquad (3.158b)$$

Here h_T is the diameter of the element T and h_E is the length of the edge E, while $\tilde{\omega}_T$ is the collection of all triangles sharing a vertex with the given triangle T and $\tilde{\omega}_E$ is the collection of all triangles sharing a vertex with E. Combining (3.158) with (3.157), we get

$$Q^*(w) \leq \sum_{T \in \mathcal{T}_h} C_1 h_T \|f\|_{0,T} \|w\|_{1,\tilde{\omega}_T} + \sum_{E \in \mathcal{E}_h} C_2 h_E^{1/2} \|[n_E \cdot \nabla u_h]_E\|_{0,E} \|w\|_{1,\tilde{\omega}_E}$$

$$\leq C_3 \|w\|_1 \left\{ \sum_{T \in \mathcal{T}_h} h_T^2 \|f\|_{0,T}^2 + \sum_{E \in \mathcal{E}_h} h_E \|[n_E \cdot \nabla u_h]_E\|_{0,E}^2 \right\}^{1/2} . \quad (3.159)$$

Here C_3 depends on $\max\{C_\tau C_1, C_2\}$, where C_τ is the maximum number of triangles that meet at any vertex. Set

$$\eta_{R,T}^2 = h_T^2 \|r_T\|_{0,T}^2 + \frac{1}{2} \sum_{E(T)} h_E \|r_E\|_{0,E}^2 \qquad (3.160a)$$

and define the *element and edge residuals*

$$r_T := f + \triangle u_h|_T, \qquad r_E := [n_E \cdot \nabla u_h]_E . \qquad (3.160b)$$

The estimates (3.159) and (3.156) prove that $\eta_{R,T}$ *is an upper error estimator*. This residual estimator was first proposed and analysed for problems in one space dimension in the classic paper [BR78]. It is also possible [Ver96] to derive a *local lower bound* related to this estimator. The lower bound relies on the fact that the residuals r_T and r_E are discrete; if f is not piecewise constant, then an additional data error indicator comes into the game.

We now move on to *estimators based on the solution of local problems*. These come from approximate solutions of the residual equation (3.155). Let us describe three important representatives of this class without going into excessive detail.

Assume that V_T is a low-dimensional space defined on a subdomain ω_T that is a small neighbourhood of the element T. Let $v_T \in V_T$ be the solution of the local residual problem

$$(\nabla v_T, \nabla w)_{\omega_T} = (f, w)_{\omega_T} - (\nabla u_h, \nabla w)_{\omega_T} \quad \forall w \in V_T$$

and set

$$\eta_{LP,T} = \|\nabla v_T\|_{0,\omega_T}.$$

It can be shown that for a good choice of the pair (ω_T, V_T), the quantity $\eta_{LP,T}$ will be a good approximation of $\|u - u_h\|_{1,T}$. Different choices of the pair (ω_T, V_T) lead to different estimators. To give some concrete examples, consider the *triangle-bubble function*

$$b_T = \begin{cases} 27\lambda_{T,1}\lambda_{T,2}\lambda_{T,3} & \text{on } T, \\ 0 & \text{on } \Omega \setminus T, \end{cases}$$

where $\lambda_{T,1}, \lambda_{T,2}$ and $\lambda_{T,3}$ are the barycentric coordinates of T. Given any edge E, let the vertices of the triangles T_1 and T_2 that contain E be enumerated in such a way that the vertices of E are numbered first. Define the *edge-bubble function* b_E by

$$b_E = \begin{cases} 4\lambda_{T_i,1}\lambda_{T_i,2} & \text{on } T_i \text{ for } i = 1,2, \\ 0 & \text{on } \Omega \setminus (T_1 \cup T_2). \end{cases}$$

We can now describe three well-known estimators.

1. *Babuška-Rheinboldt estimator:*
   ω_T is the union of those triangles T' that share a given vertex x, and $V_T = \mathrm{span}\{b_{T'}, b_E : T' \subset \omega_T, E \cap x \neq \emptyset\} \subset H_0^1(\omega_T)$;
2. *Verfürth estimator:*
   ω_T comprises the 4 triangles T' that have a common edge with the given triangle T, and $V_T = \mathrm{span}\{b_{T'}, b_E : T' \subset \omega_T, E \subset \partial T\} \subset H_0^1(\omega_T)$;
3. *modified Bank-Weiser estimator:*
   for this estimator Neumann boundary conditions, instead of Dirichlet boundary conditions, are imposed on the auxiliary problem. Set $\omega_T = T$ and $V_T = \mathrm{span}\{b_T, b_E : E \subset \partial T\}$. Let v_T be the unique solution of

$$(\nabla v_T, \nabla w)_T = (f, w)_T - \frac{1}{2} \sum_{E \in T} \int_E [n_E \cdot \nabla u_h]_E w \quad \forall w \in V_T.$$

This local problem is a discrete analogue of the Neumann problem

$$-\Delta \varphi = f \quad \text{in } T, \qquad \frac{\partial \varphi}{\partial n} = -\frac{1}{2}[n_E \cdot \nabla u_h]_E \quad \text{on } E.$$

The related equilibrated residual method (ERM) is discussed in detail in [AO00].

Because the dimensions of the local auxiliary problems for the last two estimators are small (7 and 4, respectively), they are often used in practice. See also [BW90, Ver89] for extensions of these estimators to the Stokes and Navier-Stokes equations; in the case of the model problem (3.154), it is easy to see that the auxiliary problems have unique solutions, but for more complex problems this property is far less obvious.

   As a third category we sketch the basic idea for an *estimator based on higher-order recovery of the gradient.* Suppose that one has an easily-computed approximation Gu_h of ∇u such that

$$\|\nabla u - Gu_h\|_0 \leq \beta \|\nabla(u - u_h)\|_0$$

for some constant $\beta \in [0, 1)$. Then

$$\frac{1}{1+\beta}\|Gu_h - \nabla u_h\|_0 \leq \|\nabla(u - u_h)\|_0 \leq \frac{1}{1-\beta}\|Gu_h - \nabla u_h\|_0,$$

so $\|Gu_h - \nabla u_h\|_0$ can be used as an error estimator for the energy norm. Now for some finite elements (e.g., linear elements), a superconvergent approximation Gu_h of the gradient is known [KN87], but details are not given here. For each fixed node x of the triangulation, define the weighted average (recall that ∇u_h is piecewise constant)

$$Gu_h(x) := \sum_{T \in \omega_x} \frac{|T|}{|\omega_x|} \nabla u_h|_T,$$

where ω_x is the collection of all triangles containing x. Interpolation to these nodal values defines a piecewise linear approximation $Gu_h \in V_h$. Now set

$$\eta_{Z,T} = \|Gu_h - \nabla u_h\|_{0,T}.$$

This is the *Zienkiewicz-Zhu estimator*. Carstensen [CB02a, CB02b] has pointed out that the success of averaging here can also be explained without using superconvergence. Furthermore, it is important to note that Zhang [Zha04, ZN05] introduces new averaging techniques that preserve polynomials.

So far we have sketched estimators only for the model problem (3.154). These estimators can be extended to more general elliptic problems provided that the diffusion terms dominate the convection terms. All the estimators discussed are interrelated [Ver96] and have been tested numerically on a wide class of elliptic problems [BSU94].

In recent years there has been a flurry of activity around *a posteriori* error estimates for the discontinuous Galerkin method; see [KP07] and its bibliography.

Finally, we outline the *DWR method for goal-oriented error estimation*. This relies on duality arguments, which appeared in residual-based *a posteriori* error estimation only recently. The relevance of duality has been highlighted in the review articles by Eriksson et al. [EEHJ95] and Becker and Rannacher [BR01]. See also [BR03, GS02].

Suppose that we wish to control the linear error functional $J(u - u_h)$. Bangerth and Rannacher [BR03] discuss typical practical examples of error functionals such as the mean normal flux and the drag coefficient in computational fluid mechanics. It is also possible to control global norms by setting

$$J(\varphi) = (\nabla \varphi, \nabla e) \quad \text{or} \quad J(\varphi) = (\varphi, e).$$

Given a variational equation with bilinear form $a(\cdot, \cdot)$, define the adjoint auxiliary problem

$$a(v, w) = J(v) \quad \text{for all } v \in V. \tag{3.161}$$

Appealing to Galerkin orthogonality, one obtains the error representation

$$J(u - u_h) = a(u - u_h, w) = a(u - u_h, w - w_h).$$

This formula is the basis of the DWR method. Now standard arguments in deriving residual estimators (like those sketched above) yield

$$|J(u - u_h)| \le \sum_T \|r_T\|_{0,T} \|w - w_h\|_{0,T} + \sum_E \|r_E\|_{0,E} \|w - w_h\|_{0,E}.$$

The determination of the weights multiplying the element and edge residuals here requires the solution of the adjoint problem (3.161). The advantage of

using these weights is that, analogously to a Green's function, they quantify the influences of the local residuals on the error in the target quantity J.

The DWR method can also be applied in conjunction with the discontinuous Galerkin finite element method [KR02, RW03].

For many years no proof of convergence of an adaptive algorithm was known. The first rigorous result of this type is due to Dörfler [Doe96], who used a residual error estimator and a special refinement criterion . His analysis assumes that the initial mesh is already sufficiently fine to control data oscillations, which are defined by

$$
osc(f, \mathcal{T}_h) := \left\{ \sum_{T \in \mathcal{T}_h} \| h(f - f_T) \|_{0,T}^2 \right\}^{1/2},
$$

where f_T is the mean value of f on each element T.

Later it became clear that error reduction requires conditions on the refinement, and an extended refinement criterion also takes data oscillations into account [Noc95]. Examples in [Noc95] show how significant for the convergence behaviour it is to generate new interior mesh points during the refinement process.

Binev, Dahmen and DeVore [BDdV04] prove optimal convergence rates for an adaptive algorithm with optimal complexity. In their algorithm both mesh-coarsening and mesh-refinement steps are taken into account. Stevenson [Ste07] simplifies the algorithm by combining [BDdV04] and [Noc95] in such a way that – at least for our linear model problem – coarsening steps are unnecessary.

Recently, the convergence of an adaptive discontinuous Galerkin approximation has been proved [KP07]. The first convergence results for an adaptive scheme based on the DWR method are in [DKV06].

Remark 3.124. When error estimators or refinement indicators are applied to singularly perturbed problems, they often detect layers and force mesh refinement in theses regions. But spurious global oscillations may appear if non-upwinded finite element methods are used on too coarse a mesh, and this incorrect solution will lead the estimator to suggest *global* mesh refinement.

John [Joh00] gives a detailed numerical study of the behaviour of several estimators when applied to problems with layers, using the SDFEM as a stable basic discretization. The meshes generated often turned out to be qualitatively very different from each other. *None* of the estimators examined (the gradient of the numerical solution, the Zienkiewicz-Zhu estimator, standard residual estimators for the H^1 norm and L_2 norms, estimators based on the solution of local Neumann problems using Galerkin or SDFEM, and the "robust" estimator of Verfürth [Ver98a] that we shall meet in the next section) worked satisfactorily in all tests although the test problems were not particularly difficult. It seems that when different types of layers appear in the same problem, this presents difficulties for adaptive methods.   ♣

Remark 3.125. From theoretical considerations it is clear that any good mesh for boundary or interior layers must be anisotropic. Thus an adaptive procedure designed for problems with layers should include an anisotropic refinement strategy. While several anisotropic mesh adaptation strategies are extant (see [DF04] and its bibliography), all are more or less heuristic. We do not know of any strategy for convection-diffusion problems in two dimensions where it is proved that, starting from some standard mesh, the refinement strategy is guaranteed to lead to a mesh that allows robust error estimates.

Of course, such a result would requires an error estimator suitable for an anisotropic mesh. Here some progress has been achieved in recent years – see [Kuh99, Kuh05, Pic03] – but many open problems remain.

Micheletti, Perotto and others [DMP07, FMP04] combine SDFEM, the DWR method and anisotropic interpolation error estimates to get an *a posteriori* error estimate for some target functional. They then use this information to implement a metric-based algorithm for mesh generation [CLGD06] that creates an "optimal" mesh. The numerical results obtained are interesting but the second step of the approach has a heuristic flavour.   ♣

In the singularly perturbed case the analysis of indicators is difficult because the constants occurring in the estimates usually depend on the small diffusion parameter. This is linked to the question: which norm should be used to measure the error? So far we have focussed on the H^1 norm, which is natural for second-order elliptic problems that are not singularly perturbed. But for singularly perturbed problems a weaker norm would be appropriate and we shall present some first proposals.

The next two subsections contain descriptions of several approaches specially designed for singularly perturbed problems. First, following in the footsteps of [Ang95a, San08, Ver05], we present robustness results for certain residual-type estimators in various norms. Second, we describe an estimator from [EJ93b] for the streamline-diffusion method of Section 3.2.1 (see also Section II.5 for the corresponding time-dependent version). The technique is a variant of the DWR method for which more recent results are also available, especially for pure transport problems [Ran98, HRS00].

We do not discuss recent results on upwinded mixed finite element schemes. See [KP08] for residual-type *a posteriori* estimates for our standard convection-diffusion problem and [Voh07] for a discussion of more general convection-diffusion problems that have an anisotropic diffusion-dispersion tensor.

3.6.2 Robust and Semi-Robust Residual Type Error Estimators

If the constants occurring in the estimates (3.152) are independent of ε we say the estimator is *robust*. The first robustness result for a residual-type error estimator applied to a convection-diffusion problem is due to Angermann [Ang95a]. We start our excursion into robust estimators with his approach. Consider again the singularly perturbed boundary value problem

$$-\varepsilon\triangle u + b\cdot\nabla u + cu = f \quad \text{in} \quad \Omega, \qquad (3.162a)$$
$$u = 0 \quad \text{on} \quad \Gamma, \qquad (3.162b)$$

where Ω is a two-dimensional polygonal bounded domain with boundary Γ. Assume that the coefficients b and c and the right-hand side f are sufficiently smooth, and that $c(x) \geq 0$.

Angermann proves robustness of an estimator with respect to the "graph norm"

$$||v||_{gr} := \sup_{w\in H_0^1(\Omega)} \frac{a(v,w)}{||w||_1},$$

where $a(\cdot,\cdot)$ is the bilinear form from the standard weak formulation of (3.162).

Remark 3.126. Under the hypothesis that $c - \frac{1}{2}\nabla\cdot b \geq \omega > 0$, there exist two constants C_1 and C_2 that are independent of ε, such that

$$a(w,w) \geq C_1\|w\|_\varepsilon^2 \quad \text{and} \quad |a(v,w)| \leq C_2\varepsilon^{-1/2}\|v\|_\varepsilon\|w\|_\varepsilon \quad \forall v,w \in H_0^1(\Omega),$$

where, as usual, $\|\cdot\|_\varepsilon$ is the ε-weighted H^1 norm. These inequalities imply that

$$\varepsilon^{1/2}C_1\|w\|_\varepsilon \leq \frac{C_1\|w\|_\varepsilon^2}{\|w\|_1} \leq \frac{a(w,w)}{\|w\|_1} \leq ||w||_{gr} \leq C_2\|w\|_\varepsilon.$$

Thus $||\cdot||_{gr}$ seems weaker than the ε-weighted H^1 norm on $H_0^1(\Omega)$. ♣

Angermann develops his theory for the inverse-monotonicity-preserving finite volume technique of Section 3.1, when applied to (3.162). But as we shall see, in principle the approach also works for a standard Galerkin discretization. The inverse-monotonicity-preserving finite volume technique generates the system (3.17) of discrete equations

$$\sum_{j\in\Lambda_i} \frac{\varepsilon m_{ij}}{d_{ij}} B\left(\frac{N_{ij}d_{ij}}{\varepsilon}\right)[u(P_i) - u(P_j)] + C_i m_i u(P_i) = f_i m_i$$

for $i = 1,\ldots,N$. The *a posteriori* error estimator for this method is based on a reformulation of the discrete problem as a Galerkin finite element method with a perturbed bilinear form and right-hand side. Thus we restate (3.17) above in the following way:

Find $u_h \in V_h$ (the space of piecewise linear finite elements on a weakly-acute triangulation of Ω) such that

$$a_h(u_h, v_h) = f_h(v_h) \quad \forall v_h \in V_h$$

where

$$a_h(u_h, v_h) := \varepsilon(\nabla u_h, \nabla v_h) + b_h(u_h, v_h),$$

$$b_h(u_h, v_h) := \sum_{i \in \Lambda} v_{hi} \left\{ \sum_{j \in \Lambda_i} (1 - r_{ij}) N_{ij} (u_{hj} - u_{hi}) m_{ij} + C_i u_{hi} m_i \right\},$$

$$f_h(v_h) := \sum_{i \in \Lambda} f_i v_{hi} m_i,$$

$$r_{ij} := 1 - z_{ij}^{-1} + \frac{1}{\exp(z_{ij}) - 1}, \qquad z_{ij} := \frac{N_{ij} d_{ij}}{\varepsilon},$$

where Λ is the set of all interior nodes and Λ_i was defined during the derivation of (3.17). For simplicity, we use the notation

$$v_{hi} = v_h(P_i), \qquad f_i = f(P_i).$$

To describe the error estimator, some additional notation is needed. In the circumcentric case of Figure 3.3, one can choose a triangle T_{ij} that has Γ_{ij} as an edge with P_i as the opposite vertex – this triangle is the union of two triangles $T_{ij}^{(1)}$ and $T_{ij}^{(2)}$ that have in common both the vertex P_i and exactly one half of the straight-line segment $P_i P_j$. Set $\Gamma_{ij}^k := T_{ij}^{(k)} \cap \Gamma_{ij}$ and $m_{ij}^{(k)} := \mathrm{meas}(\Gamma_{ij}^{(k)})$ for $k = 1, 2$. Let ψ_{hi} be the piecewise linear continuous function on the original triangulation that satisfies $\psi_{hi}(P_j) = \delta_{ij}$.

Angermann's estimator [Ang95a] (see [Ang91a, Ang92] for earlier slightly different versions) has the following structure:

$$\eta_l = \left\{ \sum_{i \in \Lambda} \eta_{li}^2 \right\}^{1/2} \qquad \text{for } l = 0, 1, 2, 3,$$

where

$$\eta_{0i} := \sup_{v \in H_0^1(\Omega)} \frac{a(u - u_h, \psi_{hi} v)}{\|\psi_{hi} v\|_1},$$

$$\eta_{1i} := \left\{ \sum_{j \in \Lambda_i} \sum_{k=1}^{2} \left[\frac{d_{ij}^2}{4} + (m_{ij}^{(k)})^2 \right] \int_{T_{ij}^{(k)}} [f - b \cdot \nabla u_h - c u_h]^2 dx \right\}^{1/2},$$

$$\eta_{2i} := \frac{1}{\sqrt{m_i}} \left| \int_{D_i} [f - f_i + (\nabla \cdot b - c) u_h + C_i u_{hi}] dx - \sum_{j \in \Lambda_i} u_{hi} N_{ij} m_{ij} \right|,$$

$$\eta_{3i} := \left\{ \sum_{j \in \Lambda_i} \sum_{k=1}^{2} \frac{d_{ij}}{m_{ij}^{(k)}} \left[\int_{\Gamma_{ij}^{(k)}} [(r_{ij} u_{hi} + (1 - r_{ij}) u_{hj}) N_{ij} \right. \right.$$

$$\left. \left. - (n_{ij} \cdot b) u_h] ds \right]^2 \right\}^{1/2}.$$

Roughly speaking, the estimator η_0 corresponds to the classical Section 3.6.1 approach for residual estimators or estimators based on the solution of local problems. On the other hand, the indicators η_{1i}, η_{2i} and η_{3i} correspond to the special kind of discretization used: the estimator η_1 reflects the lumping of the test functions, η_2 gives information about the approximation of the reaction term $(c - \nabla \cdot b)u$ and the right-hand side, and η_3 deals with the upwinding in the discretization of the convection term $\nabla \cdot (bu)$.

Using the graph norm, we have the following robustness result:

Theorem 3.127. *[Ang95a] Let $c - \frac{1}{2}\nabla \cdot b \geq \omega > 0$. Assume that all interior angles of all triangles of the triangulation are bounded from below by some positive constant and that the triangulation is weakly acute. Then there exist five constants $C^{(-1)}, C^{(0)}, ..., C^{(3)}$, which are independent of ε and the mesh, such that*

$$C^{(-1)}\eta_0 \leq ||u - u_h||_{gr} \leq \sum_{l=0}^{3} C^{(l)}\eta_l.$$

Remark 3.128. The definition

$$\eta_{0i} := \sup_{v \in H_0^1(\Omega)} \frac{a(u - u_h, \psi_{hi}v)}{||\psi_{hi}v||_\varepsilon}$$

and the use of the $||\cdot||_\varepsilon$ norm lead to the *unbalanced* estimate

$$\varepsilon^{1/2}C^{[-1]}\eta_0 \leq ||u - u_h||_\varepsilon \leq \varepsilon^{-1/2}\sum_{l=0}^{3} C^{[l])}\eta_l \, ;$$

see [Ang91a]. ♣

Unlike the indicators $\eta_{1i}, \eta_{2i}, \eta_{3i}$, which can either be computed exactly or approximated by quadrature rules, the indicator η_{0i} cannot be computed directly and has to be approximated. Nevertheless, we have already seen that such indicators can often be implemented through the solution of local boundary value problems; but in the present context it is very important that any such implementation uses *auxiliary boundary value problems that are not singularly perturbed*.

Let ω_{P_i} be, as in the Babuška-Rheinboldt estimator, the collection of all triangles with a given vertex P_i. Consider the following local problem: Find $e_{P_i} \in H_0^1(\omega_{P_i})$ such that

$$(\nabla e_{P_i}, \nabla w) + (e_{P_i}, w) = (f, w) - a(u_h, w) \qquad \forall w \in H_0^1(\omega_{P_i}). \qquad (3.163)$$

Lemma 3.129. *[Ang95a] One has*

$$\eta_{0i} = ||e_{P_i}||_{1,\omega_{P_i}}. \qquad (3.164)$$

Proof. First rewrite $\|e_{P_i}\|_{1,\omega_{P_i}}$ as follows:

$$
\|e_{P_i}\|_{1,\omega_{P_i}} = \sup_{w \in H_0^1(\omega_{P_i})} \frac{(\nabla e_{P_i}, \nabla w) + (e_{P_i}, w)}{\|w\|_{1,\omega_{P_i}}}
$$

$$
= \sup_{w \in H_0^1(\omega_{P_i})} \frac{(f, w) - a(u_h, w)}{\|w\|_{1,\omega_{P_i}}} = \sup_{w \in H_0^1(\omega_{P_i})} \frac{a(u - u_h, w)}{\|w\|_{1,\omega_{P_i}}}.
$$

The definition of η_{0i} implies that

$$
\eta_{0i} \leq \sup_{w \in H_0^1(\omega_{P_i})} \frac{a(u - u_h, w)}{\|w\|_{1,\omega_{P_i}}}. \tag{3.165}
$$

We now establish the opposite inequality. Let $w \in H_0^1(\omega_{P_i})$ be a function for which the supremum in (3.165) is attained (if no such function exists, choose a w that almost attains the supremum and modify the argument slightly). Then there exists a sequence $\{w^{(k)}\} \in C_0^\infty(\omega_{P_i})$ with $w^{(k)} \to w$ in $H^1(\omega_{P_i})$ as $k \to \infty$. The functions $v^{(k)} = w^{(k)}/\psi_{ih}$, extended by zero to the whole domain Ω, clearly lie in $H_0^1(\Omega)$. The continuity of the bilinear form $a(\cdot, \cdot)$ yields

$$
\eta_{0i} \geq \frac{a(u - u_h, \psi_{ih} v^{(k)})}{\|\psi_{ih} v^{(k)}\|_1} = \frac{a(u - u_h, w^{(k)})}{\|w^{(k)}\|_1} \to \frac{a(u - u_h, w)}{\|w\|_1}
$$

as $k \to \infty$. The result follows. $\square$

Note that the local problem (3.163) is not singularly perturbed, so one can apply standard methods to approximate η_{0i}. Of course one then loses the equality of (3.164). On the other hand, the symmetry of the local problem allows us to formulate both primal and complementary variational principles for (3.163) and to apply corresponding numerical methods, which leads to computable lower and upper bounds for η_{0i}; see [Ang95a] and the general discussion in [AC92] on the use of complementary variational principles for *a posteriori* error estimation.

Remark 3.130. Without referring to Angermann's work, Araya et al. [APS05, ABR07] use a related idea that introduces auxiliary problems similar to (3.163). In both these papers a variant of the SDFEM is used for stabilization. In [APS05] the residual equation is solved in a hierarchical way and in [ABR07] the estimator is based on the solution of local problems.   ♣

Can one combine the ε-weighted H^1 norm with a standard residual estimator of the type

$$
\eta_{R,T}^2 := w_T^2 \|r_T\|_{0,T}^2 + \frac{1}{2} \sum_{E(T)} w_E \|r_E\|_{0,E}^2
$$

for *a posteriori* error estimation? First, a careful determination of the dependence of the residual weights w_T and w_E on ε is necessary. It emerges [Ver98a, Ver05] that

$$\eta_T^2 := \alpha_T^2 \|r_T\|_{0,T}^2 + \sum_{E(T)} \varepsilon^{-1/2} \alpha_E \|r_E\|_{0,E}^2 \qquad (3.166a)$$

with the usual edge residual r_E, the element residual

$$r_T := (f + \varepsilon \triangle u_h - b \cdot \nabla u_h - c u_h)\big|_T \qquad (3.166b)$$

and for $S = T, E$ the weights

$$\alpha_S = \min\{h_S \varepsilon^{-1/2}, \omega^{-1/2}\}. \qquad (3.166c)$$

Here h_T and h_E are the element and edge diameters, while $\omega \geq 0$ is a constant with $c - (1/2)\operatorname{div} b \geq \omega$ and $\|c\|_\infty \leq c^* \omega$. For simplicity let us assume that f, b, c are piecewise linear (like our finite element space) as otherwise additional data error terms are present.

For the Galerkin or SDFEM discretizations with linear elements on a shape-regular mesh, under the hypothesis that $\omega = 1$, Verfürth [Ver98a] proved in the ε-weighted H^1 norm the estimates

$$\|u - u_h\|_\varepsilon \leq \left\{ \sum_T \eta_T^2 \right\}^{1/2}, \quad \eta_T \leq (C_1 + C_2 \varepsilon^{-1/2} \alpha_T)) \|u - u_h\|_{\varepsilon, \omega_T}.$$

The estimate is not robust with respect to ε, but owing to the relatively weak dependence on ε we say the estimator is *semi-robust*.

A similar technique was used in [IW99] to study the GLSFEM, resulting in weights slightly different from (3.166c), but the result was not robust with respect to Angermann's graph norm. In [AEB07] the authors construct a semi-robust estimator for the non-conforming Crouzeix-Raviart element, using stabilization by edge penalization (the CIP method).

In [San01] Sangalli proves the robustness of a certain *a posteriori* error estimator for the residual-free bubble method applied to convection-diffusion problems. The analysis uses the unusual norm

$$\|w\|_{San} := \|w\|_\varepsilon + \|b \cdot \nabla w\|_*, \quad \text{where} \quad \|\varphi\|_* = \sup \frac{\langle \varphi, v \rangle}{\|v\|_\varepsilon}. \qquad (3.167)$$

Although Sangalli's approach is devoted to residual-free bubbles, the same analysis works for the Galerkin method and the SDFEM. For the convection-diffusion problem, the residual error estimator (3.166) is robust with respect to the norm (3.167); see [Ver05].

Angermann's graph norm and the norm $\|\cdot\|_{San}$ above are defined only implicitly by an infinite-dimensional variational problem and cannot be computed exactly in practice. Recently, Sangalli [San08] pointed out that the norm

(3.167) seems to be inappropriate in the convection-dominated regime; recall the discussion of Section I.2.2.3. He proposes an improved estimator that is robust with respect to his natural norm [San05] for the advection-diffusion operator in the one-dimensional case.

Sangalli's result bears some resemblance to the results in [BC04] for wavelet methods, which rely on a non-standard variational formulation of the given problem and use an anisotropic wavelet decomposition of the residual. The associated norm, which is stronger than the standard energy norm, provides robust control over the streamline derivative of the solution. It is proved that the proposed lower error estimator is both accurate and robust. The upper estimator deviates from the true error by a factor at most $\mathcal{O}(\varepsilon^{-1/4})$, so once again semi-robustness is achieved.

Remark 3.131. (Robustness for reaction-diffusion problems) The theoretical situation is clearer for reaction-diffusion problems. The residual-based estimator (3.166) and the related estimator based on the solution of auxiliary local problems are both robust with respect to the associated energy norm [Ver98b]. A modification of the equilibrated residual method of Ainsworth and Babuska is also robust for reaction-diffusion problems [AB99].

Stevenson proved in [Ste05] the uniform convergence of a special adaptive method for the reaction-diffusion equation in the energy norm but it is unclear that the energy norm is a suitable norm for these problems because for small ε it is unable to distinguish between the typical layer function of reaction-diffusion problems and zero. It would be desirable to get robust *a posteriori* error estimates in a stronger norm.

The only result in this direction is the *a posteriori* error estimate of Kopteva [Kop08] in the L_∞ norm. For the standard finite difference method on an arbitrary rectangular mesh, she proves (with techniques closely related to finite element analyses; cf. [Noc95]) that

$$\|u_h - u\|_\infty \leq C \left(\max_i \{h_i^2 M_{1,ij}\} + \max_j \{k_j^2 M_{2,ij}\} \right).$$

Here a typical mesh rectangle has size $h_i \times k_j$ and u_h is the bilinear interpolant to the discrete approximations u_{ij} at the mesh points; furthermore,

$$M_{1,ij} \approx |D_x^2 u_{ij}| \ln(2 + \varepsilon/\kappa) + 1, \quad M_{2,ij} \approx |D_y^2 u_{ij}| \ln(2 + \varepsilon/\kappa) + 1$$

with $\kappa = \min\{\min_i h_i, \ \min_j k_j\}$. Thus we recognize the jump of the normal derivatives on edges in the *a posteriori* error estimate. The proof uses the representations

$$L(u_h - u) = f_0 - (f_1)_x - (f_2)_y$$

and

$$u_h - u = (G, f_0) + (G_x, f_1) + (G_y, f_2),$$

with a precise analysis of the dependence on ε of various norms of the Green's function G of the continuous problem. ♣

3.6.3 A Variant of the DWR Method for Streamline Diffusion

In this section we examine an adaptive streamline diffusion finite element method (SDFEM) from [EJ93b] for the convection-diffusion problem

$$-\varepsilon \triangle u + u_x = f \quad \text{in } \Omega, \tag{3.168a}$$
$$u = 0 \quad \text{on } \Gamma := \partial\Omega, \tag{3.168b}$$

where Ω is a two-dimensional bounded convex polygonal domain.

Suppose that Ω is partitioned by a shape-regular triangulation $\mathcal{T}_h$. Let V be the space of continuous piecewise linear functions on $\mathcal{T}_h$ that vanish on Γ. We seek a solution u_h of the problem

$$a_h(u_h, v_h) = f_h(v_h) \quad \forall v_h \in V_h, \tag{3.169}$$

where

$$a_h(w, v) := \sum_{T \in \mathcal{T}_h} (\hat{\varepsilon} \nabla w, \nabla v)_T - \sum_{T \in \mathcal{T}_h} (\mathrm{div}(\hat{\varepsilon} \nabla w), \delta v_x)_T + (w_x, v + \delta v_x),$$

$$f_h(v) := (f, v) + \sum_{T \in \mathcal{T}_h} \delta_T (f, v_x)_T,$$

$$h_T := \mathrm{diam}(T), \qquad \hat{\varepsilon}|_T = \hat{\varepsilon}(u_h)|_T := \max\{\varepsilon, C_2 h_T^2 |f - (u_h)_x|\},$$

$$\delta_T := C_1 \max\{0, h_T - \hat{\varepsilon}\}.$$

Here $(\cdot, \cdot)_T$ denotes the $L_2(T)$ inner product, $(\cdot, \cdot)$ is the $L_2(\Omega)$ inner product, δ is the SD parameter, and C_1 and C_2 are user-chosen positive constants.

This SDFEM is similar to that of Section II.5.1. It is a variant of the standard SDFEM of Section 3.2.1 that is *nonlinear* because the piecewise constant function $\hat{\varepsilon}$ depends on the computed solution u_h.

We treat $\hat{\varepsilon}$ as a smoothly varying function (this can be achieved by local averaging and approximation). In practice, when $\varepsilon \ll h := \max_T h_T$, one expects for example that $\hat{\varepsilon} = \mathcal{O}(h^3)$ away from layers and $\hat{\varepsilon} = \mathcal{O}(h)$ inside an exponential boundary layer.

To begin the analysis, express the error $u - u_h$ as

$$u - u_h = \rho + \theta, \tag{3.170}$$

where $\rho := u - \hat{u}, \quad \theta := \hat{u} - u_h$, and $\hat{u}$ is the solution of

$$-\mathrm{div}(\hat{\varepsilon} \nabla \hat{u}) + \hat{u}_x = f \quad \text{in } \Omega, \tag{3.171a}$$
$$\hat{u} = 0 \quad \text{on } \Gamma. \tag{3.171b}$$

Since $\hat{\varepsilon} \geq \varepsilon$, (3.171) is a regularization of (3.168). Regarding $\hat{\varepsilon}$ as known *a priori*, note that u_h is the solution obtained when the standard SDFEM is applied to (3.171). Thus θ is the error when (3.171) is solved by a *linear* method. Furthermore, the choice of $\hat{\varepsilon}$ is so suited to the mesh that one can prove almost optimal *a posteriori* estimates for θ. The estimates available for ρ are less satisfactory.

Let n be the outward-pointing unit normal to Γ. Set $\Gamma = \Gamma_- \cup \Gamma_0 \cup \Gamma_+$, where Γ_-, Γ_0 and Γ_+ denote the parts of Γ where the x-component of n is negative, zero and positive, respectively. Then Γ_- is the inflow boundary for the reduced problem $u_x = f$, while Γ_+ is the outflow boundary. The analysis assumes that $\mathrm{dist}(\Gamma_-, \Gamma_+) > 0$.

Lemma 3.132. *[EJ93b] Assume that, for some generic positive constant C, the inequality $\hat{\varepsilon}_x \leq C\hat{\varepsilon}$ holds true in some neighbourhood of Γ_- and one has $-C\min\{1, \hat{\varepsilon}\} \leq \hat{\varepsilon}_x \leq C$ on Ω. Then*

$$\|\theta\|_0 \leq \mathcal{E}_\theta(u_h, h, f) := C\left[\| \min {}^*\{1, h^2 \hat{\varepsilon}^{-1}\} R(u_h)\|_0 + (\max_{\Gamma_-} \hat{\varepsilon}^{1/2})\|f\|_0\right],$$
(3.172)

where for each $T \in \mathcal{T}_h$,

$$\min {}^*\{1, s\}|_T := \begin{cases} 1 & \text{if } T \cap \Gamma_- \text{ is nonempty,} \\ \min\{1, s\} & \text{otherwise,} \end{cases}$$

$$R(u_h)|_T := \left(|f - (u_h)_x + \nabla\hat{\varepsilon} \cdot \nabla u_h|\right)\big|_T + \hat{\varepsilon}\big|_T \left[\frac{1}{2} \sum_{\lambda \in \partial T} \left([\nabla u_h]_\lambda / h_\lambda\right)^2\right]^{1/2},$$

h_λ is the length of the edge λ of T, and $[\cdot]_\lambda$ is the jump across λ.

Proof. There are too many technical details to be given in full, so only the main ingredients of the proof are presented here. Let z be the solution of the dual problem

$$L^*_{\hat{\varepsilon}} z := -\mathrm{div}(\hat{\varepsilon}\nabla z) - z_x = \theta \quad \text{on } \Omega, \qquad z = 0 \quad \text{on } \Gamma.$$
(3.173)

Because the functional $J(\varphi) = (\varphi, \theta)$ is used we recognize the DWR method for estimating the L_2 error. One has as usual the error representation

$$\|\theta\|_0^2 = (\theta, \theta) = (\theta, L^*_{\hat{\varepsilon}} z) = (f, z) - (\hat{\varepsilon}\nabla u_h, \nabla z) - ((u_h)_x, z).$$

Instead of using some approximation z_h of z, the interpolant z^I to z from V is introduced. Using the projection property implied by (3.169), one obtains

$$\|\theta\|_0^2 = (f - (u_h)_x + \nabla\hat{\varepsilon} \cdot \nabla u_h, z - z^I - \delta z_x^I)$$
$$+ \sum_{T \in \mathcal{T}_h} \sum_{\lambda \in \partial T} \int_\lambda \hat{\varepsilon} \frac{\partial u_h}{\partial n_T} (z - z^I) \, d\lambda,$$

where n_T is the outward-pointing unit normal to ∂T.

The rest of the proof entails careful estimates of the above terms, invoking interpolation error estimates and an analogue of the stability bound (1.14) that is applied to (3.173).   □

In practice $R(u_h)$ will be small away from layers. Inside a layer, where $f - (u_h)_x$ is large, one envisages that $R(u_h) \approx |f - (u_h)_x|$, so locally

$$\min\{1, h^2 \hat{\varepsilon}^{-1}\} R(u_h) \approx \min\{R(u_h), C_2^{-1}\} \leq C_2^{-1}.$$

Hence it is reasonable to expect that $\mathcal{E}_\theta$ will be small.

The next lemma bounds ρ. No proof is given here – it proceeds analogously to the proof above via the dual problem $L_{\hat{\varepsilon}}^* w = \rho$.

Lemma 3.133. *[EJ93b] Assume that $|\hat{\varepsilon}_y| \leq C\hat{\varepsilon}^{1/2}$ and $|\hat{\varepsilon}_x| \leq C\hat{\varepsilon}$ in Ω. Then*

$$\|\rho\|_0 \leq \mathcal{E}_\rho(\hat{u}, \hat{\varepsilon}, f) := C\Big[\|(\hat{\varepsilon} - \varepsilon)\hat{u}_x\|_0 + \|((\hat{\varepsilon} - \varepsilon)\hat{u}_y)_y d_+\|_0$$
$$+ (\max_{\Gamma_-} \hat{\varepsilon}^{1/2})\|f\|_0\Big],$$

where d_+ is the distance to the outflow boundary Γ_+ in the direction (1,0).

Since $\hat{u}$ is unknown we are not quite able to compute $\mathcal{E}_\rho$, so this is not a true *a posteriori* estimate.

Theorem 3.134. *Assume that the hypotheses of Lemmas 3.132 and 3.133 are satisfied. Then*

$$\|u - u_h\|_0 \leq \mathcal{E}_\theta(u_h, h, f) + \mathcal{E}_\rho(\hat{u}, \hat{\varepsilon}, f), \tag{3.174}$$

where $\mathcal{E}_\theta$ and $\mathcal{E}_\rho$ are defined in these lemmas.

Proof. Combine (3.170) with Lemmas 3.132 and 3.133.   □

Remark 3.135. Suppose that the SDFEM is applied to (3.168). If instead of the strong stability bound (1.14) one uses the L_2-stability bound $\|u\|_0 \leq C\|f\|_0$, one can then prove the *a posteriori* estimate

$$\|u - u_h\|_0 \leq C\|R(u_h)\|_0,$$

where $R(u_h)$ is as in Lemma 3.132. Nevertheless, on meshes that are coarse relative to ε, the quantity $\|R(u_h)\|_{0,T}$ is typically large for triangles T that lie near layers. Thus $\|R(u_h)\|_0$ (without the $\min^*$ factor of (3.172)) increases as the mesh is refined even though the actual solution becomes more accurate. It is thus an unsuitable refinement indicator for an adaptive algorithm.   ♣

Eriksson and Johnson [EJ93b] consider two adaptive methods based on the above analysis. Their first method uses $\mathcal{E}_\theta(u_h, h, f)$ as a refinement indicator, and refines the mesh until $\mathcal{E}_\theta$ is below a prescribed tolerance *and* $\hat{\varepsilon} = \varepsilon$ everywhere. Assuming that the algorithm terminates, the requirement $\hat{\varepsilon} = \varepsilon$ implies that $\rho = 0$ for the final solution computed, so Lemma 3.132 shows that $\mathcal{E}_\theta$ will then provide a computable upper bound on $\|u - u_h\|_0$. If T is any triangle of the final mesh that meets an exponential boundary layer, then

$$\varepsilon = \hat{\varepsilon}|_T \geq C_2 h_T^2 \max_T |f - (u_h)_x| \geq C h_T,$$

so the layer (which has width $\mathcal{O}(\varepsilon|\ln \varepsilon|)$) will be resolved. Along a parabolic boundary layer, we expect that $|f - (u_h)_x|$ is $\mathcal{O}(1)$, whence $\varepsilon \geq C h_T^2$ and the layer (now of width $\mathcal{O}(\varepsilon^{1/2}|\ln \varepsilon|)$) is again resolved.

The second adaptive method from [EJ93b] uses the refinement indicator

$$TOL := \mathcal{E}_\theta(u_h, h, f) + C\Big[\|(\hat{\varepsilon} - \varepsilon)(u_h)_x\|_0 + \|(\hat{\varepsilon} - \varepsilon)d_+ D_y^h(u_h)_y\|_0$$
$$+ (\max_{\Gamma_-} \hat{\varepsilon}^{1/2})\|f\|_0\Big],$$

where $D_y^h(u_h)_y$ is a discrete analogue of $(u_h)_{yy}$. The indicator is clearly based on (3.174), but with a computable discrete approximation replacing $\mathcal{E}_\rho(\hat{u}, \hat{e}, f)$. This replacement means that we do not have a rigorous upper bound on $\|u - u_h\|_0$. Heuristic arguments [EJ93b] lead us to believe that $\|u - u_h\|_0$ will be of order $h^{3/8}$ and that, depending on the value of TOL prescribed by the user, the method may or may not resolve layers. For instance, suppose that only an exponential boundary layer is present and that each triangle T that meets this layer has diameter h_T. Then $(|f - (u_h)_x|)|_T = \mathcal{O}(h_T^{-1})$, so $\hat{\varepsilon}|_T = \mathcal{O}(h_T)$. Now if the boundary layer computed occupies a region Ω_γ of width γ, then $\|(u_h)_x\|_0 = \mathcal{O}(1) + \big[\int_{\Omega_\gamma}(1/h_T)^2\big]^{1/2} = \mathcal{O}(\sqrt{\gamma}/h_T)$. Hence

$$TOL \geq C\|(\hat{\varepsilon} - \varepsilon)(u_h)_x\|_0 \geq C h_T(\sqrt{\gamma}/h_T) = \mathcal{O}(\sqrt{\gamma}).$$

Thus the layer is resolved, i.e., $\gamma \leq \mathcal{O}(\varepsilon)$ if $TOL \leq \mathcal{O}(\varepsilon^{1/2})$. More generally, if TOL is $\mathcal{O}(\varepsilon^{1/4})$, then no layers are resolved; if TOL is between $\mathcal{O}(\varepsilon^{1/2})$ and $\mathcal{O}(\varepsilon^{1/4})$, then only parabolic layers are resolved; if TOL is $\mathcal{O}(\varepsilon^{1/2})$, then all layers are resolved.

Remark 3.136. The case of a homogeneous Neumann boundary condition on Γ_+, so u has a weak outflow boundary layer, is also considered in [EJ93b]. Furthermore, the authors allow ε to be a variable function of x and y.   ♣

The above approach defines the dual problem on the basis of the non-stabilized problem. In more recent papers [HRS00, Ran98] for stabilized finite element approximations of first-order hyperbolic problems that correspond to $\varepsilon = 0$, the authors consider two alternative dual problems: the formal adjoint

of the given problem and the transpose of the bilinear form for the stabilized method. It it found that the second approach is superior in the sense that it leads to sharper *a posteriori* error bounds and more economical adaptively refined meshes. We do not know any detailed studies of the related question for the SDFEM or GLSFEM applied to our standard convection-diffusion problem in the case $\varepsilon \neq 0$. It seems that the structure of the problem that is dual to the stabilized problem causes difficulties.

In [CS07] the authors consider the residual-free bubble approach and define the dual problem using the non-stabilized problem. The dual solution is approximated by the solution of a local problem on a refined mesh. The error that arises is estimated by applying the dual approach again and invoking stability estimates.

The DWR method can also be applied to the dGFEM. It turns out that it is fruitful to use the SIP version of the dGFEM for *a posteriori* error estimation because (unlike the NIP version) it is adjoint consistent; see [GHH07b, HHSS03]. In [HGH08] the authors sketch an adaptive strategy for anisotropic mesh generation; in two dimensions, a simple Cartesian refinement is used where an element marked for refinement is subdivided either anisotropically or isotropically in one of three possible ways. To choose the "optimal" refinement a trial and error philosophy is used: local error indicators are computed on the possible refinements and a decision is taken based on which of these is predicted to give the smallest error indicator.

4

Time-Dependent Problems

In this chapter we discuss convection-diffusion and reaction-diffusion problems whose solutions are time-dependent (as in Part II) and are functions of more than one space variable (as in the preceding chapters of Part III). The analysis and numerical methods used are often combinations and extensions of techniques that appeared in this earlier material, so the chapter is relatively short, but new ideas such as dimension-splitting will also make their debuts.

Let $\Omega = (0,1) \times (0,1)$ be the unit square in the (x,y)-plane, with boundary $\partial\Omega$, and set $Q = \Omega \times (0,T]$, where T is a positive constant. Consider the initial-boundary value problem

$$Lu := u_t - \varepsilon\Delta u + b \cdot \nabla u + cu = f \quad \text{on } Q, \qquad (4.1\text{a})$$

where $\Delta u := u_{xx} + u_{yy}$ and $\nabla u = (u_x, u_y)$, with initial-boundary conditions

$$u(x,y,0) = s(x,y) \quad \text{on } \Omega, \qquad (4.1\text{b})$$

$$u(x,y,t) = 0 \quad \text{on } \partial\Omega \times (0,T]. \qquad (4.1\text{c})$$

As usual ε is a parameter satisfying $0 < \varepsilon \ll 1$. The function s is assumed to be smooth on $\bar{\Omega} := [0,1] \times [0,1]$ and $b = (b_1, b_2)$, c and f are assumed to be smooth on $\bar{Q} := \bar{\Omega} \times [0,T]$. For simplicity we have taken homogeneous Dirichlet boundary conditions on the lateral surface $Q_\ell := \partial\Omega \times (0,T]$, but it is straightforward to extend most of the contents of the chapter to the case of inhomogeneous Dirichlet boundary conditions given by some smooth function.

Without loss of generality, one may assume that $c \geq \gamma > 0$ on $\bar{Q}$ for some constant γ since this can easily be obtained by the change of variable $u(x,t) = v(x,t)e^{Ct}$ with some suitable constant C.

When b is not identically zero, L is a *convection-diffusion* operator, where $-\varepsilon\Delta u$ models diffusion while $u_t + b \cdot \nabla u$ represents convection. If $b \equiv (0,0)$ on Q, then L is a *reaction-diffusion* operator.

Our discussion in this chapter will be largely confined to (4.1), which is posed in terms of two space variables. Nevertheless this will provide an adequate preparation for the reader who wishes to work with time-dependent problems with a larger number of space variables.

4.1 Analytical Behaviour of Solutions

In the convection-diffusion case, initial data from Ω and boundary data from those parts of Q_ℓ where b points into Q are transported across Q by the convective operator $u_t + b \cdot \nabla u$, while exponential boundary layers typically form at those parts of Q_ℓ where b points out of Q; cf. Theorem II.2.6.

To justify this statement, we follow Clavero et al. [CJLS98, Appendix A] in sketching a proof of an S-decomposition of the solution u of (4.1). More details can be found in [Shi92b, Shi], and in fact the spatial domains Ω considered there are n-dimensional with $n \geq 2$.

Assume that $b = (b_1, b_2) > (\beta_1, \beta_2) > (0, 0)$ on $\bar{Q}$, where β_1, β_2 are some constants. Then $\Omega \cup (Q_\ell \cap \{(x, y, t) : xy = 0\})$ is the inflow boundary of $\bar{Q}$; the remaining two sides of Q_ℓ, where $x = 1$ or $y = 1$, form the outflow boundary. That is, the inflow boundary is where the subcharacteristics of L in (x, y, t)-space enter $\bar{Q}$ and the outflow boundary is where they leave $\bar{Q}$.

Assume that the data b, c, f, s are sufficiently smooth and that enough compatibility conditions are satisfied on $\partial\Omega$ so that $u(x, y, t)$ lies in the Hölder space $C^{l+\alpha}(Q)$, which is defined analogously to the space $C^{2+\alpha}$ of Section II.2.1. (Here l is a positive integer and $0 < \alpha < 1$.) See [LSU67, Section IV.5] for a precise version of this assertion under the hypothesis that $\partial\Omega$ is smooth.

The solution u will be decomposed as $u = U + V$, where U is the smooth component of u and V comprises various layers. First, extend Ω beyond $x = 1$ and $y = 1$ to a larger domain Ω^* whose boundary is smooth except at the point $(0, 0)$. Set $Q^* = \Omega^* \times [0, T]$. Form smooth extensions b^*, c^*, f^* of the functions b, c, f to $\bar{Q}^*$ and a smooth extension s^* of s to $\bar{\Omega}^*$. Define U^* on $\bar{Q}^*$ as the solution of the initial-boundary value problem

$$U_t^* - \varepsilon \Delta U^* + b^* \cdot \nabla U^* + c^* U^* = f^* \quad \text{on } Q^*,$$
$$U^*(x, y, 0) = s^*(x, y) \quad \text{on } \Omega^*,$$
$$U^*(x, y, t) = g^*(x, y, t) \quad \text{on } \partial\Omega^* \times (0, T],$$

where the function g^* is smooth, compatible with the other data, and satisfies $g(x, y) = 0$ when $xy = 0$. Now U^* will have layers near the outflow boundary of $\bar{Q}^*$ but will otherwise be smooth; since this outflow boundary lies outside Q by our construction, defining U to be the restriction of U^* to $\bar{Q}$ yields a smooth function that satisfies $LU = f$ on Q, $U = s$ on Ω, and $U = 0$ on $Q_\ell \cap \{(x, y, t) : xy = 0\}$, the lateral part of the inflow boundary.

The layer component V satisfies $V = V_1 + V_2 + V_{12}$, where V_1 and V_2 are exponential boundary layers at the sides $x = 1$ and $y = 1$ respectively of Q_ℓ and V_{12} is an edge layer lying along $\{(1, 1, t) : 0 \leq t \leq T\}$. To define V_1, enlarge Ω slightly to Ω^{**} while removing the corner at $(1,1)$ by extending the line $x = 1$ beyond $(1,1)$ then joining it in a smooth way with the line $y = 1$. Define $\partial\Omega_x^{**}$ to be the line segment obtained when this extension is made. Form smooth extensions b^{**} etc. of our various functions to Ω^{**}. Define V_1^{**}

on $\bar{Q}^{**} := \bar{\Omega}^{**} \times [0, T]$ as the solution of the problem

$$(V_1)_t^{**} - \varepsilon \Delta V_1^{**} + b^{**} \cdot \nabla V_1^{***} + c^{**} V_1^{**} = 0 \quad \text{on } Q^{**},$$
$$V_1^{**}(x, y, 0) = 0 \quad \text{on } \Omega^{**},$$
$$V_1^{**}(x, y, t) = -U^{**}(x, y, t) \quad \text{on } \partial \Omega_x^{**} \times (0, T],$$
$$V_1^{**}(x, y, t) = 0 \quad \text{on } \partial \Omega^{**} \setminus \partial \Omega_x^{**} \times (0, T].$$

Then V_1 is the restriction of V_1^{**} to $\bar{Q}$. Similarly define V_2 by focussing on the line $y = 1$.

Finally, V_{12} is defined as the solution of the problem

$$(V_{12})_t - \varepsilon \Delta V_{12} + b \cdot \nabla V_{12} + c V_{12} = 0 \quad \text{on } Q,$$
$$V_{12}(x, y, 0) = 0 \quad \text{on } \Omega,$$
$$V_{12}(x, y, t) = -(U + V_1 + V_2)(x, y, t) \quad \text{on } \partial \Omega \times (0, T].$$

Clearly $u = U + V_1 + V_2 + V_{12}$. It can be shown [Shi92b, Shi] that there exists a constant C such that

$$\left| \frac{\partial^{i+j+k} U(x, y, t)}{\partial x^i \partial y^j \partial t^k} \right| \leq C, \tag{4.2a}$$

$$\left| \frac{\partial^{i+j+k} V_1(x, y, t)}{\partial x^i \partial y^j \partial t^k} \right| \leq C \varepsilon^{-i} \exp^{-\beta_1 (1-x)/\varepsilon}, \tag{4.2b}$$

$$\left| \frac{\partial^{i+j+k} V_2(x, y, t)}{\partial x^i \partial y^j \partial t^k} \right| \leq C \varepsilon^{-j} \exp^{-\beta_2 (1-y)/\varepsilon}, \tag{4.2c}$$

$$\left| \frac{\partial^{i+j+k} V_{12}(x, y, t)}{\partial x^i \partial y^j \partial t^k} \right| \leq C \varepsilon^{-i-j} \min \left\{ \exp^{-\beta_1 (1-x)/\varepsilon}, \exp^{-\beta_2 (1-y)/\varepsilon} \right\}, \tag{4.2d}$$

for $i + j + 2k \leq l$ and $(x, y, t) \in \bar{Q}$. Bounds like (4.2) exclude the presence of interior layers.

If $b \equiv (0, 0)$ so (4.1a) is of reaction-diffusion type, then the asymptotic structure of the solution u is quite different; cf. Remark II.2.11. An S-decomposition of u is given in [CJLS00, Shi93] which shows that u has an essentially one-dimensional parabolic boundary layer along each of the four faces of Q_ℓ and a corner parabolic layer along each of the four edges of Q_ℓ.

4.2 Finite Difference Methods

For general discussions of the ramifications that moving from one to two space variables introduces into the analysis and construction of finite difference methods for time-dependent problems, see the limpid exposition in Strikwerda [Str04, Sections 7.2 and 7.3] (though some of the comments there apply only to systems of partial differential equations) and the wide-ranging

presentation of Hundsdorfer and Verwer [HV03, Section III.6 and Chapter IV], which explicitly considers the solution of convection-dominated problems.

Much of the theory and terminology of Chapter II.3 carries over to (4.1). In particular the von Neumann L_2-stability analysis of Section II.3.1.3 can still be applied on equidistant meshes: in the difference scheme, replace the computed solution at each mesh point (x_j, y_k, t_m) by $\xi^m e^{i(j\theta_1 + k\theta_2)}$ then require that $|\xi| \leq 1$ for stability. Hindmarsh et al. [HGG84] perform a von Neumann stability analysis of various numerical methods for (4.1). Examples will also be found in [HV03] and [Str04].

In [DRH98], Donea et al. systematically examine various ways of constructing finite difference methods that are high-order accurate in time for the numerical solution of (4.1): explicit Taylor-Galerkin methods, explicit and implicit multistage methods based on Padé approximations of the exponential function, explicit and implicit Runge-Kutta-based methods, and implicit methods based on Newton-Cotes quadrature approximation of the integrated time derivative. The formal order of accuracy of each method (i.e., its rate of convergence when ε is regarded as a constant – cf. Remark II.3.15) and its stability, phase accuracy and damping error are analysed.

On $\bar{Q}$, consider tensor-product meshes

$$\{(x_i, y_j, t_k) : i = 0, \dots, M, \ j = 0, \dots, N, \ k = 0, \dots, K\}$$

that are equidistant in the t-direction with spacing τ. Set $h_x = \max_i\{x_i - x_{i-1}\}$ and $h_y = \max_j\{y_j - y_{j-1}\}$. Write $u_{i,j}^k$ for the solution computed by some difference scheme at the mesh point (x_i, y_j, t_k).

Shishkin meshes for (4.1) are constructed as follows. Assume that $b = (b_1, b_2) > (\beta_1, \beta_2)$ on $\bar{Q}$ for some positive constants β_1 and β_2. Given an even positive integer N, set

$$\lambda_x = (\kappa\varepsilon/\beta_1)\ln N \quad \text{and} \quad \lambda_y = (\kappa\varepsilon/\beta_2)\ln N,$$

where one typically takes $\kappa \geq 2$ (see Remarks I.2.99 and I.2.104). Then along the x axis, divide each of $[0, 1 - \lambda_x]$ and $[1 - \lambda_x, 1]$ into $N/2$ equal intervals. Divide the y-axis similarly using λ_y. A tensor product of these one-dimensional meshes yields a Shishkin mesh on $\bar{\Omega}$; this is the same as Figure 2.1. Take a tensor product of this two-dimensional mesh with the equidistant mesh on $[0, T]$ to yield the final mesh.

In [Shi90a] Shishkin applies a form of upwinding on this mesh and proves, under certain regularity hypotheses on the data, that the computed solution $\{u_{i,j}^k\}$ satisfies

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \leq C(N^{-1}\ln^2 N + \tau) \quad \text{for all } i, j \text{ and } k.$$

The problem considered in [Shi90a] is in fact much more general than (4.1): it is posed in $n\,(\geq 1)$ space dimensions and allows the coefficients of the differential equation to have discontinuities along a finite number of hyperplanes

parallel to the t-axis. These discontinuities cause interior layers in the solution that are handled by a generalization of the grid described above.

Generalizations to systems of two equations are examined in [Shi98b, Shi98c]. Each equation has its own small diffusion coefficient and the relative sizes of these govern the types of layer in the true solution, and also affect the rates of convergence that are proved for the computed solutions.

Boglaev [Bog05] solves a semilinear generalization of (4.1) using upwinding and a Schwarz method based on a decomposition of Ω into overlapping parallel strips; upper and lower solutions are computed that enclose the true solution u. Several theoretical results are proved but these do not include uniform convergence although numerical results are encouraging.

When the domain Ω is multi-dimensional as in (4.1), one can seek to split the problem into smaller components for reasons of efficiency when time-stepping. An encyclopedic discussion of the many ways in which this can be done is given in [HV03, Chapter IV].

We now discuss *dimension-splitting*. This concept has no previous counterpart in our book. Its basic idea, when applied to (4.1), is that instead of solving a two-dimensional (in space) problem at each discrete time step, one solves a sequence of two problems each of which is one-dimensional in space.

The particular form of dimension-splitting that we now derive, following the description in [Str04, Section 7.3], is the *alternating direction implicit (ADI) method*, which is suited to rectangular (x, y)-domains such as Ω. Write (4.1a) as $u_t = A_1 u + A_2 u$, where A_1 and A_2 are linear operators; for example one could take the splitting

$$A_1 u = \varepsilon u_{xx} - b_1 u_x - c_1 u + f_1, \quad A_2 u = \varepsilon u_{yy} - b_2 u_y - c_2 u + f_2 \qquad (4.3)$$

with $c_1 + c_2 = c$ and $f_1 + f_2 = f$. Let u^k denote the solution to (4.1) at each discrete time level $k\tau$. Then

$$\frac{u^{k+1} - u^k}{\tau} = \frac{1}{2}\left(A_1 u^{k+1} + A_1 u^k\right) + \frac{1}{2}\left(A_2 u^{k+1} + A_2 u^k\right) + \mathcal{O}(\tau^2),$$

where by $\mathcal{O}(\tau^2)$ we mean an error for fixed ε. Rewrite this as

$$\left(I - \frac{\tau}{2}A_1 - \frac{\tau}{2}A_2\right) u^{k+1} = \left(I + \frac{\tau}{2}A_1 + \frac{\tau}{2}A_2\right) u^k + \mathcal{O}(\tau^3).$$

Add $\tau^2 A_1 A_2 u^{k+1}/4$ to both sides; one can then factor, obtaining

$$\left(I - \frac{\tau}{2}A_1\right)\left(I - \frac{\tau}{2}A_2\right) u^{k+1} = \left(I + \frac{\tau}{2}A_1\right)\left(I + \frac{\tau}{2}A_2\right) u^k$$

$$+ \frac{\tau^2}{4} A_1 A_2 (u^{k+1} - u^k) + \mathcal{O}(\tau^3)$$

$$= \left(I + \frac{\tau}{2}A_1\right)\left(I + \frac{\tau}{2}A_2\right) u^k + \mathcal{O}(\tau^3).$$

This equation is solved numerically by choosing discrete approximations A_{1d} and A_{2d} of A_1 and A_2, yielding the ADI method

$$\left(I - \frac{\tau}{2}A_{1d}\right)\left(I - \frac{\tau}{2}A_{2d}\right)u_{h,\tau}^{k+1} = \left(I + \frac{\tau}{2}A_{1d}\right)\left(I + \frac{\tau}{2}A_{2d}\right)u_{h,\tau}^{k}, \qquad (4.4)$$

where $u_{h,\tau}^{k}$ is the computed solution at the time level $t = k\tau$. The advantage of (4.4) is that $u_{h,\tau}^{k+1}$ can be computed by sequentially inverting $I - (\tau/2)A_{1d}$ and $I - (\tau/2)A_{2d}$, each of which usually can be inverted much more easily than any typical discretization of the full two-dimensional differential operator in (4.1a).

One popular ADI method is the Peaceman-Rachford algorithm [PR55], where the iteration used to compute $u_{h,\tau}^{k+1}$ from $u_{h,\tau}^{k}$ is

$$\left(I - \frac{\tau}{2}A_{1d}\right)u_{h,\tau}^{k+1/2} = \left(I + \frac{\tau}{2}A_{2d}\right)u_{h,\tau}^{k}, \qquad (4.5a)$$

$$\left(I - \frac{\tau}{2}A_{2d}\right)u_{h,\tau}^{k+1} = \left(I + \frac{\tau}{2}A_{1d}\right)u_{h,\tau}^{k+1/2}. \qquad (4.5b)$$

Here one first computes $u_{h,\tau}^{k+1/2}$, then $u_{h,\tau}^{k+1}$. These formulas clarify the origin of the terminology ADI: the two steps alternate the coordinate direction that implicitly determines $u_{h,\tau}$. To see that (4.5) yields the same solution as (4.4), use (4.5b) then (4.5a):

$$\begin{aligned}
\left(I - \frac{\tau}{2}A_{1d}\right)\left(I - \frac{\tau}{2}A_{2d}\right)u_{h,\tau}^{k+1} &= \left(I - \frac{\tau}{2}A_{1d}\right)\left(I + \frac{\tau}{2}A_{1d}\right)u_{h,\tau}^{k+1/2} \\
&= \left(I + \frac{\tau}{2}A_{1d}\right)\left(I - \frac{\tau}{2}A_{1d}\right)u_{h,\tau}^{k+1/2} \\
&= \left(I + \frac{\tau}{2}A_{1d}\right)\left(I + \frac{\tau}{2}A_{1d}\right)u_{h,\tau}^{k}.
\end{aligned}$$

Note that commutativity of the operators A_{1d} and A_{2d} was not needed in this argument, unlike in some ADI methods.

We now describe some methods for (4.1) that are based on dimension-splitting. All assume that $b = (b_1, b_2) \geq (\beta_1, \beta_2) > (0, 0)$ on $\bar{Q}$.

In [CGJ06b], Clavero et al. assume that the *a priori* bounds (4.2) hold true with $l = 6$ and that $b = b(x, y)$ and $c = c(x, y)$ are independent of t. A Peaceman-Rachford ADI method based on the splitting (4.3) is applied, then a polynomial-based HODIE technique on a one-dimensional Shishkin mesh with N intervals is used to generate the discrete operators A_{1d} and A_{2d}. Assuming that $N^{-q} \leq \tau^2$ with $0 < q < 1$, it is shown in [CGJ06b] that

$$|u(x_i, y_j, t_k) - u_{i,j}^{k}| \leq C(N^{-2+q}\ln^2 N + \tau) \quad \text{for all mesh points } (x_i, y_j, t_k),$$

where C is some positive constant, but the analysis is very intricate and the amount of compatibility assumed is perhaps excessive.

A related dimension-splitting method is considered in [CJLS98] under the hypothesis that (4.2) hold true with $l = 4$. The discretization employs the fractional-step scheme

$$(I + \tau A_{1d})u_{h,\tau}^{k+1/2} = u_{h,\tau}^k + \tau f_1(x, y, t_{k+1}),$$

$$u_{h,\tau}^{k+1/2}(0, y, t_{k+1/2}) = u_{h,\tau}^{k+1/2}(1, y, t_{k+1/2}) = 0,$$

$$(I + \tau A_{2d})u_{h,\tau}^{k+1} = u_{h,\tau}^{k+1/2} + \tau f_2(x, y, t_{k+1}),$$

$$u_{h,\tau}^{k+1}(x, 0, t_k) = u_{h,\tau}^{k+1}(x, 1, t_k) = 0,$$

where A_1 and A_2 are defined by (4.3) then discretized by simple upwinding on a one-dimensional Shishkin mesh with N intervals. At each time step one need only solve a set of uncoupled tridiagonal systems. Assuming that $N^{-q} \leq \tau$ with $0 < q < 1$, the analysis leads to the inequality

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \leq C(N^{-1+q} \ln N + \tau)$$

at all mesh points, where $\{u_{i,j}^k\}$ is the computed solution and C is some constant.

The dimension-splitting of [CJL93] is also based on (4.3), but in discretizing A_1 and A_2 a one-dimensional equidistant mesh is used and on it a HODIE technique is applied to generate exponentially fitted schemes. It is shown that

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \leq C(N^{-1} + \tau)$$

at all mesh points. The analysis uses a consistency error estimate and a barrier function (compare the proof of Theorem I.2.18).

Reaction-diffusion problems

Consider now the case $b \equiv (0,0)$ in (4.1a), i.e., the differential equation becomes the time-dependent reaction-diffusion problem

$$Lu := u_t - \varepsilon \Delta u + cu = f \quad \text{on } Q, \tag{4.6}$$

with the same initial-boundary conditions as before.

In [Shi93], Shishkin considers a time-dependent reaction-diffusion problem with $n \geq 2$ space variables. Each discrete time interval $[t_k, t_{k+1}]$ is divided into n equal subintervals by the points $\{t_{k,m} : t_{k,m} = t_k + m\tau/n, m = 0, \ldots, n\}$. Splittings $f = \sum_{m=1}^n f_m$ and $c = \sum_{m=1}^n c_m$ are introduced. Then on each time subinterval $[t_{k,m}, t_{k,m+1}]$ one solves the problem

$$\frac{1}{n}u_t - \varepsilon u_{x_m x_m} + c_m u = f_m \tag{4.7}$$

using the original boundary data from Q_ℓ and initial data at $t_{k,m}$ output by the previous computation on $[t_{k,m-1}, t_{k,m}]$. To solve (4.7) numerically, a standard difference scheme on a suitable one-dimensional Shishkin mesh with N subintervals (see Remark I.2.106) approximates $u_{x_m x_m}$ and Euler backward differencing approximates u_t. Assuming sufficient regularity and compatibility

of the data (in particular, interior layers are excluded), it is proved that there exists a constant C such that

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \le C(N^{-2} \ln^2 N + \tau)$$

at all mesh points.

In [CG05] the authors consider (4.6). It is assumed that $c = c(x, y)$ is independent of t, and smoothness and compatibility assumptions are made to justify an S-decomposition of u. Then an A-stable fractional-step Runge-Kutta method is used for the time derivative, while dimension-splitting with (4.3) enables the use of a standard difference scheme on one-dimensional Shishkin meshes to generate A_{1d} and A_{2d}. Assuming that $\varepsilon \le N^{-2}$, the inequality

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \le C(N^{-4} \ln^4 N + \tau^3) \quad \text{at all mesh points}$$

is derived, where C is some constant.

The method of [BCGJ07] solves (4.6) by combining Peaceman-Rachford ADI splitting with a HODIE discretization of the spatial derivatives on a one-dimensional Shishkin mesh with N intervals. Under various hypotheses including the assumptions that $c(x, y)$ is independent of t and the operators $\varepsilon u_{xx} - c_1 u$ and $\varepsilon u_{yy} - c_2 u$ commute, where $c_1 + c_2 = c$ with each $c_i > 0$, the analysis leads to

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \le C(N^{-3} + \tau^2)$$

at all mesh points, for some constant C.

Sequential and parallel Schwarz methods for numerically solving a semilinear generalization of (4.6) on two-dimensional Shishkin meshes are analysed in [Shi97a]. Certain regularity and compatibility assumptions are made. Then at all mesh points the computed solutions satisfy

$$|u(x_i, y_j, t_k) - u_{i,j}^k| \le C(N^{-1} \ln N + \tau).$$

Shishkin meshes can also be used to handle transient layers in u at $t = 0$; these arise in the reaction-diffusion problem discussed in [Shi98a] where small parameters multiply both Δu and u_t.

4.3 Finite Element Methods

Consider now finite element methods for solving (4.1). Without loss of generality assume that

$$c - \frac{1}{2} \operatorname{div} b > 0 \text{ on } \bar{Q},$$

since this can be easily be obtained by making a change of dependent variable to $v(x, y, t) = u(x, y, t) e^{-Ct}$ for some $C > 0$. Also assume for simplicity in this

section that $b \geq (\beta, \beta) > (0,0)$ on $\bar{Q}$, although some results below hold true in more general situations.

We introduce some norm notation that is widely used for time-dependent problems. Let $g(x, y, t)$ be a suitable function defined on Q. Given a standard Sobolev norm $\|\cdot\|_{m,p}$ in $W^{m,p}(\Omega)$ with $m \geq 0$ and $1 \leq p \leq \infty$, for $1 \leq q \leq \infty$ define the $L_q(0, T; W^{m,p}(\Omega))$ norm by

$$\|g\|_{L_q(0,T;W^{m,p}(\Omega))} = \left\| \left(\|g(\cdot, \cdot, t)\|_{m,p} \right) \right\|_{L_q[0,T]}. \tag{4.8}$$

The finite element methods examined here are taken in the order of the corresponding finite element methods from earlier in the book: first, methods similar to those of Part II but with more than one space dimension, then methods based on combining time-dependence with the techniques of Chapter III.3, and finally dimension-splitting methods.

When solving (4.1) numerically, there is little current interest in the construction of analogues of the space-based methods of Section II.4.1.

The streamline diffusion finite element method of Section II.4.2.1 was based on space-time finite elements while the SDFEM of Section III.3.2.1, which dealt with stationary problems, used finite elements in space only. In the space-time framework, the first-order time and space derivatives of (4.1a) are combined into a single first-order derivative $u_z := u_t + b \cdot \nabla u$ acting in 3-dimensional (x, y, t)-space: the *material derivative*. The same technique was used in one space dimension in Section II.4.2.1. The material derivative approach is advocated by Johnson et al. [JNP84] and by Hughes and Stewart [HS96]. It is easy to generalize the SDFEM defined in equation (II.4.30) to (4.1), as we now describe, following [JNP84, Näv82].

Let $0 = t_0 < t_1 < \cdots < t_J = T$ be a subdivision of $[0, T]$, and on each strip $Q_j := \{(x, t) : x \in \Omega, \ t_j < t < t_{j+1}\}$ let $V^j \subset H^1(Q_j)$ be a finite element subspace comprising piecewise polynomials of degree $k \geq 1$ on a quasi-uniform mesh of diameter h. Set $V_0^j = \{v_h \in V^j : v = 0 \text{ on } \partial\Omega\}$. The SDFEM is applied to (4.1) on each strip Q_j successively, imposing the initial condition at $t = t_j$ weakly and the boundary condition (4.1c) strongly: for $j = 1, \ldots, J$, find $\hat{u}^j \in V_0^j$ such that

$$B_h(\hat{u}^j, \phi) := \varepsilon(\hat{u}_x^j, \phi_x)_{Q_j} - \varepsilon \sum_{K \subset Q_j} (\hat{u}_{xx}^j, \delta_K \phi_z)_K$$
$$+ (\hat{u}_z^j + c\hat{u}^j, \phi + \delta_K \phi_z)_{Q_j} + \langle \hat{u}_+^j, \phi_+ \rangle_{j-1}$$
$$= (f, \phi + \delta_K \phi_z)_{Q_j} + \langle \hat{u}_-^{j-1}, \phi_+ \rangle_{j-1} \qquad \text{for all } \phi \in V_0^j. \tag{4.9}$$

Here the $L_2(\Omega)$ inner product for $t = t_{j-1}$ is denoted by $\langle \cdot, \cdot \rangle_{j-1}$, the function $\hat{u}_-^0(x, y)$ is defined to be $s(x, y)$, and $w_\pm(x, y, t_{j-1}) := \lim_{\theta \to 0+} w(x, y, t_{j-1} \pm \theta)$ for $(x, y) \in \Omega$. The SD parameter δ_K is user-chosen and constant on each element K. The computed solution is in general discontinuous at each time level $t = t_j$.

For this SDFEM based on material derivatives one has the Galerkin orthogonality property $B_h(u - \hat{u}^j, \phi) = 0$ for all $\phi \in V_0^j$, which is a useful identity in convergence analyses.

The following analogues of Theorems II.4.11 and II.4.12 are derived in [Näv82] for (4.9). For each set $\hat{Q} \subset Q$ that is the closure of a union of space-time elements K, and each w that lies in $H^1(K)$ for all $K \subset \hat{Q}$, define

$$|||w|||_{\hat{Q}} := \left\{ \varepsilon \sum_{K \subset \hat{Q}} \|\nabla w\|_{0,K}^2 + \sum_{K \subset \hat{Q}} \delta_K \|w_z\|_{0,K}^2 + \|w\|_{0,\hat{Q}}^2 \right\}^{1/2}.$$

Write $u_{h,\tau}$ for the solution computed by (4.9).

Theorem 4.1. *(Global error bound) Assume that $\varepsilon \le h$ and $\delta_K \le C'h$ for all K and some sufficiently small constant C'. Then for all sufficiently small h (independently of ε), there exists $C > 0$ such that*

$$|||u - u_{h,\tau}|||_Q \le Ch^{k+1/2}|u|_{H^{k+1}(Q)}.$$

Theorem 4.2. *(Local error bound) Assume the hypotheses of Theorem 4.1. Let Q' and Q'' be unions of space-time elements in Q, with $Q' \subset Q''$. Assume that the inflow boundary of Q'' is a subset of the inflow boundary of Q, that the subcharacteristic direction $(b_1, b_2, 1)$ satisfies $|(b_1, b_2, 1) \cdot \nu| \ge C > 0$ on the inflow and outflow boundaries of Q'' where ν is a unit normal on $\partial Q''$, and that all points in Q that are upstream of any point on the characteristic boundary of Q'' (which may be empty) also lie in the characteristic boundary of Q''. Assume that the distance from Q' to the outflow boundary of Q'' is at least $C_2 h |\ln h|$ and to the characteristic boundary of Q'' is at least $C_2 \sqrt{h}\,|\ln h|$, where C_2 is a fixed positive constant chosen in the proof. Then there exists $C > 0$ such that*

$$|||u - u_{h,\tau}|||_{Q'} \le C\big\{ h^{k+1/2}|u|_{H^{k+1}(Q'')} + h^k \big[\|f\|_{L_2(Q)} + \|s\|_{L_2(\Omega)}\big] \big\}.$$

The assumptions on Q'' in Theorem 4.2 ensure that all points in Q that lie upstream of any point in $\bar{Q}''$ also lie in $\bar{Q}''$, so there is no "upstream cut-off".

Remark 4.3. (SDFEM applied only in space) Despite the theoretical attraction of SDFEM methods based on the material derivative, many authors prefer numerical methods that deal separately with the spatial and temporal derivatives in (4.1a): first the SDFEM of Section 2.2.3 is used to discretize $-\varepsilon \Delta u + b \cdot \nabla u + cu = f$ on some partition of Ω, which yields a system of ordinary differential equations in the independent variable t; this is then solved by some standard method for stiff systems. (Similarly, one could combine the other stabilized methods of Chapter 3 with some suitable temporal discretization.) Although Galerkin orthogonality is lost, one can now quickly extend an existing code for elliptic convection-diffusion problems to the time-dependent

situation; furthermore, the number of unknowns at each time step is in general less than when space-time elements are used. But the stability analysis of these spatial SDFEM/temporal finite difference methods raises some subtle issues, as illustrated in [BGS04], where a certain CFL condition is shown to be sufficient but unnecessary for stability.   ♣

Although one might expect the adaptive SDFEM of Section II.5.1 to generalize to higher-dimensional problems, no analysis of this seems to have been published.

Time-dependent versions of the GLSFEM of Section 3.2.2 appear in Lube et al. [LW95, LOM98] and their references. In [LW95] two forms of the stabilizing least-squares terms are analysed: one where stabilization is applied to the full differential operator L and one where it is applied only to the space derivatives. The first choice corresponds to the merging of the first-order space and time derivatives into a material derivative and yields Galerkin orthogonality, while the second is a stabilization in space followed by a separate time discretization, just like the two approaches in use for the SDFEM that we described above. Error estimates at each discrete time level are proved in [LW95] for both methods in the norm defined by

$$\left\{ \sum_{K \in \mathcal{K}} \left[\varepsilon \|\nabla v\|^2_{L_2(K)} + \|v\|^2_{L_2(K)} + \delta_K \| - \varepsilon \Delta v + b \cdot \nabla v + c v \|^2_{L_2(K)} \right] \right\}^{1/2},$$

where Ω is triangulated by $\{K : K \in \mathcal{K}\}$ and δ_K is the local GLS stabilization parameter.

Discontinuous Galerkin finite element methods for (4.1) – recall Section 3.4, where elliptic problems were examined – are still in a stage of rapid development. The examples that follow do not use a special mesh, exponential upwinding, or any other device designed to achieve uniform convergence, so their global error estimates are not uniform in ε.

In [DP01] a dGFEM is generated from a mixed formulation of (4.1) with (4.1c) modified to a homogeneous Neumann condition on the outflow boundary. The authors partition Ω by an arbitrary shape-regular mesh of diameter h and consider the semidiscrete solution u_h of their dGFEM. It is shown that

$$\|u - u_h\|_{L_2(0,T;L_2(\Omega))} \leq M h^{k+1}$$

but the dependence of M on ε is not made clear. The analysis leading to this result exploits the adjoint differential operator and shows clearly how the order of magnitude chosen for the discontinuity-penalization parameter affects the theoretical convergence rate.

A hp finite element method based on a mixed formulation of the dGFEM is used in [CCSS02] to construct a numerical method for the solution of the time-dependent problem in *one* space dimension that is stated in (4.1) of Part II. The differential operator has constant coefficients. An arbitrary space

mesh of diameter h is placed on $[0, 1]$ and piecewise polynomials of degree p are used on each mesh interval. Let u_h denote the resulting semidiscrete solution. The authors prove that there exists a constant C such that

$$\|(u - u_h)(\cdot, T)\|_{L_2[0,1]} + \sqrt{\varepsilon} \left(\int_{t=0}^T \int_{x=0}^1 (u - u_h)_x^2 dx \, dt \right)^{1/2}$$
$$\leq C \frac{h^{\min\{r,p\}+1}}{\max\{1,p\}^{r+1}} \|u^{(r+1)}\|_{\varepsilon,T}$$

for any r such that $\|u^{(r+1)}\|_{\varepsilon,T}$ is defined, where we set

$$\|w\|_{\varepsilon,T} = \max_{0 \leq t \leq T} \|w(\cdot, t)\|_{L_2(0,1)} + \int_{t=0}^1 \|w_t(\cdot, t)\|_{L_2(0,1)}$$
$$+ \sqrt{\varepsilon} \left(\int_{t=0}^T \int_{x=0}^1 w_x^2(\cdot, \cdot) \, dx \, dt \right)^{1/2}.$$

This result has optimal order in terms of h and p.

A dGFEM of asymmetric interior penalty (NIP) form with upwinding (cf. Section 3.4) is applied to (4.1) in [FŠ04], where Ω can be two-dimensional or three-dimensional; the authors consider a semidiscrete approximation (i.e., a method of lines) with piecewise polynomials of degree p on a shape-regular mesh of diameter h and it is demonstrated *inter alia* that the computed semidiscrete solution u_h satisfies

$$\|u - u_h\|_{L_\infty(0,T;L_2(\Omega))} + \sqrt{\varepsilon} \|u - u_h\|_{L_2(0,T;H^1(\Omega,\mathcal{T}))}$$
$$\leq M h^p (\sqrt{\varepsilon} + \sqrt{h}),$$

where $H^1(\Omega, \mathcal{T})$ is the broken H^1 norm (see Section 3.4) corresponding to the triangulation $\mathcal{T}$ of Ω. Here M depends on $\max_{0 \leq t \leq T} |u(\cdot, \cdot, t)|_{H^{p+1}(\Omega)}$, which is in general large when ε is small.

The time-discretization of this method by means of a dGFEM with piecewise polynomials of degree q is considered in [FHŠ07] and it is shown that

$$\|u - u_h\|_{L_2(0,T;L_2(\Omega))} + \sqrt{\varepsilon} \left[\sum_{m=1}^M \int_{I_m} \|u - u_h\|_{H^1(\Omega,\mathcal{T}_{h,m})}^2 \right]^{1/2}$$
$$\leq C h^p \left[|u|_{L_2(0,T;H^{p+1}(\Omega))} + |u|_{L_\infty(0,T;H^{p+1}(\Omega))} \right]$$
$$+ C \tau^q \left[|u|_{H^{q+1}(0,T;L_2(\Omega))} + |u|_{H^{q+1}(0,T;H^1(\Omega))} \right],$$

where u_h is the computed fully discrete solution, τ is the maximum time step and $H^1(\Omega, \mathcal{T}_{h,m})$ is the broken H^1 norm defined on the triangulation $\mathcal{T}_{h,m}$ of Ω at each time-slab $I_m := \Omega \times (t_{m-1}, t_m)$ for $m = 1, \dots, M$. Other functionals of the error $u - u^h$ that arise naturally in the dGFEM formulation are also bounded in [FHŠ07].

Sun and Wheeler [SW05] consider domains in two and three space dimensions and analyse three discontinuous Galerkin methods with interior penalties: symmetric (SIP), asymmetric (NIP) and incomplete (IIP) – this last method simply omits those terms whose sign in Section 3.4 distinguishes SIP from NIP. In (4.1c) a Robin condition is imposed at the inflow boundary and a homogeneous Neumann condition at the outflow. An arbitrary quasi-uniform mesh is used on Ω and piecewise polynomials of degree $p \geq 1$ are used in the dGFEM. The trial space is discontinuous in space but continuous in time. Writing u_h for the semidiscrete solution of any of these three locally conservative methods, it is shown that there exists a constant C such that

$$\|u - u_h\|_{L_\infty(0,T;L_2(\Omega))} + \sqrt{\varepsilon}\,\|u - u_h\|_{L_2(0,T;H^1(\Omega))}$$
$$\leq C \frac{h^{\min\{p+1,r\}-1}}{p^{r-1}} \left[\|u\|_{L_2(0,T;H^r(\Omega))} + \|u_t\|_{L_2(0,T;H^{r-1}(\Omega))} + \|s\|_{H^{r-1}(\Omega))}\right]$$

and for the SIP it is proved that

$$\|u - u_h\|_{L_2(0,T;L_2(\Omega))}$$
$$\leq C h^{\min\{p+1,r\}} \left[\frac{\|u\|_{L_2(0,T;H^r(\Omega))}}{p^{r-1}} + \frac{\|u_t\|_{L_2(0,T;H^{r-1}(\Omega))}}{p^r} + \frac{\|s\|_{H^{r-1}(\Omega))}}{p^{r-1/2}}\right].$$

These bounds hold true for any $r \geq 2$ such that the norms of the right-hand sides are defined. The second estimate is of optimal order in h. Error bounds are also proved in [SW05] for negative norms and when the mesh has hanging nodes, and numerical results illustrate the rates of convergence attained in practice by the various methods.

In [EP05] an *a posteriori* error bound in the $L_2(0,T;L_2(\Omega))$ norm is proved for an NIP dGFEM applied to a variant of (4.1) where a homogeneous Neumann condition is imposed at the outflow boundary.

For the application of the dGFEM to time-dependent nonlinear problems such as the Euler equations, consult the monograph [FFS03] and subsequent papers of Feistauer et al.

Remark 4.4. The SDFEM and GLSFEM have the drawbacks that lumping of the mass matrix is not feasible and space-time elements are needed for consistency. Methods such as stabilization by the penalization of jumps of gradients along edges [Bur05] and dGFEM permit lumping and give more flexibility in the choice of time-stepping.

Recent innovations related to these methods such as multiscale (subgrid) modelling are emerging but have not yet reached maturity; see [Cod98, CB02c, HS07] and their references.   ♣

The ELLAM of Section II.4.2.3 also generalizes to (4.1). In [Wan00], where in the spatial domain piecewise bilinear trial functions are used on an equidistant rectangular mesh with dimensions (h_x, h_y) and the discrete time-step is τ, Wang derives the $L_\infty(0,T;L_2(\Omega))$ optimal-order error estimate

$$\|u(\cdot,\cdot,t_j) - u_{h,\tau}(\cdot,\cdot,t_j)\|_{L_2(\Omega)} \leq K(h_x^2 + h_y^2 + \tau) \quad \text{for all } j;$$

this is a generalization of the bound (4.59) of Part II. Numerical results and implementation issues are examined in [WDE$^+$99].

In [BNV06a, BNV06b] a high-order Lagrange-Galerkin method is analysed in two and three space dimensions for an analogue of (4.1) where b vanishes on $\partial\Omega \times (0,T]$ and error estimates are derived.

We now describe an adaptive Lagrange-Galerkin finite element method from [HS01a] that is designed for (4.1). Like the ELLAM, this method makes explicit use of the space-time subcharacteristics of (4.1a). Take $c = 0$ in (4.1a). Each subcharacteristic $X(x,y,t';\cdot) : [0,T] \to \Omega$, where $(x,y,t') \in \Omega \times (0,T]$, satisfies

$$\frac{d}{dt}X(x,y,t';t) = b(X(x,y,t';t),t), \qquad X(x,y,t';t') = (x,y).$$

The material derivative D_t is defined by

$$D_t u(x,y,t) = \frac{\partial}{\partial t}u(x,y,t) + b(x,y,t) \cdot \nabla u(x,y,t).$$

Write (4.1) in the following weak form: find $u(x,y,t) \in V$ such that

$$\varepsilon(\nabla u(\cdot,\cdot,t), \nabla v) + (D_t u(\cdot,\cdot,t), v) = (f(\cdot,\cdot,t), v) \quad \forall v \in V \text{ and } 0 < t \leq T,$$
$$(u(\cdot,\cdot,0), v) = (s(\cdot,\cdot), v) \quad \forall v \in V,$$

where $V := H_0^1(\Omega)$ and $(\cdot,\cdot)$ is the $L_2(\Omega)$ inner product. Let $0 = t_0 < t_1 < \cdots < t_M = T$ be some subdivision of $[0,T]$. Suppose that at each discrete time level t_m we have a triangulation Ω_m of Ω and a space S_m of piecewise polynomials defined on Ω_m. Now apply a Galerkin finite element method on each Ω_m, while approximating the material derivative by the backward Euler method: for $m = 1, \ldots, M$ find $u_{h,\tau}^m \in S_m$ such that

$$\varepsilon(\nabla u_{h,\tau}^m, \nabla v) + \left(\frac{u_{h,\tau}^m - u_{h,\tau}^{m-1}(X(\cdot,\cdot,t_m;t_{m-1}))}{t_m - t_{m-1}}, v\right)$$
$$= (f^m, v) \quad \forall v \in S_m, \tag{4.10a}$$
$$(u_{h,\tau}^0, v) = (s, v) \quad \forall v \in S_0, \tag{4.10b}$$

where $f^m(\cdot,\cdot) = f(\cdot,\cdot,t_m)$. An *a posteriori* error bound for (4.10) is proved in [HS01a] in the $L_2(0,T; L_2(\Omega))$ norm defined in (4.8) and an adaptive method based on this estimate is constructed to yield a solution $u_{h,\tau}$ for which $\|u - u_{h,\tau}\|_{L_2(0,T; L_2(\Omega))}$ achieves a desired tolerance. The proof of the *a posteriori* estimate uses the methodology of [EEHJ95], which relies on Galerkin orthogonality and strong stability estimates for a problem dual to (4.10).

For further discussion of the ELLAM and its variants see [EW01].

Versions of the *moving mesh method (r-refinement)* of Section II.5.2 have been developed for higher dimensions; see in particular [AF90, Bai94, LBD$^+$02].

Several moving mesh techniques have been considered by Huang, Russell and their coworkers (see the references in [LCHR03]). Most of these methods are concerned with elliptic problems where artificial time-stepping is introduced to drive the mesh movement. In [LCHR03] a problem similar to (4.1) is considered, except that the spatial domain Ω is assumed to be bounded and open with a smooth boundary. A variant of Rothe's method is used to solve the problem numerically: the parabolic differential equation (4.1a) is first discretized in time, yielding a system of spatial problems, each of which is solved by a finite element method with rh-refinement (i.e., mesh points are moved and in addition can be added or deleted). The r-refinement aspect of this method is based on [Hua01], whose approach we now outline.

Let $x = (x_1, x_2)$ denote the physical coordinates in Ω and $\xi = (\xi_1, \xi_2)$ the computational coordinates in some reference domain Ω_c. Adaptive moving meshes for Ω can be generated as images of a reference mesh in Ω_c through a one-to-one time-dependent transformation $x = x(\xi, t')$; here t' is not the time variable of (4.1) but instead an artificial time-like variable that will facilitate r-refinement. Define the mesh adaptation functional

$$I[\xi] = \frac{1}{2} \int_\Omega \sum_{i=1}^{2} (\nabla \xi_i)^T G^{-1} \nabla \xi_i \, dx,$$

where ∇ is the gradient operator with respect to the x variables and the *monitor function* G (cf. Section I.2.5) is a user-chosen 2×2 symmetric positive definite matrix that interconnects the mesh and the physical solution. (A general discussion of the use of monitor functions in multidimensional problems can be found in [Hua06].) One seeks to minimize this functional. The Euler-Lagrange equations associated with $I[\xi]$ are

$$\nabla \cdot (G^{-1} \nabla \xi_i) = 0 \quad \text{for } i = 1, 2.$$

Then the *moving mesh partial differential equations (MMPDE)* that control the mesh movement are chosen to be the modified gradient flow equations

$$\frac{\partial \xi_i}{\partial t'} = \frac{p}{\tau'} \nabla \cdot (G^{-1} \nabla \xi_i), \quad \text{for } i = 1, 2,$$

where the user-chosen quantities τ' and p are, respectively, the artificial time step and a positive function designed to make all mesh points move with a common time scale (which makes the MMPDE easier to integrate). This system of equations determines ξ; for practical purposes it is better to transform it into a system that determines x, and after some manipulation one gets

$$\tau' \frac{\partial x}{\partial t'} = -\frac{p}{J} \sum_{i,j} \frac{\partial x}{\partial \xi_i} \frac{\partial}{\partial \xi_j} (J \nabla \xi_j \cdot G^{-1} \nabla \xi_i),$$

where J is the Jacobian of the mapping $\xi \mapsto x$. In [Hua01] Huang discusses the choices of G and p and other computational considerations, and numerical

results for the Burgers' equation analogue of (4.1) are scrutinized. For a related investigation of criteria for the optimality of constructed meshes see [CSX07].

Returning to [LCHR03], the choice of time discretization should be able to handle the addition and deletion of mesh points associated with the h-refinement. Consequently one-step time integrators are preferred to multi-step methods. Collecting the MMPDE and (4.1) – the first of which is associated with mesh movement, while the second yields the solution u – into a single system, a Galerkin finite element method applied to a weak form of the system then generates the discrete system that must be solved at each time level. Further enhancements of the method are discussed in [LCHR03] and numerical results for Burgers' equation are presented. The same problem is investigated numerically in [BMRS02], whose less technical approach also draws on the methodology of [Hua01] with monitor functions related to the one that we used in Section I.2.5.

Moving finite element methods that are closer in spirit to the method of Section II.5.2 appear in [DL02], where an error bound is proved, and in [Jim96] where Jimack shows that if both the solution of (4.1) and the solution of the moving finite element equations have steady-state solutions, then the steady-state moving finite element solution is a locally best approximation of the true steady-state solution. More recent work by the same author and coworkers (see [BHJJ06] and its references) applies similar methods to scale-invariant problems such as the porous medium equation.

In Section 4.2 we discussed dimension-splitting methods in the context of finite differences. Such methods appear much less often in a finite element framework. The reader can consult [LR95] for some general theory. The solution of a nonlinear variant of (4.1) by means of an ADI finite volume method is discussed in [WZ03].

A semilinear time-dependent reaction-diffusion problem in two space dimensions is considered in [BH07] and convergence results are proved on Shishkin meshes and on the A-meshes of Remark 3.122.

Much remains to be done in the area of numerical methods for time-dependent singularly perturbed problems; there are many unanswered questions. Surveys of previous work can be found in, e.g., [Cod98, EW01].

Part IV

The Incompressible Navier-Stokes Equations

The unsteady incompressible Navier-Stokes equations,

$$\mathbf{u}_t - \nu \triangle \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega \times (0, T],$$
$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \times (0, T],$$
$$\mathbf{u} = \mathbf{u}_b \quad \text{on } \partial\Omega \times (0, T],$$
$$\mathbf{u}(0) = \mathbf{u}_0 \quad \text{in } \Omega,$$

are widely studied as a valuable model in the highly significant research area of Computational Fluid Dynamics (CFD); see, e.g., [Fei93, FŠ04, KL04, Gal94, GR86, GS00a, GS00b, Gun89, Tem83]. In these equations $\nu := 1/\mathrm{Re}$ is the inverse of the Reynolds number, $\mathbf{u} = (u_1, \ldots, u_d)$ is the unknown velocity, p is the pressure field, $\mathbf{f} = (f_1, \ldots, f_d)$ a given body force, $\mathbf{u}_b$ a prescribed velocity field at the boundary, $\mathbf{u}_0$ the velocity field at time $t = 0$, Ω a bounded domain in $\mathbb{R}^d$ (where $d = 2$ or 3) with Lipschitz-continuous boundary, and $(0, T]$ the time interval considered.

Throughout Part IV boldface letters will be used (as above) to denote vector-valued quantities. For notational convenience when discussing spaces like $L_0^2(\Omega)$ we write L^2 instead of L_2, unlike Parts I–III. Furthermore, we sometimes use the notation $\|\cdot\|_{m,p}$ and $|\cdot|_{m,p}$ for the norm and highest-order seminorm in the Sobolev space $W^{m,p}(\Omega)$; thus for example $|\cdot|_{1,2} \equiv |\cdot|_1$ and $\|\cdot\|_{0,2} \equiv \|\cdot\|_0$.

Written out in full, these differential equations are

$$\frac{\partial u_i}{\partial t} - \nu \Delta u_i + \sum_{j=1}^{d} u_j \frac{\partial u_i}{\partial x_j} + \frac{\partial p}{\partial x_i} = f_i \quad \text{in } \Omega \times (0, T], \quad \text{for } i = 1, \ldots, d,$$

$$\sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i} = 0 \quad \text{in } \Omega \times (0, T].$$

The first d equations here model conservation of momentum while the final equation states that mass is conserved.

Implicit time discretizations of the Navier-Stokes equations lead at each time step to the Oseen problem

$$-\nu\triangle\mathbf{u} + (\mathbf{b}\cdot\nabla)\mathbf{u} + \sigma\mathbf{u} + \nabla p = \tilde{\mathbf{f}} \quad \text{in } \Omega,$$
$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega,$$
$$\mathbf{u} = \tilde{\mathbf{u}}_b \quad \text{on } \partial\Omega,$$

where $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{u}}_b$ depend on the solution at the previous time step, $\mathbf{b}$ is a given vector field with $\nabla\cdot\mathbf{b} = 0$, and $\sigma \sim 1/\triangle t$. The Oseen problem also arises when solving the nonlinear stationary Navier-Stokes equations

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega,$$
$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega,$$
$$\mathbf{u} = \mathbf{u}_b \quad \text{on } \partial\Omega$$

by a fixed-point iteration. In this case $\mathbf{b}$ corresponds to the previous iterate $\mathbf{u}^{\text{old}}$ of the velocity field and $\sigma = 0$.

Part IV will show how methods developed for convection-diffusion problems can be applied to this more complex system of equations. Compared with Parts I–III, there are additional difficulties in the numerical solution of the Oseen and Navier-Stokes problems:

- in two space dimensions, the Navier-Stokes equations with Dirichlet boundary data have, on any time interval $[0, T]$, a unique solution that is also a classical solution provided that all data of the problem are smooth enough; but in three dimensions, the existence of such solutions has been proved only for sufficiently small data or on sufficiently short intervals of time.
- to prove uniqueness of any solution of the stationary version of the Navier-Stokes equation, a smallness restriction on the data such as (1.5) below is needed in both two and three dimensions; uniqueness cannot be guaranteed for all positive ν and all data $\mathbf{f}$ and $\mathbf{u}_b$.
- the incompressibility constraint $\nabla\cdot\mathbf{u} = 0$ does not, in general, allow arbitrary approximations of the velocity and pressure fields – the approximation spaces must satisfy an inf-sup condition (the Babuška-Brezzi condition) or an additional "pressure stabilization" (like those considered in Chapters 3 and 4) has to be introduced.
- the nonlinear convection term $(\mathbf{u}\cdot\nabla)\mathbf{u}$ couples different components u_i of the solution.
- the Newton linearization of the momentum equation fails in general to be coercive; moreover, the dependence on ν of the norm of its inverse is *a priori* unknown; see Remark 1.3 below for more details.

These complications make the analysis of numerical methods for the Navier-Stokes equations a formidable task. In particular, some familiar and useful theoretical tools – e.g., the maximum principle – cannot be used. Discretizations of the incompressible Navier-Stokes problem by finite element

methods suffer in general from two main shortcomings. First, the discrete inf-sup (Babuška-Brezzi) condition is violated. Second, spurious oscillations occur because of the predominantly convective nature of the equations. Both these shortcomings are present also in the Oseen problem which is linear and uniquely solvable for all positive ν and all data $\mathbf{f}$ and $\mathbf{u}_b$. It is thus unsurprising that in the research literature, the Oseen problem is seen as a suitable test bed for the development of robust and efficient numerical methods for the incompressible Navier-Stokes equations.

Our investigation in Part IV will confine itself mainly to the Oseen equation (linear) and the stationary Navier-Stokes equations (nonlinear) with the homogeneous Dirichlet boundary condition $\mathbf{u}_b = \mathbf{0}$. For inhomogeneous boundary conditions, see [GP83, Gun96]. As regards the various stabilization techniques of Part III, we shall restrict ourselves to upwind finite element methods, methods of streamline diffusion (SDFEM) type, and local projection stabilization (LPS) methods. For the application of the continuous interior penalty (CIP) method we refer to [BFH06, BH06, Bur07] and the discontinuous Galerkin (dGFEM) approach is considered in [CKS04, CKS05b, CKS05a, CKS07].

For the unsteady Navier-Stokes equations, standard finite element methods are analysed in [HR82, HR86, HR88, HR90] and the survey papers [Ran94, Ran00, Ran04] also discuss stability issues. Applications of the SDFEM in a space-time setting are investigated in [JS86] and [HS90]. The semi-discretization of the unsteady Navier-Stokes equations using the CIP approach is studied in [BF07].

1

Existence and Uniqueness Results

We begin with the stationary incompressible Navier-Stokes equations and derive their weak formulation for the case of homogeneous boundary conditions $\mathbf{u}_b = \mathbf{0}$. Multiplying the momentum equation by a function $\mathbf{v} = (v_1, \ldots, v_d)$, where $\mathbf{v} = \mathbf{0}$ on $\partial\Omega$, then integrating over Ω and integrating the highest-order terms by parts, one obtains

$$\nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + n(\mathbf{u}, \mathbf{u}, \mathbf{v}) - (p, \nabla\cdot\mathbf{v}) = (\mathbf{f}, \mathbf{v}).$$

Here the convective term is

$$n(\mathbf{w}, \mathbf{u}, \mathbf{v}) := \int_\Omega \sum_{i,j=1}^d w_i \frac{\partial u_j}{\partial x_i} v_j \, dx \qquad (1.1)$$

and $(\cdot, \cdot)$ is used to denote the inner product in both $L^2(\Omega)$ and its vector-valued versions. The mass conservation law $\nabla\cdot\mathbf{u} = 0$ is required to hold true in $L^2(\Omega)$, which means that for arbitrary $q \in L^2(\Omega)$ one has

$$(q, \nabla\cdot u) = 0.$$

The pressure p can be determined only up to an additive constant, since $(1, \nabla\cdot\mathbf{v}) = 0$ for all $\mathbf{v}$ that vanish on the boundary $\partial\Omega$; we fix this constant by seeking a pressure p whose mean value is zero, i.e., $(p, 1) = 0$. That is, the pressure p lies in the space

$$L_0^2(\Omega) = \{\, r \in L^2(\Omega) \,:\, (r, 1) = 0 \,\}.$$

Now $L^2(\Omega) = L_0^2(\Omega) \oplus \mathrm{span}\{1\}$. If $\mathbf{v} \in H_0^1(\Omega)^d$, then $(1, \nabla\cdot\mathbf{v}) = 0$. Consequently, given $\mathbf{v} \in H_0^1(\Omega)^d$, the condition $(q, \nabla\cdot\mathbf{v}) = 0$ for all $q \in L_0^2(\Omega)$ is equivalent to the condition $(q, \nabla\cdot\mathbf{v}) = 0$ for all $q \in L^2(\Omega)$.

Setting $\mathbf{V} := H_0^1(\Omega)^d$ and $Q := L_0^2(\Omega)$, a weak formulation is:

Find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + n(\mathbf{u}, \mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \tag{1.2a}$$

$$(q, \nabla \cdot \mathbf{u}) = 0. \tag{1.2b}$$

Sobolev's embedding theorem implies that the space $H_0^1(\Omega)$ is continuously embedded in $L^6(\Omega)$ for $d \leq 4$, so the trilinear form n defined in (1.1) is continuous on $\mathbf{V} \times \mathbf{V} \times \mathbf{V}$. The pressure p can be eliminated. To do this, we introduce the subspace of divergence-free functions

$$\mathbf{W} := \{ \mathbf{v} \in \mathbf{V} \ : \ (q, \nabla \cdot \mathbf{v}) = 0 \text{ for all } q \in Q \}.$$

Equation (1.2b) implies that the component $\mathbf{u}$ of the solution $(\mathbf{u}, p)$ of (1.2) belongs to $\mathbf{W}$. Thus if $(\mathbf{u}, p) \in \mathbf{V} \times Q$ is a solution of (1.2), then $\mathbf{u}$ must also be a solution of the following simpler problem:

Find $\mathbf{u} \in \mathbf{W}$ such that for all $\mathbf{v} \in \mathbf{W}$ one has

$$\nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + n(\mathbf{u}, \mathbf{u}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}). \tag{1.3}$$

Conversely, if $\mathbf{u} \in \mathbf{W}$ is a solution of (1.3), then there exists a unique $p \in Q$ such that $(\mathbf{u}, p)$ is a solution of (1.2). Given $\mathbf{u}$, this function p is a solution of the following problem:

Find $p \in Q$ such that for all $\mathbf{v} \in \mathbf{W}^\perp$ one has

$$(p, \nabla \cdot \mathbf{v}) = \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + n(\mathbf{u}, \mathbf{u}, \mathbf{v}) - (\mathbf{f}, \mathbf{v}).$$

The *Babuška-Brezzi condition*

$$\inf_{q \in Q} \sup_{\mathbf{v} \in \mathbf{V}} \frac{(q, \nabla \cdot \mathbf{v})}{\|q\|_0 \, |\mathbf{v}|_1} \geq \beta > 0 \tag{1.4}$$

guarantees that this problem is well posed [GR86, Lemma 4.1, p. 58]. Denote the norm in the dual space $\mathbf{W}^*$ of $\mathbf{W}$ by $\| \cdot \|_*$ and denote by γ the continuity constant of the trilinear form $n : \mathbf{W} \times \mathbf{W} \times \mathbf{W} \to \mathbb{R}$; that is,

$$\|\mathbf{f}\|_* := \sup_{\mathbf{v} \in \mathbf{W}} \frac{(\mathbf{f}, \mathbf{v})}{|\mathbf{v}|_1} \quad \text{and} \quad \gamma := \sup_{\mathbf{w}, \mathbf{u}, \mathbf{v} \in \mathbf{W}} \frac{n(\mathbf{w}, \mathbf{u}, \mathbf{v})}{|\mathbf{w}|_1 \, |\mathbf{u}|_1 \, |\mathbf{v}|_1}.$$

Then we have the following existence and uniqueness result:

Theorem 1.1. *[GR86] Given a continuous linear form $\mathbf{f} : \mathbf{V} \to \mathbb{R}$, there is at least one solution $(\mathbf{u}, p)$ of (1.2). Moreover, if in addition*

$$(\gamma/\nu^2)\|\mathbf{f}\|_* < 1, \tag{1.5}$$

then this solution is unique.

To solve the nonlinear problem (1.2), a simple iterative technique can be applied:

Choose $\mathbf{u}^0 \in \mathbf{V}$. For each $k \geq 0$, given $\mathbf{u}^k \in \mathbf{V}$, find $(\mathbf{u}^{k+1}, p^{k+1}) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\nu(\nabla \mathbf{u}^{k+1}, \nabla \mathbf{v}) + n(\mathbf{u}^k, \mathbf{u}^{k+1}, \mathbf{v}) - (p^{k+1}, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \tag{1.6a}$$

$$(q, \nabla \cdot \mathbf{u}^{k+1}) = 0. \tag{1.6b}$$

Theorem 1.2. *If* $\mathbf{f} : \mathbf{V} \to \mathbb{R}$ *is a continuous linear form and* $\mathbf{u}^k \in \mathbf{V}$, *then the linear problem* (1.6) *has a solution and this solution is unique. Furthermore, if* (1.5) *is satisfied, then the sequence* $(\mathbf{u}^k, \mathbf{p}^k)$ *converges in* $\mathbf{V} \times Q$ *to the unique solution* $(\mathbf{u}, p)$ *of* (1.2) *as* $k \to \infty$.

Proof. First, $\mathbf{u}^{k+1} \in \mathbf{W}$ by (1.6b). Now we can eliminate the pressure from (1.6a) by restricting the test functions $\mathbf{v}$ to the subspace $\mathbf{W}$. This yields the linear convection-diffusion problem:

Find $\mathbf{u}^{k+1} \in \mathbf{W}$ such that for all $\mathbf{v} \in \mathbf{W}$ one has

$$\nu(\nabla \mathbf{u}^{k+1}, \nabla \mathbf{v}) + n(\mathbf{u}^k, \mathbf{u}^{k+1}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}). \tag{1.7}$$

Given a solution $\mathbf{u}^{k+1} \in \mathbf{W}$ of (1.7), the inf-sup condition (1.4) implies that there is a unique $p^{k+1} \in Q$ such that $(\mathbf{u}^{k+1}, p^{k+1})$ is a solution of (1.6). Thus (1.7) is equivalent to (1.6).

The Lax-Milgram Lemma implies that (1.7) has a unique solution $\mathbf{u}^{k+1}$. For as $\mathbf{u}^k \in \mathbf{W}$, for all $\mathbf{v} \in \mathbf{W}$ we have

$$n(\mathbf{u}^k, \mathbf{v}, \mathbf{v}) = \frac{1}{2} \int_\Omega \sum_{i,j=1}^2 u_i^k \frac{\partial v_j^2}{\partial x_i} \, dx = -\frac{1}{2} \int_\Omega \mathbf{v}^2 \, \nabla \cdot \mathbf{u}^k \, dx = 0,$$

which implies that the corresponding bilinear form $\nu(\nabla \cdot, \nabla \cdot) + n(\mathbf{u}^k, \cdot, \cdot)$ is coercive.

Defining the solution operator $P : \mathbf{W} \to \mathbf{W}$ by $\mathbf{u}^{k+1} = P\mathbf{u}^k$, one can show that P maps the closed ball $B(0, \|\mathbf{f}\|_*/\nu)$ into itself and is contractive with contraction factor $(\gamma/\nu^2)\|\mathbf{f}\|_*$ (see [GR79] for more details). Then an appeal to Banach's fixed-point theorem completes the argument. $\square$

Remark 1.3. The above theorems prove that for $\nu \geq \nu_0 > 0$, solutions exist and are unique. To analyse finite element methods when ν is arbitrarily small, the theory of nonsingular solution branches can be applied; see Section 3.2. The basic assumption of this theory is the well-posedness of the linear problem

Find $\mathbf{u} \in \mathbf{W}$ such that for all $\mathbf{v} \in \mathbf{W}$

$$B(\mathbf{u}, \mathbf{v}) := \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + n(\mathbf{u}^*, \mathbf{u}, \mathbf{v}) + n(\mathbf{u}, \mathbf{u}^*, \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \tag{1.8}$$

which is the Newton linearization of (1.3) at a branch of solutions $\mathbf{u}^* = \mathbf{u}^*(\nu)$. Unfortunately the coercivity of the bilinear form B cannot be guaranteed *a priori* since we do not know if $n(\mathbf{v}, \mathbf{u}^*, \mathbf{v}) > 0$. Consequently, it is difficult in practice to check the assumption that a branch of nonsingular solutions exists. Moreover, this technique draws on estimates of the Lipschitz constant of the solution operator of (1.8), whose dependence on ν is unknown *a priori*. Indeed, in special cases these constants can blow up exponentially – see the one-dimensional Example 3.16 in Section 3.2. These difficulties are the main obstacles to the construction of superior discretization methods for the Navier-Stokes equations in the significant regime when the Reynolds number Re = $1/\nu$ is large. ♣

Remark 1.4. From the above results where uniqueness of solutions fails only when ν is small, one might hope that standard finite element methods are adequate for moderate values of the Reynolds number, but numerical experiments show that even in this case some stabilization technique such as upwinding, streamline diffusion or local projection is still needed in order to get satisfactory results.   ♣

Finally, we consider the solvability of the weak formulation of the Oseen problem with a nonnegative constant σ, homogeneous boundary conditions $\mathbf{u}_b = \mathbf{0}$, and a given vector field $\mathbf{b}$ for which $\nabla \cdot \mathbf{b} = 0$. The weak formulation is:

Find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + ((\mathbf{b} \cdot \nabla)\mathbf{u}, \mathbf{v}) + \sigma(\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) = (\mathbf{f}, \mathbf{v}), \qquad (1.9a)$$

$$(q, \nabla \cdot \mathbf{u}) = 0. \qquad (1.9b)$$

Theorem 1.5. *Given a continuous linear form* $\mathbf{f} : \mathbf{V} \to \mathbb{R}$*, there is a unique solution* $(\mathbf{u}, p)$ *of* (1.9) *for each* $\nu > 0$.

Proof. First, eliminate the pressure by seeking a solution of the problem

Find $\mathbf{u} \in \mathbf{W}$ such that for all $\mathbf{v} \in \mathbf{W}$ one has

$$\nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + ((\mathbf{b} \cdot \nabla)\mathbf{u}, \mathbf{v}) + \sigma(\mathbf{u}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}). \qquad (1.10)$$

Here the Lax-Milgram lemma implies existence of a unique solution because of the coercivity that follows from

$$\nu(\nabla\mathbf{v}, \nabla\mathbf{v}) + ((\mathbf{b} \cdot \nabla)\mathbf{v}, \mathbf{v}) + \sigma(\mathbf{v}, \mathbf{v}) = \nu|\mathbf{v}|_1^2 + \sigma\|\mathbf{v}\|_0^2.$$

Then the reconstruction of a pressure $p \in Q$ such that the pair $(\mathbf{u}, p) \in \mathbf{V} \times Q$ satisfies (1.9) relies on the inf-sup condition (1.4).   □

2

Upwind Finite Element Method

In this chapter we consider the stationary Navier-Stokes problem. Any conforming finite element method based on the weak formulation (1.3) would require the construction of divergence-free trial functions, which is often difficult. Thus our starting point for the numerical solution of the stationary incompressible Navier-Stokes equations will be the weak formulation (1.2), which is formulated in terms of the primitive variables $\mathbf{u}$ and p. For ease of notation only the two-dimensional case is presented here but our results can be extended to problems in three dimensions. Assume that $\Omega \subset \mathbb{R}^2$ is a bounded convex polygonal domain. The solution spaces $\mathbf{V}$ and Q will be approximated by discrete spaces $\mathbf{V}_h$ and Q_h respectively. Then the standard Galerkin discretization is:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has

$$\nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + n(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h),$$
$$(q_h, \nabla \cdot \mathbf{u}_h) = 0.$$

We assume that the reader is familiar with the standard discretization techniques [GR86] used for the Navier-Stokes equations when the Reynolds number is small (i.e., when convection does not dominate). For this analysis one needs the *discrete Babuška-Brezzi condition* (cf. (1.4))

$$\inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{(q_h, \nabla \cdot \mathbf{v}_h)}{\|q_h\|_0 \, |\mathbf{v}_h|_1} \geq \beta > 0 \qquad (2.1)$$

where the positive constant β is independent of h. If the method is nonconforming (i.e., when $\mathbf{V}_h \not\subset \mathbf{V}$), the integrals in (2.1) will be computed element by element. Pairs of suitable finite element spaces $(\mathbf{V}_h, Q_h)$ can be found, e.g., in [BF91, GR86, GS00a, GS00b].

We shall use the nonconforming (P_1, P_0) element pair of Crouzeix and Raviart [CR73] which yields a relatively small number of unknowns since the shape functions are of low order. It satisfies the Babuška-Brezzi stability

condition (2.1) – modified by computing the integrals element by element – where β is independent not only of h but also of all mesh parameters, including the minimum angle in any mesh triangle [Dor95]. When the Reynolds number is large (i.e., when the convective term dominates), numerical stability will be ensured through the special upwind discretization of the convection term that was introduced in Section III.3.1; see (III.3.20c) and Figure 3.4.

Let $\mathcal{T}_h$ be a shape-regular decomposition of the domain Ω into triangles T. Denote the edges of the triangles $T \in \mathcal{T}_h$ by $\{\Gamma_i : i = 1, \ldots, N + M\}$, where the Γ_i are inner edges for $i = 1, \ldots, N$ and boundary edges (i.e., $\Gamma_i \subset \partial\Omega$) for $i = N + 1, \ldots, N + M$. For each i let B_i be the midpoint of the edge Γ_i.

The discrete spaces $\mathbf{V}_h$ and Q_h, which approximate the solution spaces $\mathbf{V}$ and Q in (1.2), are defined by

$$\mathbf{V}_h := \big\{\mathbf{v}_h \: : \: \mathbf{v}_h\big|_T \in P_1^2(T) \text{ for all } T, \ \mathbf{v}_h \text{ continuous at } B_i \text{ for } i = 1, \ldots, N,$$
$$\mathbf{v}_h(B_i) = \mathbf{0} \text{ for } i = N + 1, \ldots, N + M\big\},$$
$$Q_h := \big\{q_h \in Q \: : \: q_h\big|_T \in P_0(T) \quad \text{for all } T\big\}.$$

A typical vector-valued function $\mathbf{v}_h \in \mathbf{V}_h$ will be discontinuous on the edges Γ_i of the elements T, so $\mathbf{V}_h \not\subset \mathbf{V}$. The method is therefore nonconforming and one must extend to $\mathbf{V}_h$ the definitions of the bilinear and trilinear forms that appear in the weak formulation (1.2). This is done in a natural way by calculating the integrals element by element: for each $T \in \mathcal{T}_h$, if the functions $\mathbf{u}$ and $\mathbf{v}$ lie in $H^1(T)^2$ and q lies in $L^2(T)$, then set

$$(\nabla\mathbf{u}, \nabla\mathbf{v})_h := \sum_{T \in \mathcal{T}_h} \int_T \nabla\mathbf{u} \cdot \nabla\mathbf{v} \, dx \quad \text{and} \quad (q, \nabla \cdot \mathbf{v})_h := \sum_{T \in \mathcal{T}_h} \int_T q\nabla \cdot \mathbf{v} \, dx.$$

For the discretization of the trilinear form $n(\cdot, \cdot, \cdot)$ we do not use the element-by-element formula

$$\tilde{n}_h(\mathbf{z}, \mathbf{u}, \mathbf{v}) := \sum_{T \in \mathcal{T}_h} \int_T (\mathbf{z}\cdot\nabla)\mathbf{u}\,\mathbf{v} \, dx, \tag{2.2}$$

but instead apply the upwind technique of [OU84, ST96] that was described in Section III.3.1. This approach makes a secondary decomposition of the domain Ω into so-called lumping regions R_l where each R_l is associated with the edge Γ_l; see Figure III.3.4. Let C_T be the barycentre of the element $T \in \mathcal{T}_h$. Let $S_{T,l}$ be the triangle contained in T that has Γ_l as one of its edges and C_T as its other vertex. As in Part III, denote by Λ_l the set of all indices $k \neq l$ for which the nodes B_k and B_l belong to a common $T \in \mathcal{T}_h$; furthermore, in this case we define $\Gamma_{lk} := \partial S_{T,l} \cap \partial S_{T,k}$ to be the common edge of the triangles $S_{T,l}$ and $S_{T,k}$. For each inner edge Γ_l, the triangulation $\mathcal{T}_h$ contains two elements T and T' with $\Gamma_l = \partial T \cap \partial T'$, and we define the lumping region to be $R_l := S_{T,l} \cup S_{T',l}$. In the case of a boundary edge $\Gamma_l \subset \partial T \subset \partial\Omega$, one sets $R_l := S_{T,l}$. Define a lumping operator L_h, which maps a given function $\mathbf{v} \in \mathbf{V}_h$ into a piecewise constant function on $\cup_l R_l$, by

$$(L_h \mathbf{v})(x) := \mathbf{v}(B_l) \quad \text{if } x \in R_l.$$

To derive the upwind discretization $n_h(\cdot, \cdot, \cdot)$ of the convective term, one introduces the lumping operator L_h then appeals to Green's theorem to rewrite $n(\cdot, \cdot, \cdot)$ in terms of fluxes over the boundaries Γ_{lk} of the lumping regions R_l:

$$
\begin{aligned}
n(\mathbf{z}, \mathbf{u}, \mathbf{v}) &= \sum_{l=1}^{N+M} ((\mathbf{z} \cdot \nabla) \mathbf{u}, \mathbf{v})_{R_l} \\
&= \sum_{l=1}^{N+M} [(\nabla \cdot (\mathbf{z} \otimes \mathbf{u}), \mathbf{v})_{R_l} - (\nabla \cdot \mathbf{z}, \mathbf{u} \cdot \mathbf{v})_{R_l}] \\
&\approx \sum_{l=1}^{N+M} [(\nabla \cdot (\mathbf{z} \otimes \mathbf{u}), L_h \mathbf{v})_{R_l} - (\nabla \cdot \mathbf{z}, L_h \mathbf{u} \cdot L_h \mathbf{v})_{R_l}] \\
&= \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \mathbf{v}(B_l) \cdot \int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) (\mathbf{u} - \mathbf{u}(B_l)) \, d\gamma.
\end{aligned}
$$

Here $\mathbf{n}_{lk}$ is the unit normal on Γ_{lk} that points away from R_l and the notation $\nabla \cdot (\mathbf{z} \otimes \mathbf{u})$ means a vector whose i^{th} component is

$$(\nabla \cdot (\mathbf{z} \otimes \mathbf{u}))_i := \nabla \cdot (\mathbf{z} u_i) = \sum_{j=1}^{2} \frac{\partial (z_j u_i)}{\partial x_j}.$$

As in Part III, upwinding is achieved by replacing $\mathbf{u}$ on Γ_{lk} by a fixed upwind value $\mathbf{u}^{\text{upw}}$, i.e.,

$$\mathbf{u} \approx \mathbf{u}^{\text{upw}} := \lambda_{lk}(\mathbf{z}) \mathbf{u}(B_l) + (1 - \lambda_{lk}(\mathbf{z})) \, \mathbf{u}(B_k)$$

for some $\lambda_{lk}(\mathbf{z})$. This yields the discretization

$$
\begin{aligned}
n_h(\mathbf{z}, \mathbf{u}, \mathbf{v}) := \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \left[\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \right] (1 - \lambda_{lk}(\mathbf{z})) \\
[\, \mathbf{u}(B_k) - \mathbf{u}(B_l) \,] \cdot \mathbf{v}(B_l). \qquad (2.3)
\end{aligned}
$$

Our discrete Navier-Stokes problem is now:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has

$$\nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h)_h + n_h(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h)_h = (\mathbf{f}, \mathbf{v}_h), \qquad (2.4a)$$

$$(q_h, \nabla \cdot \mathbf{u}_h)_h = 0. \qquad (2.4b)$$

The function $\lambda_{lk}(\cdot)$ is defined as in Section III.3.1 by

$$\lambda_{lk}(\mathbf{z}) = \Phi\left(\frac{\beta_{lk}}{2\nu} \right) \quad \text{with} \quad \beta_{lk} := \int_{\Gamma_{lk}} \mathbf{z} \cdot \mathbf{n}_{lk} \, d\gamma.$$

The *weighting function* $\Phi(\cdot)$ must satisfy the following assumptions (see Section III.3.1):

(B1) $\Phi(t) = 1 - \Phi(-t)$ $\quad \forall\, t > 0$ $\quad$ and $\quad 0 \leq \Phi(t) \leq 1$ $\quad \forall\, t \in \mathbb{R}$,

(B2) $t \left[\Phi(t) - \dfrac{1}{2}\right] \geq 0$ $\quad \forall\, t \in \mathbb{R}$,

(B3) $g(t) := t\,\Phi(t)$ $\quad$ is Lipschitz continuous on $\mathbb{R}$.

Choices for $\Phi(\cdot)$ that have been used in practical computations [GRS90, RS89, Tur91, Tur99] are

$$\Phi_1(t) = \begin{cases} 1 & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \quad \text{and} \quad \Phi_2(t) = \begin{cases} (1+2t)/(2+2t) & \text{if } t \geq 0, \\ 1/(2-2t) & \text{if } t < 0, \end{cases}$$

where $\Phi_1(\cdot)$ corresponds to *simple* and $\Phi_2(\cdot)$ to *Samarskiĭ upwinding* (see Section I.2.1.2). Both functions satisfy assumptions (B1)–(B3).

We now study the convergence properties of the method (2.4). Assume that the triangulation $\mathcal{T}_h$ of the domain Ω into elements $T \in \mathcal{T}_h$ is shape-regular. Define the discrete H^1 norm $\|\cdot\|_h$ on $\mathbf{V} + \mathbf{V}_h$ by

$$\|\mathbf{v}\|_h := (\nabla \mathbf{v}, \nabla \mathbf{v})_h^{1/2}.$$

Finally, let

$$\mathbf{W}_h := \{\mathbf{v}_h \in \mathbf{V}_h \mid (q_h, \nabla \cdot \mathbf{v}_h)_h = 0 \quad \forall\, q_h \in Q_h\}$$

be the subspace of discretely divergence-free functions.

Let us take from [ST89, TT89] two lemmas and two theorems that deal with the existence of solutions and the convergence of our upwind discretization (2.4) for arbitrary Reynolds number $\mathrm{Re} = 1/\nu$. The first lemma is vital in obtaining a stability estimate for the discrete solution $\mathbf{u}_h$.

Lemma 2.1. *Assume that (B1) and (B2) hold true. Then we have*

$$n_h(\mathbf{z}_h, \mathbf{v}_h, \mathbf{v}_h) \geq 0 \quad \text{for all } \mathbf{z}_h \in \mathbf{W}_h \text{ and all } \mathbf{v}_h \in \mathbf{V}_h.$$

Proof. To simplify the notation the subscript h is omitted from $\mathbf{z}_h$ and $\mathbf{v}_h$ during the proof. The definition of n_h in (2.3) implies that

$$
\begin{aligned}
n_h(\mathbf{z}, \mathbf{v}, \mathbf{v}) = {} & \frac{1}{2} \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \left[\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk})\, d\gamma \right] (1 - \lambda_{lk}(\mathbf{z})) \\
& \qquad [\mathbf{v}(B_k) - \mathbf{v}(B_l)] \cdot \mathbf{v}(B_l) \\
& + \frac{1}{2} \sum_{k=1}^{N+M} \sum_{l \in \Lambda_k} \left[\int_{\Gamma_{kl}} (\mathbf{z} \cdot \mathbf{n}_{kl})\, d\gamma \right] (1 - \lambda_{kl}(\mathbf{z})) \\
& \qquad [\mathbf{v}(B_l) - \mathbf{v}(B_k)] \cdot \mathbf{v}(B_k).
\end{aligned}
$$

Recalling that $\Gamma_{kl} = \Gamma_{lk}$ and $\mathbf{n}_{kl} = -\mathbf{n}_{lk}$, (B1) implies that $1 - \lambda_{kl} = \lambda_{lk}$. Since both sums are over the same pairs of indices, we obtain

$$
n_h(\mathbf{z}, \mathbf{v}, \mathbf{v}) = \frac{1}{2} \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \left[\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \right]
$$

$$
[(1 - \lambda_{lk}(\mathbf{z}))\mathbf{v}(B_l) + \lambda_{lk}(\mathbf{z})\mathbf{v}(B_k)] \cdot [\mathbf{v}(B_k) - \mathbf{v}(B_l)]
$$

$$
= \frac{1}{2} \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \left[\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \right] \left(\lambda_{lk}(\mathbf{z}) - \frac{1}{2} \right) [\mathbf{v}(B_k) - \mathbf{v}(B_l)]^2
$$

$$
+ \frac{1}{4} \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \left[\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \right] \mathbf{v}^2(B_k)
$$

$$
- \frac{1}{4} \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \left[\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \right] \mathbf{v}^2(B_l)
$$

where the square of a vector means the inner product of the vector with itself. Now (B2) implies that the first sum is nonnegative. In the second sum, set $\mathbf{n}_{lk} = -\mathbf{n}_{kl}$ and $\Gamma_{lk} = \Gamma_{kl}$ then change the order of summation; on swopping the indices (k, l) for (l, k) one now sees that the second and third terms are identical. Thus

$$
n_h(\mathbf{z}, \mathbf{v}, \mathbf{v}) \geq -\frac{1}{2} \sum_{l=1}^{N+M} \left\{ \sum_{k \in \Lambda_l} \int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \right\} \mathbf{v}^2(B_l).
$$

But if T and T' are the two triangles in $\mathcal{T}_h$ that share the common node B_l, then

$$
\sum_{k \in \Lambda_l} \int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma = \int_{S_{T,l}} \nabla \cdot \mathbf{z} \, dx + \int_{S_{T',l}} \nabla \cdot \mathbf{z} \, dx
$$

$$
+ \int_{\Gamma_l} \left(\mathbf{z}|_{T'} - \mathbf{z}|_T \right) \cdot \mathbf{n}_{T,l} \, d\gamma,
$$

where $\mathbf{n}_{T,l}$ is the unit normal on Γ_l that points out from T. The first two summands here are zero since $\mathbf{z}$ is discretely divergence-free, while the last vanishes because $\mathbf{z}$ is piecewise linear and continuous at the midpoint B_l. Consequently, the inner sum

$$
\sum_{k \in \Lambda_l} \int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma
$$

in the lower bound on $n_h(\mathbf{z}, \mathbf{v}, \mathbf{v})$ is zero.   □

Remark 2.2. A careful inspection of the proof of Lemma 2.1 shows that we have the nonnegative lower bound

$$n_h(\mathbf{z}, \mathbf{v}, \mathbf{v}) \geq \frac{1}{2} \sum_{l=1}^{N+M} \sum_{k \in \Lambda_l} \int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma \left(\lambda_{lk}(\mathbf{z}) - \frac{1}{2} \right) [\mathbf{v}(B_k) - \mathbf{v}(B_l)]^2$$

for all $\mathbf{z} \in \mathbf{W}_h$ and all $\mathbf{v} \in \mathbf{V}_h$. In fact the right-hand side can be zero only if for each pair (l, k) either $\int_{\Gamma_{lk}} (\mathbf{z} \cdot \mathbf{n}_{lk}) \, d\gamma = 0$ (i.e., the flux vanishes) or $\lambda_{lk} = 1/2$ (i.e., there is no upwinding). In the standard Galerkin method one would instead work with the trilinear form

$$n_h^{\mathrm{GAL}}(\mathbf{z}, \mathbf{u}, \mathbf{v}) := \frac{1}{2} \sum_{T \in \mathcal{T}_h} \int_T ((\mathbf{z} \cdot \nabla)\mathbf{u}\,\mathbf{v} - (\mathbf{z} \cdot \nabla)\mathbf{v}\,\mathbf{u}) \, dx$$

for which

$$n_h^{\mathrm{GAL}}(\mathbf{z}, \mathbf{v}, \mathbf{v}) = 0 \qquad \text{for all } \mathbf{z} \in \mathbf{W}_h \text{ and all } \mathbf{v} \in \mathbf{V}_h,$$

which explains the improved stability properties of the upwind finite element method. ♣

The next lemma shows that $n_h(\cdot, \cdot, \cdot)$ is continuous on $\mathbf{V}_h \times \mathbf{V}_h \times \mathbf{V}_h$ and will be used to estimate the consistency error.

Lemma 2.3. *Assume that (B1) and (B3) hold true. Then n_h (which is linear in its second and third arguments) is continuous, i.e.,*

$$|\, n_h(\mathbf{z}_h^1, \mathbf{u}_h, \mathbf{v}_h) - n_h(\mathbf{z}_h^2, \mathbf{u}_h, \mathbf{v}_h) \,| \ \leq \ C \, \|\mathbf{z}_h^1 - \mathbf{z}_h^2\|_h \, \|\mathbf{u}_h\|_h \, \|\mathbf{v}_h\|_h$$

for all $\mathbf{z}_h^1, \mathbf{z}_h^2, \mathbf{u}_h, \mathbf{v}_h \in \mathbf{V}_h$, where C is independent of h and ν.

Proof. The proof is rather technical and uses the discrete version of Sobolev's embedding inequality (see [Dor95, HR82]) that gives

$$\|\mathbf{z}_h\|_{0,s} \leq C \, \|\mathbf{z}_h\|_h \quad \text{for all } \mathbf{z}_h \in \mathbf{V}_h \quad \text{and} \quad 1 \leq s \leq 6, \qquad (2.5)$$

where C depends on s and Ω. More details can be found in [ST89, ST96]. □

The solvability of the discrete Navier-Stokes problem (2.4) is guaranteed by the next result.

Theorem 2.4. *Assume that (B1),(B2) and (B3) hold true. Let $f \in (L^2(\Omega))^2$. Then the discrete problem (2.4) has at least one solution $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$.*

Proof. See [ST89]. □

The first-order derivatives of a function $\mathbf{v}_h \in \mathbf{V}_h$ do not in general lie in $(L^2(\Omega))^2$, so to formulate the next convergence result we introduce the embedding operator $E_h : \mathbf{V}_h \to (L^2(\Omega))^6$ defined by

$$(E_h \mathbf{v})(x) := (\mathbf{v}(x), \nabla \mathbf{v}(x)) \qquad \text{for all } x \notin \bigcup_{T \in \mathcal{T}_h} \partial T.$$

Theorem 2.5. *Let the assumptions of Theorem 2.4 be satisfied. Let $\{(\mathbf{u}_h, p_h)\}$ be a sequence of solutions of the discrete problem (2.4), where h tends to zero through a discrete set of values. Then there exists a subsequence $\{(\mathbf{u}_{h'}, p_{h'})\}$ and a solution $(\mathbf{u}, p) \in \mathbf{V} \times Q$ of the continuous problem (1.2) such that $\{E_{h'}\mathbf{u}_{h'}\}$ converges to $(\mathbf{u}, \nabla\mathbf{u})$ in $(L^2(\Omega))^6$ and $\{p_{h'}\}$ converges to p weakly in $L^2(\Omega)$ as $h' \to 0$. If $(\mathbf{u}, p)$ belongs to $(H^2(\Omega))^2 \times H^1(\Omega)$, then the subsequence $\{p_{h'}\}$ converges to p strongly in $L^2(\Omega)$.*

Proof. See [ST89].   □

To quantify the rate of convergence, one must estimate how well n_h (defined by (2.3)) approximates the more straightforward discretization $\tilde{n}_h$ of the convective part described in (2.2). For this we need a suitable interpolation operator $i_h : \mathbf{V} \to \mathbf{V}_h$ that satisfies $i_h\mathbf{u} \in \mathbf{W}_h$ whenever $\mathbf{u} \in \mathbf{V}$ is a divergence-free function. Such an operator can be constructed as follows. Given $\mathbf{u} \in \mathbf{V}$, define the values of $i_h\mathbf{u} \in \mathbf{V}_h$ at the nodes B_l, for $l = 1, \dots, N + M$, by

$$i_h\mathbf{u}(B_l) = \frac{1}{\text{meas}(\Gamma_l)} \int_{\Gamma_l} \mathbf{u}\, ds. \tag{2.6}$$

This determines the interpolant $i_h\mathbf{u} \in \mathbf{V}_h$ uniquely. The interpolant is stable in the broken H^1 norm, i.e.,

$$\|i_h\mathbf{u}\|_h = \left(\sum_{T \in \mathcal{T}_h} \|i_h\mathbf{u}\|_{1,T} \right)^{1/2} \leq C\|\mathbf{u}\|_1 \quad \forall \mathbf{u} \in \mathbf{V},$$

and it satisfies the usual interpolation error estimates.

The next lemma is crucial for the proof of our error estimate. It says that the consistency error introduced by using the upwind discretization n_h of the convective term instead of the Galerkin discretization $\tilde{n}_h$ is of order $\mathcal{O}(h)$.

Lemma 2.6. *Assume that $\mathbf{u} \in (H^2(\Omega))^2$ with $\nabla \cdot \mathbf{u} = 0$. Let $i_h\mathbf{u} \in \mathbf{W}_h$ be the interpolant to $\mathbf{u}$ from $\mathbf{W}_h$. Then for all $\mathbf{w}_h \in \mathbf{W}_h$ one has*

$$|\tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - n_h(i_h\mathbf{u}, i_h\mathbf{u}, \mathbf{w}_h)| \leq C\, h\, \|\mathbf{u}\|_2^2 \|\mathbf{w}_h\|_h. \tag{2.7}$$

Proof. For the details of this technical proof see [ST96].   □

Now we come to the main result in this section. It gives an optimal-order error estimate for our nonconforming (P_1, P_0) finite element approximation.

Unlike the results above, our analysis now needs the assumption that ν is bounded away from zero. For if ν tends to zero, then uniqueness of the solution of (1.2) is no longer guaranteed by Theorem 1.1, which forces us to perform a local analysis. But a basic assumption for local analysis is the existence of a branch of nonsingular solutions – which implies that the linearized problem (1.8) is stable. The authors know of no sufficient conditions that, for arbitrary $\nu > 0$, guarantee that the linearized Navier-Stokes equations (1.8) are stable

and provide concrete estimates of the dependence of the solution $\mathbf{u}$ of (1.8) on ν and $\mathbf{f}$. We therefore restrict our investigation at this stage to the case where uniqueness can be proved and for a local analysis refer the reader to the streamline diffusion method of Section 3.2.

Theorem 2.7. *Assume that (B1), (B2) and (B3) hold true. Let* $\mathbf{f} \in (\mathrm{L}^2(\Omega))^2$. *Suppose that* $\nu > \nu_0$, *where* $\nu_0 = \nu_0(\Omega, \mathbf{f}) > 0$ *is sufficiently large. Then the continuous problem* (1.2) *and the discrete problem* (2.4) *have unique solutions* $(\mathbf{u}, p)$ *and* $(\mathbf{u}_h, p_h)$ *respectively. Under the additional regularity assumption that* $(\mathbf{u}, p) \in (H^2(\Omega))^2 \times H^1(\Omega)$, *one obtains the error estimates*

$$\|\mathbf{u} - \mathbf{u}_h\|_h \leq \frac{C}{\nu_0} Mh, \qquad \|p - p_h\|_0 \leq \frac{C}{\nu_0^2} Mh,$$

where $M = M(\|\mathbf{u}\|_2, \|p\|_1)$ *is a polynomial of degree two whose coefficients are independent of* h *and* ν_0.

Proof. By Theorem 1.1 the continuous problem (1.2) has a unique solution.

We shall show that the discrete problem (2.4) also has a unique solution provided that ν is sufficiently large. Theorem 2.4 ensures existence of a solution to the discrete problem. Take $\mathbf{v}_h = \mathbf{u}_h$ and $q_h = p_h$ in (2.4) and sum, then apply Lemma 2.1 and the estimate (2.5) with $s = 2$. Hence one can see that any solution of (2.4) must satisfy the *a priori* bound

$$\|\mathbf{u}_h\|_h \leq \frac{C}{\nu} \|\mathbf{f}\|_0 \,.$$

Now suppose that the discrete problem has two different solutions, $(\mathbf{u}_h^1, p_h^1)$ and $(\mathbf{u}_h^2, p_h^2)$, say. Setting $\mathbf{v}_h = \mathbf{u}_h^1 - \mathbf{u}_h^2$ and $q_h = p_h^1 - p_h^2$ in (2.4) yields

$$0 = \nu \|\mathbf{v}_h\|_h^2 + n_h(\mathbf{u}_h^1, \mathbf{u}_h^1, \mathbf{v}_h) - n_h(\mathbf{u}_h^2, \mathbf{u}_h^2, \mathbf{v}_h)$$
$$\geq \nu \|\mathbf{v}_h\|_h^2 + n_h(\mathbf{u}_h^1, \mathbf{u}_h^2, \mathbf{v}_h) - n_h(\mathbf{u}_h^2, \mathbf{u}_h^2, \mathbf{v}_h) \,,$$

by Lemma 2.1. Recalling the continuity of n_h (Lemma 2.3) and the *a priori* estimate for $\mathbf{u}_h^2$, one infers that

$$\left(\nu - \frac{C}{\nu} \|\mathbf{f}\|_0\right) \|\mathbf{v}_h\|_h^2 \leq 0.$$

Hence, for sufficiently large $\nu > \nu_0(\Omega, \mathbf{f})$, it follows that $\|\mathbf{v}_h\|_h = 0$, i.e., $\mathbf{u}_h^1 = \mathbf{u}_h^2$. This implies that

$$(p_h^1 - p_h^2, \nabla \cdot \mathbf{w}_h)_h = 0 \quad \forall \mathbf{w}_h \in \mathbf{V}_h.$$

The uniqueness of the pressure now follows directly from the discrete Babuška-Brezzi condition (2.1), which is satisfied for our nonconforming (P_1, P_0) element pair (see, e.g., [Dor95]).

Under the regularity assumption $(\mathbf{u}, p) \in (H^2(\Omega))^2 \times H^1(\Omega)$, the exact solution satisfies the equations

$$\nu\,(\nabla\mathbf{u}, \nabla\mathbf{w}_h)_h + \tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - (p, \nabla\cdot\mathbf{w}_h)_h = (\mathbf{f}, \mathbf{w}_h) + \ell_h(\mathbf{w}_h), \qquad (2.8a)$$
$$(q_h, \nabla\cdot\mathbf{u})_h = 0, \qquad (2.8b)$$

for all $\mathbf{w}_h \in \mathbf{V}_h$ and all $q_h \in Q_h$, where

$$\ell_h(\mathbf{w}_h) = \sum_{T \in \mathcal{T}_h} \left(\nu \int_{\partial T} \frac{\partial\mathbf{u}}{\partial n} \mathbf{w}_h \, ds - \int_{\partial T} p\,(\mathbf{w}_h\cdot\mathbf{n})\, ds \right).$$

Choosing $\mathbf{w}_h = \mathbf{u}_h - \mathbf{v}_h$, where $\mathbf{v}_h \in \mathbf{W}_h$ is arbitrary, (2.4) and (2.8) together yield the identity

$$\begin{aligned}
\nu\,\|\mathbf{w}_h\|_h^2 &= \nu\,(\nabla(\mathbf{u} - \mathbf{v}_h), \nabla\mathbf{w}_h)_h + \tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - n_h(\mathbf{u}_h, \mathbf{u}_h, \mathbf{w}_h) \\
&\quad - (p - p_h, \nabla\cdot\mathbf{w}_h)_h - \ell_h(\mathbf{w}_h) \\
&= \nu\,(\nabla(\mathbf{u} - \mathbf{v}_h), \nabla\mathbf{w}_h)_h + [\tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - n_h(\mathbf{v}_h, \mathbf{v}_h, \mathbf{w}_h)] \\
&\quad + [n_h(\mathbf{v}_h, \mathbf{u}_h, \mathbf{w}_h) - n_h(\mathbf{u}_h, \mathbf{u}_h, \mathbf{w}_h)] \\
&\quad - (p - p_h, \nabla\cdot\mathbf{w}_h)_h - n_h(\mathbf{v}_h, \mathbf{w}_h, \mathbf{w}_h) - \ell_h(\mathbf{w}_h). \qquad (2.9)
\end{aligned}$$

We now estimate the terms on the right-hand side of (2.9). For each $p \in Q$ let $j_h p \in Q_h$ be the piecewise constant interpolant defined by

$$j_h p(x) := \frac{1}{\mathrm{meas}(T)} \int_T p\, dx \qquad \forall\, x \in T \in \mathcal{T}_h\,.$$

Then

$$\nu\,(\nabla(\mathbf{u} - \mathbf{v}_h), \nabla\mathbf{w}_h)_h \le \nu\,\|\mathbf{u} - \mathbf{v}_h\|_h\,\|\mathbf{w}_h\|_h,$$
$$-n_h(\mathbf{v}_h, \mathbf{w}_h, \mathbf{w}_h) \le 0 \quad \text{by Lemma 2.1},$$
$$(p - p_h, \nabla\cdot\mathbf{w}_h)_h = (p - j_h p, \nabla\cdot\mathbf{w}_h)_h \le Ch|p|_1\|\mathbf{w}_h\|_h,$$
$$|n_h(\mathbf{v}_h, \mathbf{u}_h, \mathbf{w}_h) - n_h(\mathbf{u}_h, \mathbf{u}_h, \mathbf{w}_h)| \le \frac{C}{\nu}\|\mathbf{f}\|_0\|\mathbf{w}_h\|_h^2 \le \frac{\nu}{2}\|\mathbf{w}_h\|_h^2$$

for $\nu \ge \nu_0$, provided that ν_0 is sufficiently large. It is shown in [CR73] that

$$|\ell_h(\mathbf{w}_h)| \le Ch\,(\nu|\mathbf{u}|_2 + |p|_1)\,\|\mathbf{w}_h\|_h \quad \forall\, \mathbf{w}_h \in \mathbf{V}_h\,. \qquad (2.10)$$

Taking $\mathbf{v}_h = i_h\mathbf{u} \in W_h$ (as defined in (2.6)) and using the usual interpolation error estimates, we deduce from (2.9) by means of Lemma 2.6 that

$$\|\mathbf{w}_h\|_h \le \frac{C}{\nu} M h, \qquad (2.11)$$

where $M = M(\|\mathbf{u}\|_2, \|p\|_1)$ is a polynomial of degree two. This implies that

$$\|\mathbf{u} - \mathbf{u}_h\|_h \leq \|\mathbf{u} - \mathbf{v}_h\|_h + \|\mathbf{w}_h\|_h \leq \frac{C}{\nu} Mh \leq \frac{C}{\nu_0} Mh.$$

To derive the bound on $\|p - p_h\|_0$, set $q_h := p_h - j_h p$ and, for arbitrary $\mathbf{w}_h \in \mathbf{V}_h$, estimate term by term the expression

$$-(q_h, \nabla \cdot \mathbf{w}_h)_h = -(p - j_h p, \nabla \cdot \mathbf{w}_h)_h + \nu (\nabla(\mathbf{u} - \mathbf{u}_h), \nabla \mathbf{w}_h)_h$$
$$+ \tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - n_h(\mathbf{u}_h, \mathbf{u}_h, \mathbf{w}_h) - \ell_h(\mathbf{w}_h).$$

Once again we have

$$|(p - j_h p, \nabla \cdot \mathbf{w}_h)_h| \leq Ch|p|_1 \|\mathbf{w}_h\|_h,$$
$$|\nu (\nabla(\mathbf{u} - \mathbf{u}_h), \nabla \mathbf{w}_h)_h| \leq CMh\|\mathbf{w}_h\|_h,$$

and, taking $\mathbf{v}_h = i_h \mathbf{u}$,

$$\tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - n_h(\mathbf{u}_h, \mathbf{u}_h, \mathbf{w}_h)$$
$$= \{\tilde{n}_h(\mathbf{u}, \mathbf{u}, \mathbf{w}_h) - n_h(\mathbf{v}_h, \mathbf{v}_h, \mathbf{w}_h)\} + n_h(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h, \mathbf{w}_h)$$
$$+ \{n_h(\mathbf{v}_h, \mathbf{v}_h, \mathbf{w}_h) - n_h(\mathbf{u}_h, \mathbf{v}_h, \mathbf{w}_h)\}$$
$$\leq CMh\|\mathbf{w}_h\|_h + C(\|\mathbf{u}_h\|_h + \|\mathbf{v}_h\|_h)\|\mathbf{v}_h - \mathbf{u}_h\|_h\|\mathbf{w}_h\|_h$$
$$\leq \frac{C}{\nu^2} Mh\|\mathbf{w}_h\|_h \,;$$

in deriving this estimate we invoked (2.11) to bound $\|\mathbf{v}_h - \mathbf{u}_h\|_h$, the stability inequality $\|\mathbf{v}_h\|_h \leq \|\mathbf{u}\|_h$ of the interpolation operator, and *a priori* estimates for $\|\mathbf{u}_h\|_h$ and $\|\mathbf{u}\|_h$ (viz., $|\mathbf{u}|_1 \leq C\nu^{-1}|\mathbf{f}|_*$, which follows from (1.4)). Combining these estimates with (2.10) and the discrete Babuška-Brezzi condition (2.1) yields

$$\|q_h\|_0 \leq \frac{C}{\beta\nu^2} Mh.$$

Hence

$$\|p - p_h\|_0 \leq \|p - j_h p\|_0 + \|q_h\|_0 \leq \frac{C}{\nu^2} Mh \leq \frac{C}{\nu_0^2} Mh,$$

i.e., first-order convergence of the pressure has been established.   □

Remark 2.8. The above proof reveals that three properties of the discrete trilinear form are needed to deliver the convergence result: (i) semidefiniteness (Lemma 2.1), (ii) Lipschitz continuity (Lemma 2.3), and (iii) linear consistency (Lemma 2.6). Angermann [Ang00] develops a general finite volume approach for the discretization of the trilinear form that is guaranteed to have these three properties.   ♣

Remark 2.9. Theorem 2.7 shows that we get first-order convergence for the velocity in the discrete H^1 norm and for the pressure in the L^2 norm. This is an optimal-order result for the nonconforming (P_1, P_0) element pair; in other words, the introduction of upwinding did not reduce the order of convergence

of the scheme. Our proof does however assume that the Reynolds number $\mathrm{Re} = \nu^{-1}$ is not too large, i.e., that $\nu > \nu_0$, where ν_0 depends on the continuity constant C of Lemma 2.3, the constant C in (2.5), and $\|\mathbf{f}\|_0$.   ♣

Remark 2.10. For the special case where $\mathbf{f} = \nabla\Psi$ with $\Psi \in H^1(\Omega) \cap L_0^2(\Omega)$, the solution of the continuous problem (1.2) is uniquely determined for all $\nu > 0$ since

$$(\mathbf{f}, \mathbf{v}) = (\nabla\Psi, v) = -(\Psi, \nabla \cdot \mathbf{v}) = 0 \quad \text{for all } \mathbf{v} \in \mathbf{W},$$

which implies that $\|\mathbf{f}\|_* = 0$. Consequently (1.5) is satisfied for every $\nu > 0$. We can therefore study the estimates of Theorem 2.7 for the whole range $0 < \nu \leq 1$. A careful analysis (see [DGT94] for details) shows that in this special case we now have

$$\|\mathbf{u} - \mathbf{u}_h\|_h \leq \frac{C}{\nu} M h \quad \text{and} \quad \|p - p_h\|_0 \leq CMh\left(1 + \frac{h}{\nu^2}\right)$$

with C independent of ν and h, and $M = M(\|\mathbf{u}\|_2, \|p\|_1)$. This observation is supported by a numerical example, with a discretization similar to ours but based on nonconforming quadrilateral finite elements, from [Sch94]: the H^1 error of the velocity is indeed $\mathcal{O}(\mathrm{Re})$ for a wide range of $\mathrm{Re} = 1/\nu$ when the mesh size h is fixed.   ♣

Remark 2.11. An L^2-norm error estimate for the velocity can be derived from the proof of Theorem 2.7. The interpolant $i_h\mathbf{u}$ defined by (2.6) satisfies the interpolation error estimate $\|\mathbf{u} - i_h\mathbf{u}\|_0 \leq C\,h^2\,|\mathbf{u}|_2$. Using (2.11), which was proved for $\mathbf{w}_h = \mathbf{u}_h - \mathbf{v}_h$, and the estimate (2.5) with $s = 2$, one gets the first-order L^2-norm error estimate

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq CMh,$$

where $M = M(\|\mathbf{u}\|_2, \|p\|_1)$. In [Tob89] it is shown that for the simple upwind discretization the L^2-norm error of the velocity is in general no better than $\mathcal{O}(h)$.   ♣

Remark 2.12. A similar upwind method can be constructed for the nonconforming "rotated bilinear" finite element; see [Tur91], where thorough numerical tests of this method agree with the $\mathcal{O}(h)$ convergence of velocity and pressure that is forecast in Theorem 2.7.   ♣

Remark 2.13. An alternative upwind method based on a streamfunction-vorticity formulation of the Navier-Stokes equations is developed in [For78] and analyzed in [GR82, GR86].   ♣

Remark 2.14. The method presented in this section can be extended to more general fluid-flow models. In [Dor95] the same technique is used to derive a stable finite element method for solving the Boussinesq approximation of the temperature-dependent formulation of the Navier-Stokes equations.   ♣

3
Higher-Order Methods of Streamline Diffusion Type

Sections III.3.2.1 and III.3.2.2 dealt with streamline diffusion (SDFEM) and Galerkin least squares (GLSFEM) finite element methods for scalar convection-diffusion problems. These methods try to achieve stability when convection dominates while obtaining high accuracy in subdomains that exclude boundary and interior layers. We shall now study the application of the streamline diffusion finite element method to the linearized variant of the Navier-Stokes equations, viz., the Oseen problem:

$$-\nu\Delta\mathbf{u} + (\mathbf{b}\cdot\nabla)\mathbf{u} + \sigma\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{3.1a}$$
$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega, \tag{3.1b}$$
$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega. \tag{3.1c}$$

Our method will also be applied to the full nonlinear problem

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \sigma\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{3.2a}$$
$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega, \tag{3.2b}$$
$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega. \tag{3.2c}$$

Here we assume that $\mathbf{b}$ is a smooth, divergence-free function and the constant σ is non-negative. As was pointed out in the introduction to Part IV, the term $\sigma\mathbf{u}$ arises in the time-discretization of the unsteady variants of (3.1) and (3.2); thus our analysis is relevant to these problems also. The present chapter looks only at conforming finite elements, i.e., the discrete spaces $\mathbf{V}_h$ and Q_h that approximate the velocity $\mathbf{u}$ and the pressure p are subspaces of the solution spaces in which $\mathbf{u}$ and p respectively lie.

An SDFEM based on the nonconforming (P_1, P_0) element is applied in [LT90] to (3.2) with $\sigma = 0$. See [FF92, Fra94, Lub94] for the solution of (3.1) and (3.2) by the GLSFEM.

We shall learn that the SDFEM can handle two types of instability – that caused by the dominance of convection, and that induced by discrete velocity and pressure spaces that fail to satisfy the discrete Babuška-Brezzi condition

(2.1). Thus when the SDFEM is applied to (3.1) or (3.2), the discrete spaces that approximate the velocity and the pressure can be chosen independently of each other. In Section 3.1 we obtain optimal error estimates for (3.1) for all mesh Péclet numbers, using natural norms that include in particular the L_2 norm of the pressure. Section 3.2 demonstrates that similar results are valid for the nonlinear problem, provided that it is close to a regular branch of solutions, i.e., that a linearized operator is an isomorphism; but the norm of the inverse of this linear operator will still depend on the Reynolds number and consequently the dependence of our error constants on the Reynolds number is not completely settled in the nonlinear case.

3.1 The Oseen Problem

The Oseen problem (3.1) is a linearization of the steady ($\sigma = 0$) and the unsteady ($\sigma > 0$) time-discretised Navier-Stokes equations in the bounded polyhedral domain $\Omega \subset \mathbb{R}^d$, where $d = 2$ or 3. Assume that $\mathbf{b} \in \mathbf{W}^{1,\infty}(\Omega)$ and $\nabla \cdot \mathbf{b} = 0$. Set $\mathbf{V} = \mathbf{H}_0^1(\Omega)$ and $Q = L_0^2(\Omega) := \{q \in L_2(\Omega) : (q,1) = 0\}$. Define the bilinear form A on the product space $\mathbf{V} \times Q$ by

$$A\big((\mathbf{u},p);(\mathbf{v},q)\big) := \nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + \big((\mathbf{b} \cdot \nabla)\mathbf{u}, \mathbf{v}\big) + \sigma(\mathbf{u}, \mathbf{v})$$
$$- (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}).$$

A weak formulation of the Oseen problem (3.1) reads:

Find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that one has

$$A\big((\mathbf{u},p);(\mathbf{v},q)\big) = (\mathbf{f}, \mathbf{v}) \qquad \forall(\mathbf{v}, q) \in \mathbf{V} \times Q. \tag{3.3}$$

The identity

$$\big((\mathbf{b} \cdot \nabla)\mathbf{v}, \mathbf{v}\big) = \frac{1}{2}\big((\mathbf{b} \cdot \nabla)(\mathbf{v} \cdot \mathbf{v}), 1\big) = -\frac{1}{2}(\nabla \cdot \mathbf{b}, \mathbf{v} \cdot \mathbf{v}) = 0 \quad \forall\mathbf{v} \in \mathbf{V}$$

allows us to apply the Lax-Milgram Lemma in the subspace of divergence-free functions and to find a unique velocity field $\mathbf{u}$. Then the Babuška-Brezzi condition for the pair $(\mathbf{V}, Q)$ [GR86] implies existence of a unique pressure $p \in Q$ such that $(\mathbf{u}, p)$ is a solution of (3.3). This uniqueness result was already stated in Theorem 1.5.

Let $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$ be two families of finite element spaces that correspond to a family of partitions $\mathcal{T}_h$ of Ω into polyhedral elements with maximum diameter h. Assume that each triangulation $\mathcal{T}_h$ is shape-regular. In particular, this permits the use of locally-refined meshes. Let $\mathcal{E}_h$ be the set of all inter-element boundaries in $\mathcal{T}_h$ and define

$$h_\omega := \text{diam}(\omega) = \sup_{x,y\in\omega} |x - y| \quad \text{for each } \omega \in \mathcal{T}_h \cup \mathcal{E}_h.$$

The shape-regularity of the triangulation $\mathcal{T}_h$ implies that the ratio h_T/h_E (where $E \in \mathcal{E}_h$, $T \in \mathcal{T}_h$, and $E \subset \partial T$) is bounded independently of h, T and E. For any $E \in \mathcal{E}_h$ with $E = T_{1E} \cap T_{2E}$, where $T_{1E}, T_{2E} \in \mathcal{T}_h$, and any $q \in L_2(\Omega)$ with $q|_{T_i} \in \mathcal{C}(\overline{T_i})$ for $i = 1, 2$, we use $[q]_E$ to denote the jump of q across E (in a fixed direction).

Regarding the approximation properties of the finite element pair $(\mathbf{V}_h, Q_h)$, we assume that the following interpolation error estimates and local inverse inequalities are fulfilled. Let $k \geq 1$ and $l \geq 0$ and let two interpolation operators $I_h : \mathbf{V} \to \mathbf{V}_h$ and $J_h : Q \to Q_h$ exist such that for all $T \in \mathcal{T}_h$ and all $E \in \mathcal{E}_h$ (where $E = T_{1E} \cap T_{2E}$ with $T_{1E}, T_{2E} \in \mathcal{T}_h$) one has for $0 \leq m \leq 2$, $\max\{m, 2\} \leq s \leq k + 1$, $0 \leq i \leq 1$ and $2 \leq j \leq l + 1$ the bounds

$$\|\mathbf{u} - I_h\mathbf{u}\|_{m,T} \leq c_1 h_T^{s-m} |\mathbf{u}|_{s,T} \qquad \text{for all } \mathbf{u} \in H^s(T)^d, \tag{3.4a}$$

$$\|\mathbf{u} - I_h\mathbf{u}\|_{0,E} \leq c_2 h_E^{s-1/2} |\mathbf{u}|_{s,T_{1E} \cup T_{2E}} \quad \text{for all } \mathbf{u} \in H^s(T_{1E} \cup T_{2E})^d, \tag{3.4b}$$

$$\|p - J_h p\|_{i,T} \leq c_3 h_T^{j-i} |p|_{j,T} \qquad \text{for all } p \in H^j(T), \tag{3.4c}$$

$$\|p - J_h p\|_{0,E} \leq c_4 h_E^{j-1/2} |p|_{j,T_{1E} \cup T_{2E}} \quad \text{for all } p \in H^j(T_{1E} \cup T_{2E}), \tag{3.4d}$$

$$\|\Delta \mathbf{v}_h\|_{0,T} \leq \mu_{\text{inv}} h_T^{-1} \|\nabla \mathbf{v}_h\|_{0,T} \qquad \text{for all } \mathbf{v}_h \in \mathbf{V}_h, \tag{3.4e}$$

$$\|\nabla^m p_h\|_{0,T} \leq c_5 h_T^{-m} \|p_h\|_{0,T} \qquad \text{for all } p_h \in Q_h, \tag{3.4f}$$

$$\|[p_h]_E\|_{0,E} \leq c_6 h_E^{-1/2} \|p_h\|_{0,T_{1E} \cup T_{2E}} \quad \text{for all } p_h \in Q_h. \tag{3.4g}$$

These conditions are satisfied if $\mathbf{V}_h$ and Q_h comprise piecewise polynomials of degrees at most k and l respectively, while $I_h : H^2(\Omega)^d \to \mathbb{R}^d$ and $J_h : H^2(\Omega) \to \mathbb{R}$ are the standard nodal interpolation operators. To prove the inf-sup condition of Lemma 3.4, however, one needs an interpolation operator $i_h : H^1(\Omega)^d \to \mathbb{R}^d$ that is defined on the larger space $H^1(\Omega)^d$. For non-smooth functions, the existence theory of interpolation operators that yield estimates like (3.4) is well established in the literature [Ape99, Clé75, SZ90]. One example is discussed in the next remark.

Remark 3.1. (Scott-Zhang interpolant) As an example of an interpolation operator for non-smooth functions, let us construct the Scott-Zhang operator for a space Y_h that comprises continuous piecewise polynomial functions of degree at most k on a simplicial mesh. Let φ_i, $i \in I$, be the standard nodal basis functions with respect to the nodes p_i in the finite element space Y_h and define

$$i_h u(x) := \sum_{i \in I} a_i \varphi_i(x)$$

where the real numbers a_i, $i \in I$, are yet to be specified. The standard Lagrange interpolant of a continuous function u is given by $a_i := u(p_i)$, $i \in I$, but we are interested in interpolation operators that are defined on a larger set of functions. For inner nodes p_i of a cell T, replace $a_i := u(p_i)$ by $a_i := (\pi_T u)(p_i)$ where $\pi_T : L_2(T) \to P_r(T)$ is the $L_2(T)$ projection. If p_i is a boundary node of one or more cells $T \in \mathcal{T}_h$, choose some $(d-1)$-dimensional face E_i of one

of these elements. In the case of a boundary node $p_i \in \partial\Omega$, we restrict the choice of E_i to boundary faces $E_i \subset \partial\Omega$. Then, for all boundary nodes p_i of one or more cells we set $a_i := (\pi_{E_i} u)(p_i)$ where $\pi_{E_i} : L_2(E_i) \to P_r(E_i)$ is the $L_2(E_i)$ projection. Unlike the Lagrange interpolant this L_2-based interpolation is defined for each function $u \in H^1(\Omega)$, since the trace of each such function on $E_i \subset T$ belongs to $L_2(E_i)$. Moreover, if $u|_{\partial\Omega}$ is a continuous
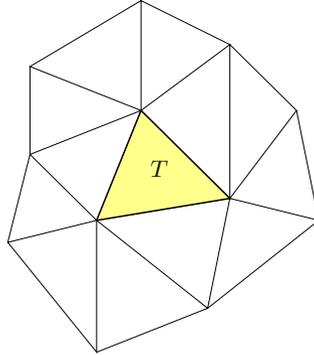
Fig. 3.1. A possible neighbourhood $\omega(T)$ of a cell T for the Scott-Zhang operator

piecewise polynomial function of degree at most k, then $i_h u = u$ on $\partial\Omega$. This property guarantees that homogeneous Dirichlet data will be interpolated in the correct manner. The Scott-Zhang operator satisfies

$$\|\mathbf{u} - i_h\mathbf{u}\|_{m,T} \leq c_1 h_T^{s-m} |\mathbf{u}|_{s,\omega(T)} \quad \text{for all } \mathbf{u} \in H^s(\omega(T))^d, \tag{3.5a}$$

$$\|\mathbf{u} - i_h\mathbf{u}\|_{0,E} \leq c_2 h_E^{1/2} |\mathbf{u}|_{1,\omega(T)} \quad \text{for all } E \subset \partial T, \ \mathbf{u} \in H^1(\omega(T))^d, \tag{3.5b}$$

for $\max\{1,m\} \leq s \leq k$, $0 \leq m \leq 2$, where $\omega(T)$ is a certain local neighbourhood of a cell T as drawn in Figure 3.1 [SZ90].

An alternative interpolant is the Clément operator of Section III.3.6.   ♣

The streamline diffusion finite element method (SDFEM) for solving the Oseen problem (3.3) is obtained by adding to (3.3) both a least-squares control of the divergence and, on each element, a weak form of the momentum equation using test functions of the form $(\mathbf{b} \cdot \nabla)\mathbf{v} + \nabla q$ for $(\mathbf{v}, q) \in \mathbf{V} \times Q$:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in V_h \times Q_h$ one has

$$A_\delta\left((\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)\right) = L_\delta\left((\mathbf{v}_h, q_h)\right), \tag{3.6}$$

where

$$A_\delta\left((\mathbf{w},r),(\mathbf{v},q)\right) := \nu(\nabla\mathbf{w},\nabla\mathbf{v}) + ((\mathbf{b}\cdot\nabla)\mathbf{w} + \sigma\mathbf{w},\mathbf{v}) - (r,\nabla\cdot\mathbf{v})$$

$$+ (q,\nabla\cdot\mathbf{w}) + \mu(\nabla\cdot\mathbf{w},\nabla\cdot\mathbf{v}) + \gamma\sum_{E\in\mathcal{E}_h} h_E([r]_E,[q]_E)_E$$

$$+ \sum_{T\in\mathcal{T}} \delta_T(-\nu\Delta\mathbf{w} + (\mathbf{b}\cdot\nabla)\mathbf{w} + \sigma\mathbf{w} + \nabla r,(\mathbf{b}\cdot\nabla)\mathbf{v} + \nabla q)_T,$$

$$L_\delta\left((\mathbf{v},q)\right) := (f,\mathbf{v}) + \sum_{T\in\mathcal{T}_h} \delta_T(\mathbf{f},(\mathbf{b}\cdot\nabla)\mathbf{v} + \nabla q)_T,$$

and $\mu \geq 0$, $\delta_T > 0$, $\delta := \max_T \delta_T$ and $\gamma > 0$ are parameters that will be determined later.

Remark 3.2. The pressure jumps across inter-element boundaries $E \in \mathcal{E}_h$ are present in $A_\delta(\cdot,\cdot)$ to allow discontinuous pressure approximations. As in [FS91], they can be omitted if one has $k \geq d$, where k is the polynomial degree of the velocity space V_h. ♣

Remark 3.3. On setting $\mathbf{b} = \mu = 0$, the SDFEM (3.6) reduces to the Petrov-Galerkin discretization of the Stokes equations considered in [HFB86, HFB87]. If we add a term $-\nu\Delta\mathbf{v}$ to the $(\cdot,\cdot)_T$ inner products in $A_\delta(\cdot,\cdot)$ and $L_\delta(\cdot)$, and choose suitable signs for the coefficients, then in the case $\nu = 1$ and $b = \mu = 0$ the SDFEM (3.6) corresponds to the Petrov-Galerkin discretizations of the Stokes problem considered in [DW89, FH88]. ♣

The SDFEM (3.6) is a residual-based method and so is consistent, i.e.,

$$A_\delta((\mathbf{u},p),(\mathbf{w}_h,r_h)) = L_\delta((\mathbf{w}_h,r_h)) \quad \text{for all } (\mathbf{w}_h,r_h) \in \mathbf{V}_h \times Q_h,$$

provided that the exact solution $(\mathbf{u},p)$ of (3.3) satisfies the local regularity condition

$$\mathbf{f} + \nu\Delta\mathbf{u} - (\mathbf{b}\cdot\nabla)\mathbf{u} - \sigma\mathbf{u} - \nabla p \in L_2(T)^d \quad \text{for each } T \in \mathcal{T}_h.$$

In this case one has the projection property

$$A_\delta((\mathbf{u}-\mathbf{u}_h,p-p_h),(\mathbf{w}_h,r_h)) = 0 \qquad \text{for all } (\mathbf{w}_h,r_h) \in \mathbf{V}_h \times Q_h. \quad (3.7)$$

We introduce the mesh-dependent norm

$$|||(\mathbf{v},q)|||_h := \left\{ \nu|\mathbf{v}|_1^2 + \sigma\|\mathbf{v}\|_0^2 + \nu\|q\|_0^2 + \mu\|\nabla\cdot\mathbf{v}\|_0^2 \right.$$

$$\left. + \gamma\sum_{E\in\mathcal{E}_h} h_E\|[q]_E\|_{0,E}^2 + \sum_{T\in\mathcal{T}_h} \delta_T\|(\mathbf{b}\cdot\nabla)\mathbf{v} + \nabla q\|_{0,T}^2 \right\}^{1/2}.$$

Lemma 3.4. *Assume that $\sigma \geq 0$, $\mu \geq 0$, $\gamma > 0$ and that for a fixed positive constant δ_0 the local SD parameter δ_T is chosen such that*

$$0 < \delta_0 h_T^2 \leq \delta_T \leq \min\left\{\delta, \frac{h_T^2}{2\nu\mu_{inv}^2}\right\} \qquad and \qquad 0 \leq \sigma\,\delta_T \leq \frac{1}{2}\,.$$

Then

$$\inf_{\substack{(\mathbf{v}_h,q_h)\in\mathbf{V}_h\times Q_h \\ |||(\mathbf{v}_h,q_h)|||_h=1}} \sup_{\substack{(\mathbf{w}_h,r_h)\in\mathbf{V}_h\times Q_h \\ |||(\mathbf{w}_h,r_h)|||_h=1}} A_\delta((\mathbf{v}_h,q_h),(\mathbf{w}_h,r_h)) \;\geq\; \beta,$$

*where the positive constant β is independent of ν and h. Moreover, $\beta = \mathcal{O}(\delta_0)$
as $\delta_0 \to 0$.*

Proof. Consider an arbitrary point $(\mathbf{v}_h,q_h) \in \mathbf{V}_h \times Q_h$ and for brevity put

$$X := \left\{ \sum_{T\in\mathcal{T}_h} \delta_T \|(\mathbf{b}\cdot\nabla)\mathbf{v}_h + \nabla q_h\|_{0,T}^2 \right\}^{1/2},$$

$$Y := \left\{ \gamma \sum_{E\in\mathcal{E}_h} h_E \|[q_h]_E\|_{0,E}^2 \right\}^{1/2}, \qquad Z := \mu^{1/2}\|\nabla\cdot\mathbf{v}_h\|_0.$$

Then

$$A_\delta((\mathbf{v}_h,q_h),(\mathbf{v}_h,q_h)) = \nu|\mathbf{v}_h|_1^2 + ((\mathbf{b}\cdot\nabla)\mathbf{v}_h,\mathbf{v}_h) + \sigma\|\mathbf{v}_h\|_0^2 + X^2 + Y^2 + Z^2$$
$$+ \sum_{T\in\mathcal{T}_h} \delta_T(\sigma\mathbf{v}_h - \nu\Delta\mathbf{v}_h, (\mathbf{b}\cdot\nabla)\mathbf{v}_h + \nabla q_h)_T.$$

Applying the inverse estimate (3.4e) and the hypothesis on the upper bound
of δ_T, we can absorb the last sum into the other terms:

$$\left| \sum_{T\in\mathcal{T}_h} \delta_T(\sigma\mathbf{v}_h - \nu\Delta\mathbf{v}_h, (\mathbf{b}\cdot\nabla)\mathbf{v}_h + \nabla q_h)_T \right|$$
$$\leq \sigma \sum_{T\in\mathcal{T}_h} \sigma\delta_T\|\mathbf{v}_h\|_{0,T}^2 + \nu \sum_{T\in\mathcal{T}_h} \nu\delta_T h_T^{-2}\mu_{inv}^2 |\mathbf{v}_h|_{1,T}^2 + \frac{1}{2}X^2$$
$$\leq \frac{1}{2}\left(\sigma\|\mathbf{v}_h\|_0^2 + \nu|\mathbf{v}_h|_1^2 + X^2 \right).$$

Recalling that $\nabla\cdot\mathbf{b} = 0$, one obtains

$$A_\delta((\mathbf{v}_h,q_h),(\mathbf{v}_h,q_h)) \geq \frac{1}{2}\left[\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + X^2 + Y^2 + Z^2 \right]. \qquad (3.8)$$

This inequality shows that A_δ is coercive over $\mathbf{V}_h\times Q_h$ with respect to the
norm defined by the square root of the right-hand side. Nevertheless we are
interested in proving error estimates in a more natural norm for this problem
– for example, one that includes the L_2 norm of the pressure, which is missing
from (3.8).

It will be shown that there exists a constant M, which is independent of
ν and h, such that for each $(\mathbf{v}_h,q_h) \in \mathbf{V}_h \times Q_h$ one can choose an element
$(\mathbf{w}_h,0) \in \mathbf{V}_h \times Q_h$ for which

$$A_\delta((\mathbf{v}_h, q_h), (\mathbf{w}_h, 0)) \geq \frac{3}{8}\|q_h\|_0^2 - 2M^2\Big(|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + X^2 + Y^2 + Z^2\Big).$$
(3.9)

By [GR86, Chapter I, Lemma 4.1], there exist $\mathbf{w} \in \mathbf{V}$ and a constant c_Ω that depends only on the geometry of Ω, such that $\nabla \cdot \mathbf{w} = -q_h$ and $\|\mathbf{w}\|_1 \leq c_\Omega \|q_h\|_0$. Set $\mathbf{w}_h = i_h \mathbf{w}$ where i_h is the Scott-Zhang operator of Remark 3.1. By (3.5a) with $m = s = 1$, one has $\|w_h\|_1 \leq c'_\Omega \|q_h\|_0$ with $c'_\Omega := (1 + c_1)c_\Omega$. Moreover, the convective term is a continuous trilinear form on $\mathbf{V} \times \mathbf{V} \times \mathbf{V}$, so

$$|((\mathbf{u} \cdot \nabla)\mathbf{v}, \mathbf{w})| \leq M_0 |\mathbf{u}|_1 |\mathbf{v}|_1 |\mathbf{w}|_1 \quad \text{for all } \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{V},$$

where the constant M_0 depends only on the geometry of Ω.

Integrating by parts on each element, one obtains

$$A_\delta\big((\mathbf{v}_h, q_h), (\mathbf{w}_h, 0)\big) =$$

$$\sigma(\mathbf{v}_h, \mathbf{w}_h) + \nu(\nabla \mathbf{v}_h, \nabla \mathbf{w}_h) + ((\mathbf{b} \cdot \nabla)\mathbf{v}_h, \mathbf{w}_h) - (q_h, \nabla \cdot \mathbf{w})$$

$$- \sum_{T \in \mathcal{T}_h} (\nabla q_h, \mathbf{w} - \mathbf{w}_h)_T + \sum_{E \in \mathcal{E}_h} ([q_h]_E, (\mathbf{w} - \mathbf{w}_h) \cdot \mathbf{n}_E)_E$$

$$+ \sum_{T \in \mathcal{T}_h} \delta_T(\sigma \mathbf{v}_h - \nu \Delta \mathbf{v}_h + (\mathbf{b} \cdot \nabla)\mathbf{v}_h + \nabla q_h, (\mathbf{b} \cdot \nabla)\mathbf{w}_h)_T$$

$$+ \mu(\nabla \cdot \mathbf{v}_h, \nabla \cdot \mathbf{w}_h).$$

Now

$$|\sigma(\mathbf{v}_h, \mathbf{w}_h)| \leq \sigma c'_\Omega \|\mathbf{v}_h\|_0 \|q_h\|_0,$$
$$|\nu(\nabla \mathbf{v}_h, \nabla \mathbf{w}_h)| \leq c'_\Omega |\mathbf{v}_h|_1 \|q_h\|_0,$$
$$|((\mathbf{b} \cdot \nabla)\mathbf{v}_h, \mathbf{w}_h)| \leq M_0 c'_\Omega |\mathbf{b}|_1 |\mathbf{v}_h|_1 \|q_h\|_0,$$

and $-(q_h, \nabla \cdot \mathbf{w}) = \|q_h\|_0^2$. Furthermore, recalling the properties (3.5) of the interpolation operator and the choice of δ_T, we get

$$\Big| \sum_{T \in \mathcal{T}_h} (\nabla q_h, \mathbf{w} - \mathbf{w}_h)_T \Big|$$

$$\leq \sum_{T \in \mathcal{T}_h} \Big(\|(\mathbf{b} \cdot \nabla)\mathbf{v}_h\|_{0,T} + \|(\mathbf{b} \cdot \nabla)\mathbf{v}_h + \nabla q_h\|_{0,T} \Big) c_1 h_T |\mathbf{w}|_{1,T}$$

$$\leq c_1 c_\Omega h \|\mathbf{b}\|_\infty |\mathbf{v}_h|_1 \|q_h\|_0 + X c_1 \Big(\sum_{T \in \mathcal{T}_h} \delta_T^{-1} h_T^2 |\mathbf{w}|_{1,T}^2 \Big)^{1/2}$$

$$\leq c_1 c_\Omega h \|\mathbf{b}\|_\infty |\mathbf{v}_h|_1 \|q_h\|_0 + c_1 \delta_0^{-1/2} c_\Omega X \|q_h\|_0,$$

$$\Big| \sum_{E \in \mathcal{E}_h} ([q_h]_E, (\mathbf{w} - \mathbf{w}_h) \cdot \mathbf{n}_E)_E \Big| \leq \sum_{E \in \mathcal{E}_h} \|[q_h]_E\|_{0,E} \, c_2 h_E^{1/2} |\mathbf{w}|_{1,T_{1E} \cup T_{2E}},$$

$$\leq 3 Y c_2 \gamma^{-1/2} c_\Omega \|q_h\|_0,$$

$$\left| \sum_{T \in \mathcal{T}_h} \delta_T (\sigma \mathbf{v}_h - \nu \Delta \mathbf{v}_h + (\mathbf{b} \cdot \nabla) \mathbf{v}_h + \nabla q_h \ , (\mathbf{b} \cdot \nabla) \mathbf{w}_h)_T \right|$$

$$\leq \sigma \, \delta c'_\Omega \|\mathbf{b}\|_\infty \, \|\mathbf{v}_h\|_0 \, \|q_h\|_0 + (2\mu_{\mathrm{inv}})^{-1} c'_\Omega h \|\mathbf{b}\|_\infty \, |\mathbf{v}_h|_1 \, \|q_h\|_0$$

$$+ X c'_\Omega \delta^{1/2} \|\mathbf{b}\|_\infty \, \|q_h\|_0,$$

and

$$|\mu(\nabla \cdot \mathbf{v}_h, \nabla \cdot \mathbf{w}_h)| \leq Z \mu^{1/2} \|\nabla \cdot \mathbf{w}_h\|_0 \leq Z \mu^{1/2} d^{1/2} c'_\Omega \, \|q_h\|_0.$$

Combining all the above estimates, we arrive at

$$A_\delta\big((\mathbf{v}_h, q_h), (\mathbf{w}_h, 0)\big) \geq \|q_h\|_0^2 - M \|q_h\|_0 \left(\sigma^{1/2} \|\mathbf{v}_h\|_0 + |\mathbf{v}_h|_1 + X + Y + Z \right)$$

$$\geq \frac{3}{8} \|q_h\|_0^2 - 2M^2 \left(|\mathbf{v}_h|_1^2 + \sigma \|\mathbf{v}_h\|_0^2 + X^2 + Y^2 + Z^2 \right),$$

where $M = M_1 + M_2 \delta^{1/2}$, and M_1 and M_2 are independent of ν, h and δ but depend on μ, γ, σ, $\|\mathbf{b}\|_\infty$ and $|\mathbf{b}|_1$. This completes the proof of (3.9).

Next, multiply (3.8) by $1 - \rho\nu$ and (3.9) by $\rho\nu$, where $\rho \geq 0$, then add the ensuing inequalities. This yields

$$A_\delta\big((\mathbf{v}_h, q_h), (\mathbf{z}_h, r_h)\big) \geq \frac{3}{8} \rho\nu \|q_h\|_0^2 + \left(\frac{1 - \rho\nu}{2} - 2M^2 \rho \right) \cdot$$

$$\left(\nu |\mathbf{v}_h|_1^2 + \sigma \|\mathbf{v}_h\|_0^2 + X^2 + Y^2 + Z^2 \right)$$

$$\geq \min \left\{ \frac{3}{8} \rho, \frac{1 - \rho\nu}{2} - 2M^2 \rho \right\} |||(\mathbf{v}_h, q_h)|||_h^2,$$

where

$$(\mathbf{z}_h, r_h) := \big((1 - \rho\nu) \mathbf{v}_h + \rho\nu \mathbf{w}_h, (1 - \rho\nu) q_h \big).$$

The minimum is maximized by taking $\rho = 4/(3 + 4\nu + 16M^2)$. Then

$$A_\delta\big((\mathbf{v}_h, q_h), (\mathbf{z}_h, r_h)\big) \geq \frac{3}{6 + 8\nu + 32M^2} |||(\mathbf{v}_h, q_h)|||_h^2$$

$$\geq \frac{3}{14 + 32M^2} |||(\mathbf{v}_h, q_h)|||_h^2. \tag{3.10}$$

On the other hand, a careful study of M shows that

$$|||(\mathbf{z}_h, r_h)|||_h \leq (1 - \rho\nu) |||(\mathbf{v}_h, q_h)|||_h + \rho\nu |||(\mathbf{w}_h, 0)|||_h$$

$$\leq |||(\mathbf{v}_h, q_h)|||_h + 4\rho\nu M \|q_h\|_0$$

$$\leq (1 + 4\rho M) |||(\mathbf{v}_h, q_h)|||_h$$

$$\leq 2 |||(\mathbf{v}_h, q_h)|||_h, \tag{3.11}$$

where we used the bound $\rho M \leq 1/(2\sqrt{3}) < 1/4$. Combining (3.10) and (3.11) yields the desired estimate

$$A_\delta\big((\mathbf{v}_h, q_h), (\mathbf{z}_h, r_h)\big) \geq \beta \, |||(\mathbf{v}_h, q_h)|||_h \, |||(\mathbf{z}_h, r_h)|||_h, \quad \text{with } \beta = \frac{3}{28 + 64M^2}.$$

A detailed investigation of M shows that $M^2 = \mathcal{O}(\delta_0^{-1})$; the asymptotic behaviour of β relative to δ_0 follows. $\square$

Remark 3.5. If we use continuous pressure approximations (i.e., $Q_h \subset H^1(\Omega)$), then the pressure jumps across inter-element boundaries vanish. In this case Lemma 3.4 is still valid with $\gamma = 0$. ♣

Lemma implies that problem (3.6) has a unique solution $(\mathbf{u}_h, p_h)$. We now derive an error estimate for this solution. To this end, assume that the solution of the Oseen problem (3.3) is sufficiently regular.

Theorem 3.6. *Let* $(\mathbf{u}, p) \in V \times Q$ *and* $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ *be the unique solutions of problems (3.3) and (3.6) respectively. Assume that* δ_T *satisfies*

$$0 < \delta_0 h_T^2 \leq \delta_T \leq \min\left\{\delta, \frac{h_T^2}{2\nu\mu_{inv}^2}\right\} \quad \text{and} \quad 0 \leq \sigma\,\delta_T \leq \frac{1}{2}$$

for some positive constant δ_0, *that* $\mu > 0$, $\gamma > 0$ *and that* $(\mathbf{u}, p)$ *lies in* $H^{k+1}(\Omega)^d \times H^{l+1}(\Omega)$ *for some* $k \geq 1$ *and some* $l \geq 0$. *Then one has the error estimate*

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||_h \leq E_u\,h^k\|\mathbf{u}\|_{k+1} + E_p\,h^l\|p\|_{l+1} \tag{3.12}$$

where

$$E_u \leq C\left(\nu^{1/2} + \sigma^{1/2}h + \mu^{1/2} + \delta^{1/2} + \delta_0^{-1/2} + \gamma^{-1/2} + \delta^{1/2}\sigma h\right),$$

$$E_p \leq C\left(\nu^{1/2}h + \gamma^{1/2}h + \delta^{1/2} + \eta h \min\left\{\mu^{-1/2}, \nu^{-1/2}\right\}\right).$$

Here C *is independent of* ν *and* h. *Moreover,* $\eta = 0$ *if* $\nabla \cdot V_h \subset Q_h$; *otherwise* $\eta = 1$.

Proof. Put $\mathbf{v}_h = I_h\mathbf{u}$ and denote by q_h the L_2 projection of p onto Q_h. Since $k \geq 1$, the interpolant I_h can be the standard nodal interpolant; it is unnecessary to introduce the Scott-Zhang interpolant as in the proof of Lemma 3.4. Comparing $p - q_h$ with $p - J_h p$ and using the inverse estimates (3.4f) and (3.4g), one checks easily that $p - q_h$ satisfies the error estimates (3.4c) and (3.4d) with modified constants. Thus we have the interpolation error estimate

$$|||(\mathbf{u} - \mathbf{v}_h, p - q_h)|||_h \leq Ch^k(\nu^{1/2} + \sigma^{1/2}h + \mu^{1/2} + \delta^{1/2})\|\mathbf{u}\|_{k+1}$$
$$+ Ch^l(\nu^{1/2}h + \gamma^{1/2}h + \delta^{1/2})\|p\|_{l+1}.$$

On the other hand, Lemma 3.4 and the projection property (3.7) imply that

$$|||(\mathbf{u}_h - \mathbf{v}_h, p_h - q_h)|||_h$$
$$\leq \frac{1}{\beta} \sup_{\substack{(\mathbf{w}_h, r_h) \in V_h \times Q_h \\ |||(\mathbf{w}_h, r_h)|||_h = 1}} A_\delta((\mathbf{u}_h - \mathbf{v}_h, p_h - q_h), (\mathbf{w}_h, r_h))$$
$$= \frac{1}{\beta} \sup_{\substack{(\mathbf{w}_h, r_h) \in V_h \times Q_h \\ |||(\mathbf{w}_h, r_h)|||_h = 1}} A_\delta((\mathbf{u} - \mathbf{v}_h, p - q_h), (w_h, r_h)).$$

To bound the right-hand side of this inequality, consider an arbitrary element $(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h$ with $|||(\mathbf{w}_h, r_h)|||_h = 1$ and estimate separately the various terms in $A_\delta((\mathbf{u} - \mathbf{v}_h, p - q_h), (\mathbf{w}_h, r_h))$:

$$\sigma(\mathbf{u} - \mathbf{v}_h, \mathbf{w}_h) \leq C\, \sigma^{1/2} h^{k+1} \|\mathbf{u}\|_{k+1} |||(\mathbf{w}_h, r_h)|||_h,$$

$$\nu(\nabla(\mathbf{u} - \mathbf{v}_h), \nabla \mathbf{w}_h) \leq C\, \nu^{1/2} h^k \|\mathbf{u}\|_{k+1} |||(\mathbf{w}_h, r_h)|||_h,$$

$$\sum_{T \in \mathcal{T}_h} \delta_T(\sigma(\mathbf{u} - \mathbf{v}_h) - \nu \Delta(\mathbf{u} - \mathbf{v}_h), (\mathbf{b} \cdot \nabla)\mathbf{w}_h + \nabla r_h)_T$$
$$\leq C\, h^k (\delta^{1/2}\sigma h + \nu^{1/2}) \|\mathbf{u}\|_{k+1} |||(\mathbf{w}_h, r_h)|||_h,$$

$$\sum_{T \in \mathcal{T}_h} \delta_T((\mathbf{b} \cdot \nabla)(\mathbf{u} - \mathbf{v}_h) + \nabla(p - q_h), (\mathbf{b} \cdot \nabla)\mathbf{w}_h + \nabla r_h)_T$$
$$\leq C\, \delta^{1/2}\{h^k \|\mathbf{u}\|_{k+1} + h^l \|p\|_{l+1}\} |||(\mathbf{w}_h, r_h)|||_h,$$

$$\gamma \sum_{E \in \mathcal{E}_h} h_E([p - q_h]_E, [r_h]_E)_E$$
$$\leq \{\gamma \sum_{E \in \mathcal{E}_h} h_E \|[p - q_h]_E\|_{0,E}^2\}^{1/2} |||(\mathbf{w}_h, r_h)|||_h$$
$$\leq C\, \gamma^{1/2} h^{l+1} \|p\|_{l+1} |||(\mathbf{w}_h, r_h)|||_h,$$

$$((\mathbf{b} \cdot \nabla)(\mathbf{u} - \mathbf{v}_h), \mathbf{w}_h) + (r_h, \nabla \cdot (\mathbf{u} - \mathbf{v}_h))$$
$$= -\sum_{T \in \mathcal{T}_h} ((\mathbf{b} \cdot \nabla)\mathbf{w}_h + \nabla r_h, \mathbf{u} - \mathbf{v}_h)_T + \sum_{E \in \mathcal{E}_h} ((\mathbf{u} - \mathbf{v}_h) \cdot \mathbf{n}_E, [r_h]_E)_E$$
$$\leq C\, h^k (\delta_0^{-1/2} + \gamma^{-1/2}) \|\mathbf{u}\|_{k+1} |||(\mathbf{w}_h, r_h)|||_h,$$

$$\mu(\nabla \cdot (\mathbf{u} - \mathbf{v}_h), \nabla \cdot \mathbf{w}_h) \leq C\, \mu^{1/2} h^k \|\mathbf{u}\|_{k+1} |||(\mathbf{w}_h, r_h)|||_h,$$

$$-(p - q_h, \nabla \cdot \mathbf{w}_h) \leq \eta \|p - q_h\|_0 \min\{d^{1/2} |\mathbf{w}_h|_1, \|\nabla \cdot \mathbf{w}_h\|_0\}$$
$$\leq C\, \eta\, \min\{\mu^{-1/2}, \nu^{-1/2}\}\, h^{l+1} \|p\|_{l+1} |||(\mathbf{w}_h, r_h)|||_h.$$

These estimates and a triangle inequality complete the proof. □

Remark 3.7. The mesh-dependent norm $|||(\cdot, \cdot)|||_h$ controls the quantities

$$\nu^{1/2} |\mathbf{u}|_1 + \sigma^{1/2} \|\mathbf{u}\|_0 \quad \text{and} \quad \mu^{1/2} \|\nabla \cdot \mathbf{u}\|_0.$$

If $l = k-1$, i.e., if $\mathbf{V}_h$ and Q_h consist of piecewise polynomials of degrees k and $k-1$ respectively, then (3.12) implies that the choices $\delta_T \sim h_T^2$, $\mu \sim$ constant and $\sigma \sim$ constant are optimal. For these choices, the right-hand side of the error estimate (3.12) has the form

$$C\, h^k \Big[\big(1 + \nu^{1/2} + h + h^2\big) \|\mathbf{u}\|_{k+1} + \big(1 + \nu^{1/2}\big) \|p\|_k \Big], \tag{3.13}$$

so the estimate is optimal, independently of the mesh Péclet number $\nu^{-1}h$. Of course the Sobolev norms $\|\mathbf{u}\|_{k+1}$ and $\|p\|_k$ still, in general, depend on ν. The above choice of the SD parameter δ_T differs from that made in scalar convection-diffusion equations where (see (III.3.38) in Section III.3.2.1) the usual choice of δ_T is

$$\delta_T \sim \begin{cases} h_T^2/\nu & \text{if } \nu \ge h_T, \\ h_T & \text{if } \nu < h_T. \end{cases}$$

This change is due to the coupling of two different phenomena – dominant convection and the incompressibility condition. See also Remark 3.8 below.

If $\sigma = Ch^{-\rho}$ with $\rho > 0$, which occurs when the Oseen problem (3.3) is generated by a time-discretization of the corresponding unsteady problem, then provided $\rho < 2$ (so that the hypotheses on δ_T in Theorem 3.6 are satisfied) one sees that (3.12) gives a better L_2-error estimate for the velocity than the case $\sigma = 1$ because of the σ-weighting in $\|\|\cdot\|\|_h$. ♣

Remark 3.8. When using equal-order interpolation (i.e., $k = l$), the estimate (3.12) implies that the choices $\mu \sim h$, $\sigma \sim h^{-1}$ and $\delta_T \sim \min\{h_T, h_T^2/\nu\}$ are optimal. This recovers the "classical" choice (III.3.38) of the SD parameter for scalar convection-diffusion problems. In the interesting case $\nu < h$, setting $\delta_T \sim h_T$ in the proof of Theorem 3.6 yields an estimate of the form

$$\|\|(\mathbf{u} - \mathbf{u}_h, p - p_h)\|\|_h \le C\, h^{k+1/2} \left(\|\mathbf{u}\|_{k+1} + \|p\|_{k+1}\right),$$

which is better than the bound of (3.13). Thus equal-order interpolation is suitable for situations where the pressure is sufficiently smooth. In general, however, regularity theory tells us that $\|\mathbf{u}\|_{m+1}$ and $\|p\|_m$ are comparable. Equal-order interpolation with $k = l = m - 1$ then yields only an $\mathcal{O}(h^{m-1/2})$ error estimate, which is inferior to the choice $k - 1 = l = m - 1$ for which one has an $\mathcal{O}(h^m)$ error estimate. ♣

Remark 3.9. When continuous pressure approximations are used, i.e., when $Q_h \subset H^1(\Omega)$, the pressure jumps across inter-element boundaries vanish. One can then repeat the proof of Theorem 3.6 with $\gamma = 0$ and derive the error estimate (3.12) without the terms that contain γ. ♣

Remark 3.10. Discontinous pressure approximations of degree $k - 1$ and continuous velocity approximations of degree k usually give $\nabla \cdot \mathbf{V}_h \subset Q_h$, so $\eta = 0$ in Theorem 3.6. Thus one can choose $\mu = 0$, i.e., the least-squares term of the divergence can be removed from the SDFEM (3.6). Note that such pairs of finite elements do not in general satisfy the Babuška-Brezzi condition on shape-regular families of meshes. It has been shown recently [Qin94, SV85, Zha05] that the Scott-Vogelius element of continuous velocity approximations of degree $k \ge d$ and discontinuous pressure approximations of degree $k - 1$ satisfies the Babuška-Brezzi condition on special macro-type triangulations. ♣

3.2 The Navier-Stokes Problem

This section extends our analysis of the SDFEM to the nonlinear Navier-Stokes equations written in velocity-pressure form:

$$\tilde{\sigma}\mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla\tilde{p} = \tilde{\mathbf{f}} \quad \text{in } \Omega,$$
$$\nabla\cdot\mathbf{u} = 0 \quad \text{in } \Omega,$$
$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega,$$

where Ω is a bounded polyhedral domain in $\mathbb{R}^d$, $d = 2$ or 3, and $\tilde{\mathbf{f}} \in L_2(\Omega)^d$. The restriction to polyhedral domains and homogeneous Dirichlet boundary conditions is made only to simplify the exposition.

In contrast to the linear case of Section 3.1, we shall consider here a scaled form of the Navier-Stokes equations that is better suited to the approximation of non-singular branches of solutions to nonlinear problems [BRR80]. Thus set $\tilde{p} = \nu p$, $\tilde{\mathbf{f}} = \nu\mathbf{f}$, $\tilde{\sigma} = \sigma\nu$ and $\lambda = \nu^{-1}$, and the above equations become

$$\sigma\mathbf{u} - \Delta\mathbf{u} + \lambda((\mathbf{u}\cdot\nabla)\mathbf{u}) + \nabla p = \mathbf{f} \qquad \text{in } \Omega, \tag{3.14a}$$
$$\nabla\cdot\mathbf{u} = 0 \qquad \text{in } \Omega, \tag{3.14b}$$
$$\mathbf{u} = \mathbf{0} \qquad \text{on } \partial\Omega . \tag{3.14c}$$

Define the spaces $\mathbf{V}$ and Q by

$$\mathbf{V} := H_0^1(\Omega)^d \quad \text{and} \quad Q := L_0^2(\Omega) := \{q \in L_2(\Omega) : (q, 1) = 0\}.$$

Then the weak formulation of the scaled Navier-Stokes problem (3.14) is:

Find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\sigma(\mathbf{u}, \mathbf{v}) + (\nabla\mathbf{u}, \nabla\mathbf{v}) + \lambda((\mathbf{u}\cdot\nabla)\mathbf{u}, \mathbf{v}) - (p, \nabla\cdot\mathbf{v}) = (\mathbf{f}, \mathbf{v}), \tag{3.15a}$$
$$(q, \nabla\cdot\mathbf{u}) = 0. \tag{3.15b}$$

Let the finite element trial and test spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$ satisfy the interpolation error and inverse inequalities (3.4). We also assume that a discrete Sobolev inequality of type

$$\|\mathbf{v}_h\|_\infty \le c_7\, h^{-\rho}\|\nabla\mathbf{v}_h\|_0 \quad \forall\mathbf{v}_h \in \mathbf{V}_h \tag{3.16}$$

with $\rho > 0$ for $d = 2$ and $\rho = 1/2$ for $d = 3$ holds. Indeed, in [OR79] for piecewise linears on a quasiuniform mesh

$$\|\mathbf{v}_h\|_\infty \le c_7\, |\ln h|^{-1/2}\|\nabla\mathbf{v}_h\|_0 \quad \forall\mathbf{v}_h \in \mathbf{V}_h$$

when $d = 2$ has been proven. The generalization to piecewise polynomials of degree less than or equal to r is given in [Kop98] with a constant c_7 depending on r. Moreover, when replacing h by $h_{min} = \min_{\mathcal{T}_h} h_T$ this inequality is true on any mesh [Kop98]. The three-dimensional case, $d = 3$ is studied in [HR82]

on quasi uniform meshes for conforming and nonconforming finite element spaces.

The SDFEM for (3.14) is obtained by adding to (3.14) both a least-squares control of the divergence and, on each element, a weak form of the momentum equation, using test functions of the form $\lambda(\mathbf{u} \cdot \nabla)\mathbf{v} + \nabla q$:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has

$$\sigma(\mathbf{u}_h, \mathbf{v}_h) + (\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + \lambda((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot \mathbf{u}_h)$$

$$+\delta \sum_{T \in \mathcal{T}_h} h_T^2 (\sigma \mathbf{u}_h - \Delta \mathbf{u}_h + \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla p_h, \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h + \nabla q_h)_T$$

$$+\delta \sum_{E \in \mathcal{E}_h} h_E([p_h]_E, [q_h]_E)_E + \mu(\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h)$$

$$= (\mathbf{f}, \mathbf{v}_h) + \delta \sum_{T \in \mathcal{T}_h} h_T^2 (\mathbf{f}, \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h + \nabla q_h)_T . \qquad (3.17)$$

Here the case $\delta = \mu = 0$ corresponds to the standard finite element method for problem (3.14). Note that, unlike that method, no Babuška-Brezzi condition is imposed on the spaces $\mathbf{V}_h$ and Q_h.

To obtain the unscaled form of (3.17), multiply (3.17) by ν then set $\tilde{p}_h = \nu p_h$, $\tilde{\mathbf{f}} = \nu \mathbf{f}$, $\tilde{\sigma} = \nu \sigma$, $\tilde{\delta} = \lambda \delta$, $\tilde{\mu} = \nu \mu$, $\tilde{\sigma} = \nu \sigma$ and $\tilde{q}_h = \nu q_h$:

Find $(\mathbf{u}_h, \tilde{p}_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has

$$\tilde{\sigma}(\mathbf{u}_h, \mathbf{v}_h) + \nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + ((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h) - (\tilde{p}_h, \nabla \cdot \mathbf{v}_h) + (\tilde{q}_h, \nabla \cdot \mathbf{u}_h)$$

$$+ \tilde{\delta} \sum_{T \in \mathcal{T}_h} h_T^2 (\tilde{\sigma} \mathbf{u}_h - \nu \Delta \mathbf{u}_h + (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla \tilde{p}_h, (\mathbf{u}_h \cdot \nabla)\mathbf{v}_h + \nabla \tilde{q}_h)_T$$

$$+ \tilde{\delta} \sum_{E \in \mathcal{E}_h} h_E([\tilde{p}_h]_E, [\tilde{q}_h]_E)_E + \tilde{\mu}(\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h)$$

$$= (\tilde{\mathbf{f}}, \mathbf{v}_h) + \tilde{\delta} \sum_{T \in \mathcal{T}_h} h_T^2 (\tilde{f}, (\mathbf{u}_h \cdot \nabla)\mathbf{v}_h + \nabla \tilde{q}_h)_T . \qquad (3.18)$$

Comparing (3.18) with its linear analogue (3.6), one observes that in (3.6) we have $\sigma = \tilde{\delta}$ and $\delta_T = \tilde{\delta} h_T^2$. This agrees with the parameter choices discussed in Remarks 3.7 and 3.8. Corresponding to the linear case (Lemma 3.4), assume in (3.17) that $\mu \geq 0$, $\delta > 0$ and that δ satisfies

$$\delta \leq \frac{1}{2} \mu_{\text{inv}}^{-2} \qquad \text{and} \qquad \delta \sigma h_T^2 \leq \frac{1}{2} . \qquad (3.19)$$

Theorem 3.11. *Assume that the finite element spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$ satisfy (3.4) and inequality (3.16), and that δ satisfies (3.19). Assume that for some constant $\tilde{C}$ one has*

$$\lambda h^{1-\rho} \left(\|\mathbf{f}\|_{-1}^2 + \delta \sum_{T \in \mathcal{T}_h} h_T^2 \|\mathbf{f}\|_{0,T}^2 \right)^{1/2} \leq \tilde{C},$$

where ρ is as in inequality (3.16). Then there exists a constant C, which is independent of h and λ, such that the SDFEM problem (3.17) has at least one solution $(\mathbf{u}_h, p_h)$. Moreover, the solution of (3.17) is unique if λ is sufficiently small.

Proof. Define an operator $\mathcal{P} : \mathbf{V}_h \to Q_h$ by

$$\delta \sum_{T \in \mathcal{T}_h} h_T^2 (\nabla \mathcal{P}(\mathbf{u}_h), \nabla q_h)_T + \delta \sum_{E \in \mathcal{E}_h} h_E ([\mathcal{P}(\mathbf{u}_h)]_E, [q_h]_E)_E \tag{3.20}$$

$$= -(q_h, \nabla \cdot \mathbf{u}_h) - \delta \sum_{T \in \mathcal{T}_h} h_T^2 (\sigma \mathbf{u}_h - \Delta \mathbf{u}_h + \lambda (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \mathbf{f}, \nabla q_h)_T$$

for all $\mathbf{u}_h \in \mathbf{V}_h$ and all $q_h \in Q_h$, and an operator $\mathcal{N} : \mathbf{V}_h \to \mathbf{V}_h$ by

$$(\mathcal{N}(\mathbf{u}_h), \mathbf{v}_h) = \sigma(\mathbf{u}_h, \mathbf{v}_h) + (\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + \lambda((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h) \tag{3.21}$$
$$- (\mathcal{P}(\mathbf{u}_h), \nabla \cdot \mathbf{v}_h) - (\mathbf{f}, \mathbf{v}_h) + \mu(\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h)$$
$$+ \delta \sum_{T \in \mathcal{T}_h} h_T^2 (\sigma \mathbf{u}_h - \Delta \mathbf{u}_h + \lambda (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla \mathcal{P}(\mathbf{u}_h) - \mathbf{f}, \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h)_T$$

for all $\mathbf{u}_h$ and $\mathbf{v}_h \in \mathbf{V}_h$. Clearly $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ is a solution of (3.17) if and only if $\mathcal{N}(\mathbf{u}_h) = 0$ and $p_h = \mathcal{P}(\mathbf{u}_h)$.

Let $\mathbf{u}_h \in \mathbf{V}_h$ satisfy $\sigma \|\mathbf{u}_h\|_0^2 + |\mathbf{u}_h|_1^2 = R^2$, where $R > 0$ is arbitrary, and use the abbreviations

$$F := \left[\|\mathbf{f}\|_{-1}^2 + \delta \sum_{T \in \mathcal{T}_h} h_T^2 \|\mathbf{f}\|_{0,T}^2 \right]^{1/2}, \qquad Y := \left[\delta \sum_{E \in \mathcal{E}_h} h_E \|[\mathcal{P}(\mathbf{u}_h)]_E\|_{0,E}^2 \right]^{1/2},$$

$$X := \left[\delta \sum_{T \in \mathcal{T}_h} h_T^2 \|\lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla \mathcal{P}(\mathbf{u}_h)\|_{0,T}^2 \right]^{1/2}.$$

Equations (3.20) and (3.21) and the conditions (3.19) then imply that

$(\mathcal{N}(\mathbf{u}_h), \mathbf{u}_h)$

$$= \sigma\|\mathbf{u}_h\|_0^2 + |\mathbf{u}_h|_1^2 + \lambda((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{u}_h)$$
$$\quad - (\mathcal{P}(\mathbf{u}_h), \nabla \cdot \mathbf{u}_h) - (\mathbf{f}, \mathbf{u}_h) + \mu\|\nabla \cdot \mathbf{u}_h\|_0^2$$
$$\quad + \delta \sum_{T \in \mathcal{T}_h} h_T^2(\sigma\mathbf{u}_h - \Delta\mathbf{u}_h + \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla\mathcal{P}(\mathbf{u}_h) - \mathbf{f} \ , \ \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h)_T$$

$$\geq R^2 - \frac{\lambda}{2}(\nabla \cdot \mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{u}_h) - (\mathbf{f}, \mathbf{u}_h) + Y^2$$
$$+ \delta \sum_{T \in \mathcal{T}_h} h_T^2(\sigma\mathbf{u}_h - \Delta\mathbf{u}_h + \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla\mathcal{P}(\mathbf{u}_h) - \mathbf{f}, \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla\mathcal{P}(\mathbf{u}_h))_T$$

$$\geq R^2 - \frac{\lambda}{2}(\nabla \cdot \mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{u}_h) - \|\mathbf{f}\|_{-1}R + X^2 + Y^2$$
$$- \delta \sum_{T \in \mathcal{T}_h} h_T^2 \left[\sigma\|\mathbf{u}_h\|_{0,T} + \mu_{\text{inv}}h_T^{-1}\|\nabla\mathbf{u}_h\|_{0,T} + \|\mathbf{f}\|_{0,T}\right].$$

$$\|\lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla\mathcal{P}(\mathbf{u}_h)\|_{0,T}$$

$$\geq \frac{1}{4}R^2 + \frac{1}{4}X^2 + Y^2 - F^2 - \frac{\lambda}{2}(\nabla \cdot \mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{u}_h). \tag{3.22}$$

Next, we estimate the term $(\nabla \cdot \mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{u}_h)$. From (3.4), (3.16), (3.20) and the continuity of $\mathbf{u}_h$, one sees that

$$|(\nabla \cdot \mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{u}_h)|$$
$$\leq |(\nabla \cdot \mathbf{u}_h, \mathbf{u}_h \cdot \mathbf{u}_h - J_h(\mathbf{u}_h \cdot \mathbf{u}_h))| + |(\nabla \cdot \mathbf{u}_h, J_h(\mathbf{u}_h \cdot \mathbf{u}_h))|$$
$$\leq d^{1/2}Rc_3h|\mathbf{u}_h \cdot \mathbf{u}_h|_1$$
$$+ \left|\delta \sum_{T \in \mathcal{T}_h} h_T^2(\sigma\mathbf{u}_h - \Delta\mathbf{u}_h + \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla\mathcal{P}(\mathbf{u}_h) - \mathbf{f} \ , \ \nabla J_h(\mathbf{u}_h \cdot \mathbf{u}_h))_T\right|$$
$$+ \left|\delta \sum_{E \in \mathcal{E}_h} h_E([\mathcal{P}(\mathbf{u}_h)]_E \ , \ [\mathbf{u}_h \cdot \mathbf{u}_h - J_h(\mathbf{u}_h \cdot \mathbf{u}_h)]_E)_E\right|$$
$$\leq d^{1/2}c_3c_7h^{1-\rho}R^3 + c_4h|\mathbf{u}_h \cdot \mathbf{u}_h|_1\delta^{1/2}Y$$
$$+ (1 + c_3)h|\mathbf{u}_h \cdot \mathbf{u}_h|_1\{\delta^{1/2}X + \delta^{1/2}F + \delta\mu_{\text{inv}}|\mathbf{u}_h|_1 + \delta\sigma h\|\mathbf{u}_h\|_0\}$$
$$\leq 2Mh^{1-\rho}R^3 + \gamma h^{1-\rho}R^2[X + Y + F], \tag{3.23}$$

where

$$M := \max\left\{\frac{1}{2}c_7[\delta(1 + c_3)(\mu_{\text{inv}} + h\sigma^{1/2}) + d^{1/2}c_3], \ \delta^{1/2}c_7(1 + c_3 + c_4)\right\}.$$

Combining (3.22) and (3.23), we obtain

$$(\mathcal{N}(\mathbf{u}_h), \mathbf{u}_h) \geq \frac{1}{4}R^2 - M\lambda h^{1-\rho}R^3 - M^2\lambda^2h^{2(1-\rho)}R^4 - 2F^2.$$

Now assume that $64M\lambda h^{1-\rho}F \leq 1$ and put $R = \dfrac{1}{8M\lambda h^{1-\rho}}$. Then

$$(\mathcal{N}(\mathbf{u}_h), \mathbf{u}_h) \geq \frac{7}{4096 M^2 \lambda^2 h^{2(1-\rho)}} - 2F^2$$

$$= \frac{1}{4096 M^2 \lambda^2 h^{2(1-\rho)}} \left[7 - 2(64 M \lambda h^{1-\rho} F)^2 \right] > 0 .$$

Using this inequality, a variant of Brouwer's fixed-point theorem implies existence of a $\mathbf{u}_h \in \mathbf{V}_h$ with $\sigma \|\mathbf{u}_h\|_0^2 + |\mathbf{u}_h|_1^2 \leq R^2$ and $\mathcal{N}(\mathbf{u}_h) = 0$. Thus the SDFEM problem (3.17) has at least one solution $(\mathbf{u}_h, p_h)$. The remaining part of the theorem follows from Banach's fixed-point theorem using the same arguments as for standard finite element methods; cf. [GR86].   □

Remark 3.12. If one replaces the term $\lambda((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h)$ in the SDFEM (3.17) by its anti-symmetric analogue

$$\frac{\lambda}{2}[((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h) - ((\mathbf{u}_h \cdot \nabla)\mathbf{v}_h, \mathbf{u}_h)],$$

the hypothesis on $\lambda h^{1-\rho}$ in Theorem 3.11 can be dropped.   ♣

Theorem 1.1 informs us that the problem (3.15) has at least one solution, which is unique provided λ is sufficiently small, and that it can be written in the operator form

$$F(\lambda, \mathbf{u}_\lambda, p_\lambda) := (\mathbf{u}_\lambda, p_\lambda) + TG(\lambda, \mathbf{u}_\lambda, p_\lambda) = 0. \qquad (3.24)$$

Here the linear functional $G(\lambda, \mathbf{u}, p) \in \mathbf{V}^* \times Q$ is defined by

$$< G(\lambda, \mathbf{u}, p), (\mathbf{v}, q) > := \lambda((\mathbf{u} \cdot \nabla)\mathbf{u}, \mathbf{v}) - (\mathbf{f}, \mathbf{v})$$

for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$, and the generalized Stokes operator $T : \mathbf{V}^* \times Q \to \mathbf{V} \times Q$ associates with each $(\mathbf{w}, r) \in \mathbf{V}^* \times Q$ the unique solution $(\mathbf{u}, p) = T(\mathbf{w}, r)$ in $\mathbf{V} \times Q$ of

$$\sigma(\mathbf{u}, \mathbf{v}) + (\nabla \mathbf{u}, \nabla \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) = < \mathbf{w}, \mathbf{v} >,$$
$$(q, \nabla \cdot \mathbf{u}) = (r, q),$$

for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$.

In what follows, we assume that one has a compact interval $\Lambda \subset \mathbb{R}$ and a continuous branch $\lambda \mapsto (\mathbf{u}_\lambda, p_\lambda)$ of solutions of the problem (3.24) that is regular, in the sense that for each $\lambda \in \Lambda$ the Fréchet derivative $D_{(\mathbf{u},p)}F(\lambda, \mathbf{u}_\lambda, p_\lambda)$ of $F(\lambda, \cdot, \cdot)$ at $(\mathbf{u}_\lambda, p_\lambda)$ is a homeomorphism of $\mathbf{V} \times Q$ onto itself. We shall prove that the SDFEM problem (3.17) has a unique solution $(\mathbf{u}_{h,\lambda}, p_{h,\lambda})$ in a neighbourhood of the solution branch $(\mathbf{u}_\lambda, p_\lambda)$ and derive error estimates for $(\mathbf{u}_\lambda - \mathbf{u}_{h,\lambda}, p_\lambda - p_{h,\lambda})$. The analysis follows [BRR80] for the approximation of non-singular branches of solutions to nonlinear problems and [GR86] for standard finite element methods applied to the Navier-Stokes equations.

Let $T_h : L_2(\Omega)^d \times Q \to \mathbf{V}_h \times Q_h$ be the discrete Stokes operator that associates with each $(\mathbf{w}, r) \in L_2(\Omega)^d \times Q$ the unique solution $(\mathbf{u}_h, p_h) = T_h((\mathbf{w}, r))$ in $\mathbf{V}_h \times Q_h$ of

$$\sigma(\mathbf{u}_h, \mathbf{v}_h) + (\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot \mathbf{u}_h)$$

$$+ \delta \sum_{T \in \mathcal{T}_h} h_T^2 (\sigma \mathbf{u}_h - \Delta \mathbf{u}_h + \nabla p_h, \nabla q_h)_T$$

$$+ \delta \sum_{E \in \mathcal{E}_h} h_E([p_h]_E, [q_h]_E)_E + \mu(\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h)$$

$$= (\mathbf{w}, \mathbf{v}_h) + \delta \sum_{T \in \mathcal{T}_h} h_T^2 (\mathbf{w}, \nabla q_h)_T + (r, q_h) \quad \text{for all } (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h.$$

Define the operator $G_h : \Lambda \times \mathbf{V}_h \times Q_h \to \mathbf{V}_h \times Q_h$ by

$$(G_h(\lambda, \mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)) := < G(\lambda, \mathbf{u}_h, p_h), (\mathbf{v}_h, q_h) >$$

$$+ \delta \sum_{T \in \mathcal{T}_h} h_T^2 (-\mathbf{f} + \sigma \mathbf{u}_h - \Delta \mathbf{u}_h + \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla p_h, \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h)_T$$

$$- \delta^2 \sum_{T \in \mathcal{T}_h} h_T^4 (-\mathbf{f} + \sigma \mathbf{u}_h - \Delta \mathbf{u}_h + \lambda(\mathbf{u}_h \cdot \nabla)\mathbf{u}_h + \nabla p_h, \lambda(\mathbf{u}_h \cdot \nabla)\nabla q_h)_T \quad (3.25)$$

for all $(\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$. Using these operators, the SDFEM problem (3.17) can be written, analogously to (3.24), as

$$F_h(\lambda, \mathbf{u}_{h,\lambda}, p_{h,\lambda}) := (\mathbf{u}_{h,\lambda}, p_{h,\lambda}) + T_h G_h(\lambda, \mathbf{u}_{h,\lambda}, p_{h,\lambda}) = 0. \quad (3.26)$$

To formulate our main result, introduce the mesh-dependent norm

$$|(\mathbf{v}, q)|_h := \left(\sigma \|\mathbf{v}\|_0^2 + |\mathbf{v}|_1^2 + \|q\|_0^2 \right.$$

$$\left. + \delta \sum_{T \in \mathcal{T}_h} h_T^2 |q|_{1,T}^2 + \delta \sum_{E \in \mathcal{E}_h} h_E \|[q]_E\|_{0,E}^2 \right)^{1/2}.$$

The assumption $\lambda \in \Lambda$, where Λ is compact, implies that ν is bounded away from zero, so we do not consider the behaviour of (3.17) as $\nu \to 0$. Thus $|(\cdot, \cdot)|_h$ resembles the norm $|||(\cdot, \cdot)|||_h$ that was used in Section 3.1.

Lemma 3.13. *[TV96] Assume that $(\mathbf{u}_\lambda, p_\lambda) \in H^{k+1}(\Omega)^d \times H^k(\Omega)$ for some $k \geq 1$. Then there exists a constant C, which is independent of h and λ, such that*

$$|F_h(\lambda, I_h \mathbf{u}_\lambda, J_h p_\lambda)|_h \leq Ch^k \{K_k + \lambda K_k^2 + \lambda^2 h^2 K_k^3\}, \quad (3.27a)$$

$$\|D_{(\mathbf{u},p)} F(\lambda, \mathbf{u}_\lambda, p_\lambda) - D_{(\mathbf{u},p)} F(\lambda, I_h \mathbf{u}_\lambda, J_h p_\lambda)\|_{\mathcal{L}(\mathbf{V} \times Q)} \leq C \lambda h K_1, \quad (3.27b)$$

$$\|D_{(\mathbf{u},p)} F_h(\lambda, I_h \mathbf{u}_\lambda, J_h p_\lambda) - D_{(\mathbf{u},p)} F(\lambda, I_h \mathbf{u}_\lambda, J_h p_\lambda)\|_{\mathcal{L}(\mathbf{V}_h \times Q_h)}$$
$$\leq C \lambda h K_1 \{1 + \lambda h K_1\}, \quad (3.27c)$$

$$\|D_{(\mathbf{u},p)} F_h(\lambda, \mathbf{u}_h, p_h) - D_{(\mathbf{u},p)} F_h(\lambda, v_h, q_h)\|_{\mathcal{L}(\mathbf{V}_h \times Q_h)}$$
$$\leq C \lambda |(\mathbf{u}_h - \mathbf{v}_h, p_h - q_h)|_h, \quad (3.27d)$$

where

$$K_l := \sup_{\lambda \in \Lambda} \max\{\|\mathbf{f}\|_{l-1}, \|\mathbf{u}_\lambda\|_{l+1}, \|p_\lambda\|_l\} \quad \text{for } 1 \leq l \leq k .$$

In this lemma the operator norms $\| \cdot \|_{\mathcal{L}(\mathbf{V} \times Q)}$ and $\| \cdot \|_{\mathcal{L}(\mathbf{V}_h \times Q_h)}$ are induced by the norm $|(\mathbf{v}, q)| = (|\mathbf{v}|_1^2 + \|q\|_0^2)^{1/2}$ on $\mathbf{V} \times Q$ and by our mesh-dependent norm $|(\cdot, \cdot)|_h$ on $\mathbf{V}_h \times Q_h$.

Theorem 3.14. *Let $\Lambda \subset \mathbb{R}$ be a given compact interval. For $\lambda \in \Lambda$, assume that the problem (3.24) has a regular branch of solutions $\lambda \mapsto (\mathbf{u}_\lambda, p_\lambda)$ and that $(\mathbf{u}_\lambda, p_\lambda) \in H^{k+1}(\Omega)^d \times H^k(\Omega)$ for some $k \geq 1$. Then there is a positive constant $h_0(\Lambda)$ such that for all $h \in (0, h_0(\Lambda)]$ the problem (3.17) has a unique branch of solutions $\lambda \mapsto (\mathbf{u}_{h,\lambda}, p_{h,\lambda})$ in a neighbourhood of $(\mathbf{u}_\lambda, p_\lambda)$. Moreover, the error estimate*

$$\sup_{\lambda \in \Lambda} |(\mathbf{u}_{h,\lambda} - \mathbf{u}_\lambda \,, p_{h,\lambda} - p_\lambda)|_h \leq M(K, \Lambda)\, h^k \tag{3.28}$$

holds true, where $K = \max_{1 \leq l \leq k} K_l$.

Proof. From (3.27b) and (3.34) one can deduce that $D_{(\mathbf{u},p)} F_h(\lambda, I_h \mathbf{u}_\lambda, J_h p_\lambda)$, the Fréchet derivative of $F_h(\lambda, \cdot, \cdot)$ at $(I_h \mathbf{u}_\lambda, J_h p_\lambda)$, is a homeomorphism of $\mathbf{V}_h \times Q_h$ onto itself for each $\lambda \in \Lambda$, provided that $h \sup_{\lambda \in \Lambda} \lambda$ is sufficiently small. Thus (3.26) can be recast as a fixed-point equation for the operator Φ defined by

$$\Phi(\mathbf{u}_{h,\lambda}, p_{h,\lambda}) := (\mathbf{u}_{h,\lambda}, p_{h,\lambda}) + D_{(\mathbf{u},p)} F_h(\lambda, I_h \mathbf{u}_\lambda, J_h p_\lambda)^{-1}\, F_h(\lambda, \mathbf{u}_{h,\lambda}, p_{h,\lambda}).$$

The bounds (3.27a) and (3.27d) imply that $\Phi : \mathbf{V}_h \times Q_h \to \mathbf{V}_h \times Q_h$ is contractive and maps the ball $B((I_h \mathbf{u}_{h,\lambda}, J_h p_{h,\lambda}), R)$ of radius $R(h, \Lambda) = \mathcal{O}(h^k)$ into itself. Hence, by Banach's fixed-point theorem, (3.26) has a unique solution $(\mathbf{u}_{h,\lambda}, p_{h,\lambda})$ in this ball. The triangle inequality

$$|(\mathbf{u}_{h,\lambda} - \mathbf{u}_\lambda \,, p_{h,\lambda} - p_\lambda)|_h$$
$$\leq |(\mathbf{u}_{h,\lambda} - I_h \mathbf{u}_\lambda \,, p_{h,\lambda} - J_h p_\lambda)|_h + |(I_h \mathbf{u}_\lambda - \mathbf{u}_\lambda \,, J_h p_\lambda - p_\lambda)|_h$$

and the estimates (3.4a)–(3.4d) then yield (3.28).   □

Remark 3.15. For fixed λ, instead of (3.28) one obtains the estimate

$$|(\mathbf{u}_{h,\lambda} - \mathbf{u}_\lambda \,, p_{h,\lambda} - p_\lambda)|_h \leq M(K, \lambda)\, h^k.$$

Now $M(K, \lambda) \sim \|D_{(\mathbf{u},p)} F(\lambda, \mathbf{u}_\lambda, p_\lambda)^{-1}\| \lambda^2 K^3$, but unfortunately the behaviour of $\|D_{(\mathbf{u},p)} F(\lambda, \mathbf{u}_\lambda, p_\lambda)^{-1}\|$ as a function of $\lambda = 1/\nu$ is in general unknown. The simple one-dimensional Example 3.16 shows that in the worst case $\|D_{(\mathbf{u},p)} F(\lambda, \mathbf{u}_\lambda, p_\lambda)^{-1}\|$ may depend exponentially on λ. On the other hand, results of [JR94, JRB95a, JRB95b] show that – for certain classes of

flow problems and particular perturbations – the norm of the solution of the linearized problem may depend only linearly on λ. This operator norm was estimated in [Tob81] for a class of scalar convection-diffusion equations, using the maximum principle in a framework of Hölder spaces; in the worst case a linear dependence on λ was proved.

Example 3.16. Consider Burgers' equation, which is the one-dimensional version of the Navier-Stokes equations:

$$-\nu u'' + uu' = f \quad \text{in } (-1,1), \quad u(-1) = 1, \ u(1) = -1.$$

If $f \equiv 0$ then the exact solution of this problem is $u(x) = -2\nu\alpha_\nu \tanh(\alpha_\nu x)$, where α_ν is the unique positive solution of $2\nu\alpha_\nu \tanh(\alpha_\nu) = 1$. When linearized about u, the original problem becomes

$$Lv := -\nu v'' + uv' + vu' = f \quad \text{in } (-1,1), \quad v(-1) = v(1) = 0.$$

This has the unique solution

$$v(x) = -\lambda e^{\lambda U(x)} \int_{-1}^{x} e^{-\lambda U(t)} \left(\int_{0}^{t} f(s)ds + c \right) dt,$$

where U is a primitive of u, c is determined by the condition $v(1) = 0$, and $\lambda = 1/\nu$. For the particular choice

$$f(s) = \cosh(\alpha_\nu s)$$

the solution is
$$v(x) = \frac{\cosh^3(\alpha_\nu) - \cosh^3(\alpha_\nu x)}{3\nu\alpha_\nu^2 \cosh^2(\alpha_\nu x)}.$$

Now $L : H_0^1 \to H^{-1}$, with

$$\|L^{-1}\| = \sup_{g \in H^{-1}} \frac{|L^{-1}g|_1}{\|g\|_{-1}} \geq \frac{|v|_1}{\|f\|_{-1}}.$$

But $\|v\|_\infty \leq \sqrt{2}\,|v|_1$ and $\|f\|_{-1} \leq \sqrt{2}\,\|f\|_\infty$, so

$$\|L^{-1}\| \geq \frac{\|v\|_\infty}{2\|f\|_\infty} \geq \frac{\sinh^2 \alpha_\nu}{6\nu\alpha_\nu^2}.$$

Hence, in the most interesting and challenging case when $\lambda \gg 1$, the lower bound behaves like $2e^\lambda/(3\lambda)$, i.e., it grows exponentially in λ. ♣

Remark 3.17. If $\lambda((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h)$ in the SDFEM (3.17) is replaced by its anti-symmetric analogue

$$\frac{\lambda}{2}[((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h) - ((\mathbf{u}_h \cdot \nabla)\mathbf{v}_h, \mathbf{u}_h)],$$

then the condition that $\lambda h^{1-\rho}$ be small can be omitted from the existence result of Theorem 3.11. This does however introduce an additional term

$$\frac{\lambda}{2}\delta \sum_{T\in\mathcal{T}_h} h_T^2 \big[((\mathbf{u}_h \cdot \nabla)\nabla q_h, \mathbf{u}_h)_T + ((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \nabla q_h)_T\big]$$

into the operator G_h of (3.25) and an additional term $\lambda h^{d/2}K_1^2$ in (3.27a). Theorem 3.14 consequently gives only an $\mathcal{O}(h^{d/2})$ error estimate. This order of convergence cannot be improved by assuming higher-order regularity of the solution of problem (3.15), so there is no point in using high-order finite element spaces. Note that these additional terms vanish when using piecewise constant approximations of the pressure. A more detailed analysis of the additional terms in the case of piecewise linear pressure approximations yields an $\mathcal{O}(h^{2-\kappa})$ error estimate, where $\kappa > 0$ can be arbitrarily small if $d = 2$, and $\kappa = 0$ if $d = 3$.   ♣

4

Local Projection Stabilization for Equal-Order Interpolation

Local projection stabilization (LPS) was introduced in Part III, Chapter 3 for a scalar convection-diffusion equation. It will now be extended to the Oseen system

$$-\nu \Delta \mathbf{u} + (\mathbf{b} \cdot \nabla)\mathbf{u} + \sigma \mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega \subset \mathbb{R}^d, \tag{4.1a}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{4.1b}$$

$$\mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega. \tag{4.1c}$$

As we saw in Chapter 3, the streamline diffusion method (SDFEM) can handle two types of instabilities: that caused by a violation of the discrete inf-sup (Babuška-Brezzi) condition (2.1) and that due to dominant convection. The SDFEM combines the Pressure Stabilized Petrov-Galerkin (PSPG) approach (testing the residual against ∇q) with the Streamline Upwind Petrov-Galerkin (SUPG) technique (testing the residual against $(\mathbf{b} \cdot \nabla)\mathbf{v}$); see for example [BH82, FF92, HFB86, TBA$^+$92].

Despite the extensive theoretical and practical development of the SD-FEM, a fundamental flaw in the method – in particular for higher-order interpolations – is that various terms must be added to the weak formulation to guarantee its consistency. Moreover, the requirement of consistency leads to undesirable effects when using residual-based stabilization methods like the SDFEM in optimal control problems; see the discussion at the beginning of Section III.3.3. LPS relaxes the consistency requirement while preserving the main features of the SDFEM approach; in particular, one can use equal-order interpolation without worrying about the Babuška-Brezzi condition. Furthermore, LPS allows us to separate velocity and pressure in the stabilization terms, which for systems of equations means that one can avoid non-physical couplings.

Stabilization by local projection is introduced for the Stokes problem in [BB01], extended to the transport equation in [BB04], and analysed for low-order discretizations of the Oseen equations in [BB06]. Some variants and applications are discussed in [BR06a, BR06b]. The method has been sucessfully

used in many different areas; see for example [EAE06, BB01, BB06, BR06a, BR06b, BR07, BV07, EG04, GMQ06, Gue99a, Gue99b, Gue01a, Gue01b, JK06a, JKL06, KR05].

In LPS, one uses a projection $\pi_h : Y_h \rightarrow D_h$ from the finite element space Y_h, which approximates velocity and pressure, into a discontinuous space D_h. Stabilization of the standard Galerkin method is then achieved by adding terms that give a weighted L^2 control over the fluctuations $(\mathrm{id} - \pi_h)$ of the gradients of the quantity of interest. In the error analysis of the LPS, the key idea – as expounded in Section III.3.3.1 – is the construction of an interpolant operator that maps into Y_h and possesses a particular orthogonality property with respect to the discontinuous space D_h.

Our exposition here begins by recalling in Section 4.1 the weak formulation of the Oseen equations, its standard Galerkin discretization and LPS in an abstract setting. In the same setting, Section 4.2 is devoted to the convergence analysis of LPS; a special interpolant will be constructed that satisfies a local inf-sup condition. On proving the independence of its stability from the Reynolds number and an approximate Galerkin orthogonality identity, we deduce optimal *a priori* error estimates. The application of this theory in a framework of two-level methods is studied in Section 4.3, where the focus is on defining pairs of finite element spaces that satisfy the local inf-sup condition of Section 4.2. The analysis is extended in Section 4.4 to spaces that are defined on the same mesh; one begins from the space D_h then constructs the space Y_h by enriching standard finite element spaces. It is well known that stabilized methods can also be derived in a variational multiscale framework [HS07, Hug95, Tob06] – a scale separation of the underlying finite element spaces shows that it is sufficient to stabilize only the fine scale fluctuations. This produces a stabilizing term that gives a weighted L^2 control over the gradients of fluctuations instead of the fluctuations of gradients [EG04, Gue99a]. The relation between this subgrid modelling approach and LPS will be discussed in Section 4.5.

4.1 Local Projection Stabilization in an Abstract Setting

We briefly recall the weak formulation and unique solvability of the Oseen problem (4.1). Assume that $\mathbf{b} \in \mathbf{W}^{1,\infty}(\Omega)$ with $\nabla \cdot \mathbf{b} = 0$. Set $\mathbf{V} = \mathbf{H}_0^1(\Omega)$ and $Q = L_0^2(\Omega) = \{q \in L^2(\Omega) : (q,1) = 0\}$. On the product space $\mathbf{V} \times Q$, introduce the bilinear form

$$A\big((\mathbf{u},p);(\mathbf{v},q)\big) := \nu(\nabla\mathbf{u},\nabla\mathbf{v}) + \big((\mathbf{b}\cdot\nabla)\mathbf{u},\mathbf{v}\big) + \sigma(\mathbf{u},\mathbf{v})$$
$$-(p,\nabla\cdot\mathbf{v}) + (q,\nabla\cdot\mathbf{u}).$$

Then a weak formulation of the Oseen problem (4.1) is:

Find $(\mathbf{u},p) \in \mathbf{V} \times Q$ such that

$$A\big((\mathbf{u}, p); (\mathbf{v}, q)\big) = (\mathbf{f}, \mathbf{v}) \qquad \forall (\mathbf{v}, q) \in \mathbf{V} \times Q. \tag{4.2}$$

Recall from Theorem 1.5 that (4.2) has a unique solution for all positive ν.

The LPS is based on two finite element spaces, the approximation space Y_h and the projection space D_h, which can be defined on different families of meshes. Here the projection space D_h is associated with a shape-regular decomposition $\mathcal{M}_h$ of Ω into macro-elements $M \in \mathcal{M}_h$. Then we use certain refinement rules to generate from the family of macro-elements $\mathcal{M}_h$ a family of shape-regular decompositions $T \in \mathcal{T}_h$, in such a way that for each cell $T \in \mathcal{T}_h$ there is a macro-element $M \in \mathcal{M}_h$ with $T \subset M$ and

$$h_T \sim h_M \qquad \forall T \subset M, \quad \forall M \in \mathcal{M}_h.$$

The spaces used for the approximation of the velocity $\mathbf{u} \in \mathbf{V}$ and the pressure $p \in Q$ will be based on the decomposition $\mathcal{T}_h$. Our analysis allows the possibility that $\mathcal{M}_h = \mathcal{T}_h$.

Let $Y_h \subset H^1(\Omega)$ be a finite element space of continuous piecewise polynomial functions defined over $\mathcal{T}_h$. For simplicity of presentation we consider the case of equal-order interpolation, so the velocity and pressure are approximated by the respective spaces $\mathbf{V}_h := Y_h^d \cap \mathbf{V}$ and $Q_h := Y_h \cap Q$. Let D_h be a discontinuous finite element space defined on the macro-decomposition $\mathcal{M}_h$ and set $D_h(M) = \{q_h\big|_M \ : \ q_h \in D_h\}$. Let $\pi_M : L^2(M) \to D_h(M)$ be a local projection then define the projection $\pi_h : L^2(\Omega) \to D_h$ by $(\pi_h w)\big|_M := \pi_M\big(w\big|_M\big)$. The fluctuation operator $\kappa_h : L^2(\Omega) \to L^2(\Omega)$ associated with the projection π_h is $\kappa_h := \mathrm{id} - \pi_h$, where $\mathrm{id} : L^2(\Omega) \to L^2(\Omega)$ is the identity mapping. These operators will be applied to vector-valued functions component by component, and this usage is indicated by boldface notation, e.g., $\boldsymbol{\pi}_h : \mathbf{L}^2(\Omega) \to \mathbf{D}_h$ and $\boldsymbol{\kappa}_h : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$.

Our stabilized scheme is:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has

$$A\big((\mathbf{u}_h, p_h); (\mathbf{v}_h, q_h)\big) + S_h\big((\mathbf{u}_h, p); (\mathbf{v}_h, q_h)\big) = (\mathbf{f}, \mathbf{v}_h) \tag{4.3a}$$

where the stabilization term is

$$S_h\big((\mathbf{u}_h, p_h); (\mathbf{v}_h, q_h)\big) := \sum_{M \in \mathcal{M}_h} \Big(\tau_M(\boldsymbol{\kappa}_h(\mathbf{b} \cdot \nabla)\mathbf{u}_h, \boldsymbol{\kappa}_h(\mathbf{b} \cdot \nabla)\mathbf{v}_h)_M$$

$$+ \mu_M(\kappa_h \nabla \cdot \mathbf{u}_h, \kappa_h \nabla \cdot \mathbf{v}_h)_M + \alpha_M(\boldsymbol{\kappa}_h \nabla p_h, \boldsymbol{\kappa}_h \nabla q_h)_M\Big), \tag{4.3b}$$

and the user-chosen constants τ_M, μ_M, and α_M are yet to be specified – an optimal mesh-dependent choice will follow from the error analysis of the method. Existence, uniqueness, and convergence properties of discrete solutions $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ will be studied in the next section.

4.2 Convergence Analysis

4.2.1 The Special Interpolant

In Section III.3.3.1 we saw that the crucial ingredient in the error analysis of the LPS applied to convection-diffusion equations is the existence of an interpolant $j_h : H^2(\Omega) \to Y_h$ such that the error $w - j_h w$ is L^2-orthogonal to D_h and the standard approximation properties hold true. In Theorem III.3.71 the existence of such an interpolation operator was established provided that one has existence of an interpolation operator $i_h : H^2(\Omega) \to Y_h$ with the desired approximation properties and the pair (Y_h, D_h) satisfies the local inf-sup condition (3.90). For the stability of the LPS method (4.3) we need interpolation operators $j_h : H^1(\Omega) \to Y_h$ that are defined on the larger space $H^1(\Omega)$, have the standard approximation properties, and exhibit the orthogonality $(w - j_h w) \perp D_h$ in $L^2(\Omega)$. The first two requirements can be satisfied by employing interpolation operators for non-smooth functions such as the Scott-Zhang operator of Remark III.3.1. Modifying this operator as in the proof of Theorem III.3.71 yields a interpolation operator j_h that satisfies all three requirements. When applied to vector-valued functions component by component, we indicate this by using boldface notation, e.g., $\mathbf{j}_h : \mathbf{V} \to Y_h^d \cap \mathbf{V}$. Thus for an approximation space Y_h that contains piecewise polynomials of degree $r \in \mathbb{N}$, one can assume the existence of interpolation operators $j_h : H^1(\Omega) \to Y_h$ and $\mathbf{j}_h : \mathbf{V} \to \mathbf{V}_h$ that satisfy the orthogonality and approximation properties

$$(w - j_h w, q_h) = 0 \qquad \forall q_h \in D_h,\ \forall w \in H^1(\Omega), \tag{4.4a}$$

$$(\mathbf{w} - \mathbf{j}_h \mathbf{w}, \mathbf{q}_h) = 0 \qquad \forall \mathbf{q}_h \in \mathbf{D}_h,\ \forall \mathbf{w} \in \mathbf{V}. \tag{4.4b}$$

For all $w \in H^l(\Omega)$, $1 \le l \le r+1$, $\forall M \in \mathcal{M}_h$, one has

$$\|w - j_h w\|_{0,M} + h_M |w - j_h w|_{1,M} \le C\, h_M^l \|w\|_{l,\Lambda(M)}.$$

For all $\mathbf{w} \in \mathbf{V} \cap \mathbf{H}^l(\Omega)$, $1 \le l \le r+1$, $\forall M \in \mathcal{M}_h$ one has

$$\|\mathbf{w} - \mathbf{j}_h \mathbf{w}\|_{0,M} + h_M |\mathbf{w} - \mathbf{j}_h \mathbf{w}|_{1,M} \le C\, h_M^l \|\mathbf{w}\|_{l,\Lambda(M)}. \tag{4.5}$$

Here $\Lambda(M) := \cup_{T \in M} \omega(T)$ is a local neighbourhood of M that is generated from the local neighbourhoods $\omega(T)$ that appear in the interpolation error estimates (3.5) of the Scott-Zhang operator; see Remark 3.1 and [Ape99, Clé75, SZ90] for more details.

Remark 4.1. On setting $q_h = 1$ in (4.4a), it follows that $(j_h w, 1) = (w, 1)$ for all $w \in H^1(\Omega)$. Thus $j_h : H^1(\Omega) \cap Q \to Q_h$. ♣

4.2.2 Stability

On the product space $\mathbf{V} \times Q$, define the mesh-dependent norm

$$|||(\mathbf{v}, q)||| := \left(\nu |\mathbf{v}|_1^2 + \sigma \|\mathbf{v}\|_0^2 + (\nu + \sigma) \|q\|_0^2 + S_h\big((\mathbf{v}, q); (\mathbf{v}, q)\big) \right)^{1/2}. \quad (4.6)$$

Lemma 4.2. *Assume that* $\max\{\nu, \sigma, \tau_M, \mu_M, h_M^2/\alpha_M\} \leq C$ *for all* $M \in \mathcal{M}_h$ *and that there are interpolation operators* $j_h : H^1(\Omega) \to Y_h$ *and* $\mathbf{j}_h : \mathbf{V} \to \mathbf{V}_h$ *satisfying* (4.4)–(4.5). *Then there is a positive constant* β_2, *which is independent of* ν *and* h, *such that*

$$\inf_{(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{(A + S_h)\big((\mathbf{v}_h, q_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{v}_h, q_h)||| \; |||(\mathbf{w}_h, r_h)|||} \geq \beta_2 > 0. \quad (4.7)$$

Proof. Consider an arbitrary pair $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$. Choosing $(\mathbf{w}_h, r_h) = (\mathbf{v}_h, q_h)$, one has

$$(A + S_h)\big((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\big) = \nu |\mathbf{v}_h|_1^2 + \sigma \|\mathbf{v}_h\|_0^2 + S_h\big((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\big)$$

since $\nabla \cdot \mathbf{b} = 0$ and $\mathbf{v}_h = 0$ on $\partial\Omega$.

Now we make a different choice of $(\mathbf{w}_h, r_h)$ to get some L^2-norm control over the pressure. For each $q_h \in Q_h$, the continuous Babuška-Brezzi condition guarantees the existence of a function $\mathbf{v}_{q_h} \in \mathbf{V}$ such that

$$(\nabla \cdot \mathbf{v}_{q_h}, q_h) = -(q_h, q_h) \quad \text{and} \quad \|\mathbf{v}_{q_h}\|_1 \leq C \|q_h\|_0. \quad (4.8)$$

Choose $(\mathbf{w}_h, r_h) = (\mathbf{j}_h \mathbf{v}_{q_h}, 0)$ where $\mathbf{j}_h$ satisfies (4.4b) and (4.5). This yields

$$A\big((\mathbf{v}_h, q_h); (\mathbf{j}_h \mathbf{v}_{q_h}, 0)\big) = \|q_h\|_0^2 - (q_h, \nabla \cdot (\mathbf{j}_h \mathbf{v}_{q_h} - \mathbf{v}_{q_h})) + ((\mathbf{b} \cdot \nabla)\mathbf{v}_h, \mathbf{j}_h \mathbf{v}_{q_h})$$
$$+ \nu(\nabla \mathbf{v}_h, \nabla \mathbf{j}_h \mathbf{v}_{q_h}) + \sigma(\mathbf{v}_h, \mathbf{j}_h \mathbf{v}_{q_h}). \quad (4.9)$$

We estimate the last four terms on the right-hand side. Integrating the first by parts gives

$$-\big(q_h, \nabla \cdot (\mathbf{j}_h \mathbf{v}_{q_h} - \mathbf{v}_{q_h})\big) = \big(\nabla q_h, \mathbf{j}_h \mathbf{v}_{q_h} - \mathbf{v}_{q_h}\big) = \big(\boldsymbol{\kappa}_h \nabla q_h, \mathbf{j}_h \mathbf{v}_{q_h} - \mathbf{v}_{q_h}\big),$$

so

$$\left| \big(q_h, \nabla \cdot (\mathbf{j}_h \mathbf{v}_{q_h} - \mathbf{v}_{q_h})\big) \right| \leq \left(\sum_{M \in \mathcal{M}_h} \alpha_M \|\boldsymbol{\kappa}_h \nabla q_h\|_{0,M}^2 \right)^{1/2} \cdot$$

$$\left(\sum_{M \in \mathcal{M}_h} \frac{1}{\alpha_M} \|\mathbf{j}_h \mathbf{v}_{q_h} - \mathbf{v}_{q_h}\|_{0,M}^2 \right)^{1/2}$$

$$\leq C \left[S_h\big((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\big) \right]^{1/2} \|\mathbf{v}_{q_h}\|_1$$

$$\leq C \left[S_h\big((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\big) \right]^{1/2} \|q_h\|_0$$

$$\leq \frac{\|q_h\|_0^2}{8} + C \, S_h\big((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\big). \quad (4.10)$$

Integrating the third term in (4.9) by parts, invoking the H^1 stability of $\mathbf{j}_h$ which follows from (4.5) for $l = 1$, and taking (4.8) into account, we obtain

$$|((\mathbf{b} \cdot \nabla)\mathbf{v}_h, \mathbf{j}_h\mathbf{v}_{q_h})| = |(\mathbf{v}_h, (\mathbf{b} \cdot \nabla)\mathbf{j}_h\mathbf{v}_{q_h})| \leq C \|\mathbf{v}_h\|_0 |\mathbf{j}_h\mathbf{v}_{q_h}|_1$$

$$\leq \frac{\|q_h\|_0^2}{8} + C \|\mathbf{v}_h\|_0^2. \tag{4.11}$$

To estimate the remaining terms in (4.9), use $\max\{\nu, \sigma\} \leq C$ to get

$$\left|\nu(\nabla\mathbf{v}_h, \nabla\mathbf{j}_h\mathbf{v}_{q_h}) + \sigma(\mathbf{v}_h, \mathbf{j}_h\mathbf{v}_{q_h})\right| \leq \left(\nu|\mathbf{v}_h|_1 + \sigma\|\mathbf{v}_h\|_0\right)\|\mathbf{j}_h\mathbf{v}_{q_h}\|_1$$

$$\leq C\left(\nu^{1/2}|\mathbf{v}_h|_1 + \sigma^{1/2}\|\mathbf{v}_h\|_0\right)\|q_h\|_0$$

$$\leq \frac{\|q_h\|_0^2}{8} + C\left(\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2\right).$$

The Cauchy-Schwarz inequality and the L^2 stability of $\boldsymbol{\kappa}_h$ give

$$\left|S_h\left((\mathbf{v}_h, q_h); (\mathbf{j}_h\mathbf{v}_{q_h}, 0)\right)\right| \leq C \left(S_h\left((\mathbf{v}_h, 0); (\mathbf{v}_h, 0)\right)\right)^{1/2} |\mathbf{j}_h\mathbf{v}_{q_h}|_1$$

$$\leq C \left(S_h\left((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\right)\right)^{1/2} \|q_h\|_0$$

$$\leq \frac{\|q_h\|_0^2}{8} + C\, S_h\left((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\right). \tag{4.12}$$

Let

$$X := \left(\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + S_h\left((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\right)\right)^{1/2}$$

denote the part of the triple norm without L^2 control over the pressure. Combining (4.10)–(4.12) with (4.9), one has

$$(A + S_h)\left((\mathbf{v}_h, q_h); (\mathbf{j}_h\mathbf{v}_{q_h}, 0)\right) \geq \frac{\|q_h\|_0^2}{2} - C\, X^2 - C \|\mathbf{v}_h\|_0^2. \tag{4.13}$$

Now multiply (4.13) by $2(\nu + \sigma)$ and invoke Poincaré's inequality to get

$$2(\nu + \sigma)\|\mathbf{v}_h\|_0^2 \leq C\left(\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2\right).$$

Hence

$$(A + S_h)\left((\mathbf{v}_h, q_h); 2(\nu + \sigma)(\mathbf{j}_h\mathbf{v}_{q_h}, 0)\right) \geq (\nu + \sigma)\|q_h\|_0^2 - C_1\, X^2$$

with a suitable constant C_1.

For an arbitrary $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$, choose

$$(\mathbf{w}_h, r_h) := (\mathbf{v}_h, q_h) + \frac{2(\nu + \sigma)}{1 + C_1}(\mathbf{j}_h\mathbf{v}_{q_h}, 0) \in \mathbf{V}_h \times Q_h.$$

Then

$$(A + S_h)\big((\mathbf{v}_h, q_h); (\mathbf{w}_h, r_h)\big) \geq \frac{(\nu + \sigma)}{1 + C_1} \|q_h\|_0^2 + \left(1 - \frac{C_1}{1 + C_1}\right) X^2$$

$$= \frac{1}{1 + C_1} |||(\mathbf{v}_h, q_h)|||^2 \qquad (4.14)$$

and

$$|||(\mathbf{w}_h, r_h)||| \leq |||(\mathbf{v}_h, q_h)||| + \frac{2(\nu + \sigma)}{1 + C_1} |||(\mathbf{j}_h \mathbf{v}_{q_h}, 0)|||$$

$$\leq |||(\mathbf{v}_h, q_h)||| + C(\nu + \sigma)\|\mathbf{j}_h \mathbf{v}_{q_h}\|_1$$

$$\leq |||(\mathbf{v}_h, q_h)||| + C(\nu + \sigma)\|q_h\|_0 \leq C_2 |||(\mathbf{v}_h, q_h)|||. \qquad (4.15)$$

Now (4.7) follows from (4.14) and (4.15) with $\beta_2 = 1/(C_2(1 + C_1))$. $\quad\square$

Remark 4.3. For $\sigma > 0$ we have control over the L^2 norms of pressure and velocity *uniformly in* $\nu > 0$. In the case $\sigma = 0$ this control is lost as $\nu \to 0$ because of the presence of the convection term: now (4.11) is no longer useful since $\|q_h\|_0$ has disappeared from $|||(\mathbf{v}_h, q_h)|||$. If we consider the Stokes problem (i.e., set $b \equiv \sigma \equiv 0$ in (4.1)), then a careful investigation shows that one still has control over the L^2 norm of the pressure with a constant independent of ν and h. $\quad\clubsuit$

Remark 4.4. The unique solvability of the stabilized discrete problem (4.3) is immediate from Lemma 4.2. $\quad\clubsuit$

4.2.3 Consistency Error

Unlike residual-based stabilization schemes, LPS does not have the Galerkin orthogonality property. Consequently we investigate its consistency error.

Lemma 4.5. *Let* $(\mathbf{u}, p) \in \mathbf{V} \times Q$ *and* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ *be the solutions of* (4.2) *and* (4.3) *respectively. Then for all* $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ *one has*

$$(A + S_h)((\mathbf{u} - \mathbf{u}_h, p - p_h); (\mathbf{v}_h, q_h)) = S_h((\mathbf{u}, p); (\mathbf{v}_h, q_h)).$$

Proof. Simply subtract (4.3a) from (4.2). $\quad\square$

To estimate the consistency error, assume that $\mathbf{b}$ is sufficiently smooth in the sense that

$$\mathbf{b}\big|_M \in \mathbf{W}^{r,\infty}(M) \quad \forall M \in \mathcal{M}_h, \qquad \max_{M \in \mathcal{M}_h} \|\mathbf{b}\|_{r,\infty,M} \leq C. \qquad (4.16)$$

Lemma 4.6. *Suppose that for the fluctuation operator* κ_h *one has*

$$\|\kappa_h q\|_{0,M} \leq C\, h_M^l |q|_{l,M} \qquad \forall q \in H^l(M), \forall M \in \mathcal{M}_h, \ 0 \leq l \leq r. \qquad (4.17)$$

and $\mathbf{b}$ *satisfies* (4.16). *Assume that* $(\mathbf{u}, p) \in \mathbf{H}^{r+1}(\Omega) \times H^{r+1}(\Omega)$. *Then*

$$\left|S_h\big((\mathbf{u},p);(\mathbf{v}_h,q_h)\big)\right| \le C\Bigg(\sum_{M\in\mathcal{M}_h} h_M^{2r}\Big[\big(\tau_M\|\mathbf{b}\|_{r,\infty,M}^2 + \mu_M\big)\|\mathbf{u}\|_{r+1,M}^2$$

$$+\,\alpha_M\|p\|_{r+1,M}^2\Big]\Bigg)^{1/2}|||(\mathbf{v}_h,q_h)|||\quad(4.18)$$

for all $(\mathbf{v}_h,q_h)\in\mathbf{V}_h\times Q_h$.

Proof. From the definition of the stabilizing term we get

$$|S_h((\mathbf{u},p);(\mathbf{v}_h,q_h))| \le \Big(S_h((\mathbf{u},p);(\mathbf{u},p))\Big)^{1/2}\Big(S_h((\mathbf{v}_h,q_h);(\mathbf{v}_h,q_h))\Big)^{1/2}$$

$$\le \Big(S_h((\mathbf{u},p);(\mathbf{u},p))\Big)^{1/2}|||(\mathbf{v}_h,q_h)|||.$$

The approximation properties of $\boldsymbol{\kappa}_h$ imply that

$$S_h\big((\mathbf{u},p);(\mathbf{u},p)\big)$$
$$\le C\sum_{M\in\mathcal{M}_h} h_M^{2r}\big(\tau_M|(\mathbf{b}\cdot\nabla)\mathbf{u}|_{r,M}^2 + \mu_M|\nabla\cdot\mathbf{u}|_{r,M}^2 + \alpha_M|\nabla p|_{r,M}^2\big)$$

and (4.18) follows. $\square$

Remark 4.7. The assumption $\mathbf{b}|_M \in \mathbf{W}^{r,\infty}(M)$ is rather restrictive in the framework of the Navier-Stokes model, since $\mathbf{b}$ corresponds to a finite element function that is in general non-smooth across element borders. But in the case $\mathcal{M}_h = \mathcal{T}_h$ the macro-cells are element cells and this assumption should not be a problem. Another way to relax the smoothness assumption on $\mathbf{b}$ is by means of a modified stabilization term; see Theorem 4.10.   ♣

4.2.4 A priori Error Estimate

Stability and consistency engender an *a priori* error estimate in the usual way. An important aspect of this bound is that the constant multiplier C is independent of the viscosity ν and the mesh size h.

Theorem 4.8. *Let* $(\mathbf{u},p) \in \big(\mathbf{H}_0^1(\Omega)\cap\mathbf{H}^{r+1}(\Omega)\big)\times\big(L_0^2(\Omega)\cap H^{r+1}(\Omega)\big)$ *be the solution of (4.2) and* $(\mathbf{u}_h,p_h)\in\mathbf{V}_h\times Q_h$ *be the solution of the LPS method (4.3). Then there is a positive constant* C, *which is independent of* ν *and* h, *such that*

$$|||(\mathbf{u}-\mathbf{u}_h,p-p_h)||| \le C\Bigg[\sum_{M\in\mathcal{M}_h} h_M^{2r}\Big(\nu + h_M^2\,\sigma + h_M^2\,\tau_M^{-1} + \tau_M\|\mathbf{b}\|_{r,\infty,M}^2$$

$$+\,h_M^2\,\mu_M^{-1} + \mu_M + h_M^2\,\alpha_M^{-1} + \alpha_M\Big)\Big(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2\Big)\Bigg]^{1/2}.$$

The choices $\tau_M \sim h_M/\|\mathbf{b}\|_{r,\infty,M}$, $\mu_M \sim h_M$, and $\alpha_M \sim h_M$ are asymptotically optimal and lead to

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||$$

$$\leq C\left(\sum_{M \in \mathcal{M}_h} (\nu + h_M) h_M^{2r} \left(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \right) \right)^{1/2}.$$

Proof. Lemma 4.2 implies that

$$|||(\mathbf{j}_h \mathbf{u} - \mathbf{u}_h, j_h p - p_h)|||$$

$$\leq \frac{1}{\beta_2} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{(A + S_h)\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}_h, j_h p - p_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||}$$

$$\leq \frac{1}{\beta_2} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{(A + S_h)\big((\mathbf{u} - \mathbf{u}_h, p - p_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||}$$

$$+ \frac{1}{\beta_2} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{(A + S_h)\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}, j_h p - p); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||}.$$

Using Lemmas 4.5 and 4.6, we bound the first term as follows:

$$\frac{(A + S_h)\big((\mathbf{u} - \mathbf{u}_h, p - p_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||} = \frac{S_h\big((\mathbf{u}, p); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||}$$

$$\leq C\left(\sum_{M \in \mathcal{M}_h} h_M^{2r} \left[\left(\tau_M \|\mathbf{b}\|_{r,\infty,M}^2 + \mu_M \right) \|\mathbf{u}\|_{r+1,M}^2 + \alpha_M \|p\|_{r+1,M}^2 \right] \right)^{1/2}.$$

To estimate the second term above, consider separately each individual term in the expression $(A + S_h)\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}, j_h p - p); (\mathbf{w}_h, r_h)\big)$. The treatment of

$$\big| \nu(\nabla(\mathbf{j}_h\mathbf{u} - \mathbf{u}), \nabla \mathbf{w}_h) + \sigma(\mathbf{j}_h\mathbf{u} - \mathbf{u}, \mathbf{w}_h) \big|$$

is standard and leads to the bound

$$C\left(\sum_{M \in \mathcal{M}_h} h_M^{2r} \left(\nu + \sigma\, h_M^2 \right) \|\mathbf{u}\|_{r+1,\Lambda(M)}^2 \right)^{1/2} |||(\mathbf{w}_h, r_h)|||.$$

When dealing with the next three terms, we use the special interpolants that satisfy (4.4)–(4.5). Integrating by parts, one gets

$$\left| \big((\mathbf{b} \cdot \nabla)(\mathbf{j}_h\mathbf{u} - \mathbf{u}), \mathbf{w}_h\big) \right| = \big| (\mathbf{j}_h\mathbf{u} - \mathbf{u}, (\mathbf{b} \cdot \nabla)\mathbf{w}_h) \big|$$

$$= \big| (\mathbf{j}_h\mathbf{u} - \mathbf{u}, \boldsymbol{\kappa}_h(\mathbf{b} \cdot \nabla)\mathbf{w}_h) \big| \tag{4.19}$$

$$\leq C\left(\sum_{M \in \mathcal{M}_h} h_M^{2r+2} \tau_M^{-1} \|\mathbf{u}\|_{r+1,\Lambda(M)}^2 \right)^{1/2} \left(S_h\big((\mathbf{w}_h, 0); (\mathbf{w}_h, 0)\big) \right)^{1/2},$$

$$\left|(p - j_h p, \nabla \cdot \mathbf{w}_h)\right| = \left|(p - j_h p, \kappa_h \nabla \cdot \mathbf{w}_h)\right| \tag{4.20}$$

$$\leq C \left(\sum_{M \in \mathcal{M}_h} h_M^{2r+2} \, \mu_M^{-1} \, \|p\|_{r+1, \Lambda(M)}^2 \right)^{1/2} \left(S_h\big((\mathbf{w}_h, 0); (\mathbf{w}_h, 0)\big) \right)^{1/2},$$

$$\left|(r_h, \nabla \cdot (j_h \mathbf{u} - \mathbf{u}))\right| = \left|(\nabla r_h, j_h \mathbf{u} - \mathbf{u})\right| = \left|(\kappa_h \nabla r_h, j_h \mathbf{u} - \mathbf{u})\right|$$

$$\leq C \left(\sum_{M \in \mathcal{M}_h} h_M^{2r+2} \, \alpha_M^{-1} \, \|\mathbf{u}\|_{r+1, \Lambda(M)}^2 \right)^{1/2} \left(S_h\big((0, r_h); (0, r_h)\big) \right)^{1/2}.$$

Finally, one has

$$\frac{\left| S_h\big((j_h \mathbf{u} - \mathbf{u}, j_h p - p); (\mathbf{w}_h, r_h)\big) \right|}{|||(\mathbf{w}_h, r_h)|||}$$

$$\leq \left[S_h\big((j_h \mathbf{u} - \mathbf{u}, j_h p - p); (j_h \mathbf{u} - \mathbf{u}, j_h p - p)\big) \right]^{1/2} \frac{\left(S_h\big((\mathbf{w}_h, r_h); (\mathbf{w}_h, r_h)\big) \right)^{1/2}}{|||(\mathbf{w}_h, r_h)|||}$$

$$\leq C \left(\sum_{M \in \mathcal{M}_h} h_M^{2r} \Big[\big(\tau_M \|\mathbf{b}\|_{0,\infty,M}^2 + \mu_M\big) \|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \alpha_M \|p\|_{r+1,\Lambda(M)}^2 \Big] \right)^{1/2}.$$

Collecting all the above estimates, we have shown that

$$|||(j_h \mathbf{u} - \mathbf{u}_h, j_h p - p_h)|||$$

$$\leq C \left(\sum_{M \in \mathcal{M}_h} h_M^{2r} \Big[\nu + h_M^2 \, \sigma + h_M^2 \, \tau_M^{-1} + \tau_M \|\mathbf{b}\|_{r,\infty,M}^2 + h_M^2 \, \mu_M^{-1} + \mu_M \right.$$

$$\left. + h_M^2 \, \alpha_M^{-1} + \alpha_M \Big] \Big(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \Big) \right)^{1/2}.$$

By using the triangle inequality

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)||| \leq |||(\mathbf{u} - j_h \mathbf{u}, p - j_h p)||| + |||(j_h \mathbf{u} - \mathbf{u}_h, j_h p - p_h)|||$$

and the approximation property

$$|||(\mathbf{u} - j_h \mathbf{u}, p - j_h p)||| \leq C \left(\sum_{M \in \mathcal{M}_h} h_M^{2r} \Big[\nu + h_M^2 \, \sigma + (\nu + \sigma) \, h_M^2 \right.$$

$$\left. + \tau_M \|\mathbf{b}\|_{0,\infty,M}^2 + \mu_M + \alpha_M \Big] \Big(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \Big) \right)^{1/2},$$

one arrives at

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)||| \leq C \left[\sum_{M \in \mathcal{M}_h} h_M^{2r} \left(\nu + h_M^2 \, \sigma + h_M^2 \, \tau_M^{-1} + \tau_M \|\mathbf{b}\|_{r,\infty,M}^2 \right. \right.$$

$$\left. \left. + h_M^2 \, \mu_M^{-1} + \mu_M + h_M^2 \, \alpha_M^{-1} + \alpha_M \right) \left(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \right) \right]^{1/2}$$

which proves the first estimate of the theorem. Minimizing the upper bound results in the choices $\tau_M \sim h_M / \|\mathbf{b}\|_{r,\infty,M}$, $\mu_M \sim h_M$, and $\alpha_M \sim h_M$, which together imply the second error bound. □

Remark 4.9. Comparing LPS with the SDFEM, in terms of the norm

$$(\mathbf{v}, q) \mapsto \left[\nu |\mathbf{v}|_1^2 + \sigma \|\mathbf{v}\|_0^2 + (\nu + \sigma) \|q\|_0^2 \right]^{1/2}$$

both approaches attain the same rate of convergence in the case of equal-order interpolation [TV96]. Moreover, the LPS gives additional control over

$$\left[\sum_{M \in \mathcal{M}_h} \left(\tau_M \|\boldsymbol{\kappa}_h((\mathbf{b} \cdot \nabla)\mathbf{v})\|_{0,M}^2 + \alpha_M \|\boldsymbol{\kappa}_h(\nabla q)\|_{0,M}^2 + \mu_M \|\boldsymbol{\kappa}_h(\nabla \cdot \mathbf{v})\|_{0,M}^2 \right) \right]^{1/2}$$

whereas for the SDFEM one controls

$$\left[\sum_{T \in \mathcal{T}_h} \left(\delta_T \|(\mathbf{b} \cdot \nabla)\mathbf{v} + \nabla q\|_{0,T}^2 + \mu_T \|\nabla \cdot \mathbf{v}\|_{0,T}^2 \right) \right]^{1/2} .$$

It has been shown recently [ML07] that the SDFEM also gives control over the terms $\|(\mathbf{b} \cdot \nabla)\mathbf{v}\|_{0,T}$ and $\|\nabla q\|_{0,T}$ if $\sigma > 0$ and the parameters δ_T are chosen appropriately. This corresponds to the LPS method where an additional (separate) control is guaranteed over the fluctuations of these quantities. ♣

Finally, we discuss two slightly modified approaches that produce the same error estimates as those of Theorem 4.8. The first of these is to replace the stabilizing term S_h from (4.3b) by

$$S_h^1((\mathbf{u}_h, p_h); (\mathbf{v}_h, q_h))$$
$$:= \sum_{M \in \mathcal{M}_h} \left(\tau_M (\boldsymbol{\kappa}_h(\nabla \mathbf{u}_h), \boldsymbol{\kappa}_h(\nabla \mathbf{v}_h))_M + \alpha_M (\boldsymbol{\kappa}_h \nabla p_h, \boldsymbol{\kappa}_h \nabla q_h)_M \right) \quad (4.21)$$

which gives control over the fluctuations of the gradients of the velocities instead of separate control over the fluctuations of the derivatives in the streamline direction and the divergence. In the second modification, we replace the stabilizing term S_h from (4.3b) by a term S_h^2 that is spectrally equivalent, i.e., $S_h^2 \sim S_h$. Note that the choice of the parameters τ_M, μ_M, and α_M defining S_h influences the selection of possible stabilizing terms S_h^2. When replacing S_h by S_h^i in (4.6) for $i = 1, 2$, two new mesh-dependent norms appear which will be denoted by $|||(\cdot, \cdot)|||_1$ and $|||(\cdot, \cdot)|||_2$.

Theorem 4.10. *Let* $(\mathbf{u}, p) \in \left(\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^{r+1}(\Omega)\right) \times \left(L_0^2(\Omega) \cap H^{r+1}(\Omega)\right)$ *be the weak solution of* (4.2) *and let* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ *be the solution of the LPS method* (4.3) *with* S_h *replaced by* S_h^1. *Then for* $\sigma > 0$ *there is a positive constant* C, *which is independent of* ν, *such that*

$$
|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||_1 \le C\Bigg(\sum_{M \in \mathcal{M}_h} h_M^{2r} \Big[\nu + h_M^2 \left(\sigma + \sigma^{-1}|b|_{1,\infty,M}^2 \right) + h_M^2 \, \alpha_M^{-1}
$$

$$
+ \alpha_M + h_M^2 \, \tau_M^{-1} \left(1 + \|\mathbf{b}\|_{0,\infty,M}^2 \right) + \tau_M \Big] \left(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \right) \Bigg)^{1/2}.
$$

The choices $\tau_M \sim h_M \sqrt{1 + \|\mathbf{b}\|_{0,\infty,M}^2}$ *and* $\alpha_M \sim h_M$ *are asymptotically optimal and lead to*

$$
|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||_1
$$

$$
\le C_\sigma \left[\sum_{M \in \mathcal{M}_h} (\nu + h_M) \, h_M^{2r} \left(\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \right) \right]^{1/2} \qquad (4.22)
$$

with a constant C_σ *that is independent of* ν *but depends on* σ.

Proof. A careful check shows that Lemma 4.2, with S_h and $|||(\cdot, \cdot)|||$ replaced by S_h^1 and $|||(\cdot, \cdot)|||_1$, remains valid. Furthermore, the additional smoothness hypothesis on $\mathbf{b}$ in Lemma 4.6 can be discarded since now the approximation properties of the fluctuation give already

$$
S_h^1\big((\mathbf{u}, p); (\mathbf{u}, p)\big) \le C \sum_{M \in \mathcal{M}_h} h_M^{2r} \Big(\tau_M \, |\nabla \mathbf{u}|_{r,M}^2 + \alpha_M \, |\nabla p|_{r,M}^2 \Big).
$$

The estimates (4.19) and (4.20) in the proof of Theorem 4.8 have to be modified. Consider first (4.20):

$$
\big|(p - j_h p, \nabla \cdot \mathbf{w}_h)\big| = \big|(p - j_h p, \kappa_h \nabla \cdot \mathbf{w}_h)\big|
$$

$$
\le C \sum_{M \in \mathcal{M}_h} h_M^{r+1} \, \tau_M^{-1/2} \|p\|_{r+1,\Lambda(M)} \, \tau_M^{1/2} \|\kappa_h \nabla \cdot \mathbf{w}_h\|_{0,M}
$$

$$
\le C \Bigg(\sum_{M \in \mathcal{M}_h} h_M^{2r+2} \, \tau_M^{-1} \, \|p\|_{r+1,\Lambda(M)}^2 \Bigg)^{1/2} \Big(S_h^1\big((\mathbf{w}_h, 0); (\mathbf{w}_h, 0)\big) \Big)^{1/2}.
$$

The treatment of (4.19) needs more care. Begin as in the proof of Theorem 4.8:

$$
\big|((\mathbf{b} \cdot \nabla)(j_h \mathbf{u} - \mathbf{u}), \mathbf{w}_h)\big| = \big|(j_h \mathbf{u} - \mathbf{u}, (\mathbf{b} \cdot \nabla)\mathbf{w}_h)\big| = \big|(j_h \mathbf{u} - \mathbf{u}, \kappa_h(\mathbf{b} \cdot \nabla)\mathbf{w}_h)\big|
$$

$$
\le C \sum_{M \in \mathcal{M}_h} h_M^{r+1} \|\mathbf{u}\|_{r+1,\Lambda(M)} \, \big\|\kappa_h(\mathbf{b} \cdot \nabla)\mathbf{w}_h\big\|_{0,M}.
$$

Let $\overline{\mathbf{b}}$ be the L^2 projection of $\mathbf{b}$ into the space of piecewise constant functions with respect to the macro-decomposition $\mathcal{M}_h$. Using the L^2 stability of $\boldsymbol{\kappa}_h$, an inverse inequality, and $\boldsymbol{\kappa}_h(\overline{\mathbf{b}} \cdot \nabla)\mathbf{w}_h = \overline{\mathbf{b}} \cdot \boldsymbol{\kappa}_h(\nabla \mathbf{w}_h)$, we get

$$
\begin{aligned}
\left\|\boldsymbol{\kappa}_h(\mathbf{b} \cdot \nabla)\mathbf{w}_h\right\|_{0,M} &\leq \left\|\boldsymbol{\kappa}_h((\mathbf{b} - \overline{\mathbf{b}}) \cdot \nabla)\mathbf{w}_h\right\|_{0,M} + \left\|\boldsymbol{\kappa}_h(\overline{\mathbf{b}} \cdot \nabla)\mathbf{w}_h\right\|_{0,M} \\
&\leq C\, h_M |\mathbf{b}|_{1,\infty,M} \|\nabla \mathbf{w}_h\|_{0,M} + \|\mathbf{b}\|_{0,\infty,M} \left\|\boldsymbol{\kappa}_h(\nabla \mathbf{w}_h)\right\|_{0,M} \\
&\leq C |\mathbf{b}|_{1,\infty,M} \|\mathbf{w}_h\|_{0,M} + \|\mathbf{b}\|_{0,\infty,M} \left\|\boldsymbol{\kappa}_h(\nabla \mathbf{w}_h)\right\|_{0,M}.
\end{aligned}
$$

Substituting this into our previous inequality and recalling that $\sigma > 0$, we obtain

$$
\begin{aligned}
\left|((\mathbf{b} \cdot \nabla)(\mathbf{j}_h\mathbf{u} - \mathbf{u}), \mathbf{w}_h)\right| &\leq C \sum_{M \in \mathcal{M}_h} h_M^{r+1} \|\mathbf{u}\|_{r+1,\Lambda(M)} \\
&\qquad \left(|\mathbf{b}|_{1,\infty,M} \|\mathbf{w}_h\|_{0,M} + \|\mathbf{b}\|_{0,\infty,M} \|\boldsymbol{\kappa}_h(\nabla \mathbf{w}_h)\|_{0,M} \right) \\
&\leq C \left[\sum_{M \in \mathcal{M}_h} h_M^{2r} \left(h_M^2 \sigma^{-1} |\mathbf{b}|_{1,\infty,M}^2 + h_M^2\, \tau_M^{-1} \|\mathbf{b}\|_{0,\infty,M}^2 \right) \|\mathbf{u}\|_{r+1,\Lambda(M)}^2 \right]^{1/2} \\
&\qquad \left[\sigma \|\mathbf{w}_h\|_0^2 + S_h^1\big((\mathbf{w}_h,0);(\mathbf{w}_h,0)\big) \right]^{1/2}.
\end{aligned}
$$

The remaining terms can be estimated as in the proof of Theorem 4.8. The calculation culminates in

$$
\begin{aligned}
|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||_1 \leq C\Bigg\{ &\sum_{M \in \mathcal{M}_h} h_M^{2r} \Big[\nu + h_M^2 \left(\sigma + \sigma^{-1} |\mathbf{b}|_{1,\infty,M}^2 \right) + h_M^2\, \alpha_M^{-1} \\
&+ \alpha_M + h_M^2\, \tau_M^{-1} (1 + \|\mathbf{b}\|_{0,\infty,M}^2) + \tau_M \Big] \left[\|\mathbf{u}\|_{r+1,\Lambda(M)}^2 + \|p\|_{r+1,\Lambda(M)}^2 \right] \Bigg\}^{1/2}
\end{aligned}
$$

which is the first statement of the theorem. Minimizing the upper bound gives $\tau_M \sim h_M \sqrt{1 + \|\mathbf{b}\|_{0,\infty,M}^2}$ and $\alpha_M \sim h_M$, which implies (4.22). $\qquad\square$

We return to the second modification which replaces S_h by a spectrally equivalent stabilization term $S_h^2 \sim S_h$. Assume that the consistency estimate

$$
\left| S_h^2\big((\mathbf{u},p);(\mathbf{v}_h,q_h)\big) \right| \leq C\, h^{r+1/2}\big(\|\mathbf{u}\|_{r+1} + \|p\|_{r+1}\big)\, |||(\mathbf{v}_h,q_h)||| \qquad (4.23)
$$

and the approximation property

$$
\begin{aligned}
\Big| S_h^2\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}, j_h p - p);(\mathbf{v}_h,q_h)\big) \Big| & \\
&\hspace{-6em}\leq C\, h^{r+1/2}\big(\|\mathbf{u}\|_{r+1} + \|p\|_{r+1}\big)\, |||(\mathbf{v}_h,q_h)||| \qquad (4.24)
\end{aligned}
$$

are satisfied for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$.

Theorem 4.11. *Let* $(\mathbf{u}, p) \in \left(\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^{r+1}(\Omega)\right) \times \left(L_0^2(\Omega) \cap H^{r+1}(\Omega)\right)$ *be the solution of* (4.2) *and let* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ *be the solution of the LPS method* (4.3a) *with* S_h *replaced by an* S_h^2 *that satisfies* $S_h^2 \sim S_h$, (4.23), *and* (4.24). *Assume also that* $\tau_M, \mu_M, \alpha_M \sim h_M$ *for all* $M \in \mathcal{M}_h$. *Then there is a positive constant* C, *which is independent of* ν *and* h, *such that*

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||_2 \leq C\left(\nu^{1/2} + h^{1/2}\right) h^r \left(\|\mathbf{u}\|_{r+1} + \|p\|_{r+1}\right).$$

Proof. A careful check shows that Lemma 4.2, with S_h and $|||(\cdot, \cdot)|||$ replaced by S_h^2 and $|||(\cdot, \cdot)|||_2$, retains its validity. Lemma 4.6 is replaced by (4.23). Now following the line of argument of Theorem 4.8 and bounding S_h by CS_h^2, one gets

$$|||(\mathbf{j}_h\mathbf{u} - \mathbf{u}_h, j_h p - p_h)|||_2 \leq C(\nu^{1/2} + h^{1/2}) h^r \left(\|\mathbf{u}\|_{r+1} + \|p\|_{r+1}\right).$$

The desired result then follows from the triangle inequality. $\square$

4.3 Local Projection onto Coarse-Mesh Spaces

The next two sections examine a particular implementation of LPS: we specify an approximation space Y_h and a projection space D_h for which special interpolants satisfying (4.4)–(4.5) exist. Recall that the velocity space and pressure space are defined by $\mathbf{V}_h = Y_h^d \cap \mathbf{V}$ and $Q_h = Y_h \cap Q$, respectively. For simplicity of notation we assume that the fluctuation operator $\kappa_h : L^2(\Omega) \to D_h$ is given by $\kappa_h = \mathrm{id} - \pi_h$ where $\pi_h : L^2(\Omega) \to D_h$ is the L^2 projection.

The notation $\mathcal{M}_h = \mathcal{T}_{2h}$ will be used to indicate that the partition $\mathcal{T}_h$ of the domain Ω is generated by a suitable refinement of a given, shape-regular macro-decomposition $\mathcal{M}_h$. For more details and for the proofs we refer to [MST07]. The method has been extended to anisotropic meshes in [Bra].

4.3.1 Simplices

Refine each macro-simplex $M \in \mathcal{M}_h$ in $\mathbb{R}^d$ through the common simplicial subdivision by means of $(d-1)$-dimensional simplices whose vertices are the barycentre and each subset of $d-1$ vertices of M; see Figure III.3.8 for the cases $d = 2, 3$. For the approximation of velocity and pressure, choose the finite element space of continuous piecewise polynomials of degree at most r on $\mathcal{T}_h$. The projection space comprises discontinuous piecewise polynomials of degree at most $r-1$ on $\mathcal{T}_{2h}$. These choices are summarized by writing $(Y_h, D_h) = (P_{r,h}, P_{r-1,2h}^{\mathrm{disc}})$ where

$$P_{r,h} := \{v \in H^1(\Omega) : v|_K \in P_r(K) \quad \forall K \in \mathcal{T}_h\},$$
$$P_{r-1,2h}^{\mathrm{disc}} := \{v \in L^2(\Omega) : v|_M \in P_{r-1}(M) \quad \forall M \in \mathcal{T}_{2h}\}.$$

Lemma 4.12. *Define the LPS method by setting $(Y_h, D_h) = (P_{r,h}, P_{r-1,2h}^{disc})$ with an arbitrary but fixed polynomial degree $r \in \mathbb{N}$. Then on shape-regular simplicial meshes there exist interpolation operators that satisfy (4.4)–(4.5) and the fluctuation operator satisfies (4.17).*
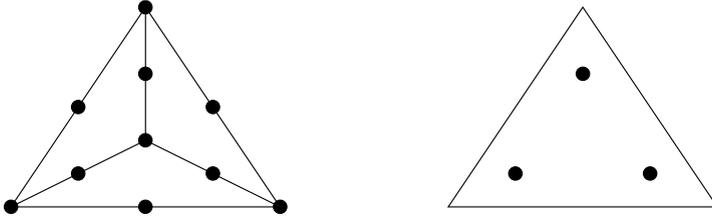
Fig. 4.1. Degrees of freedoms for the approximation space (left) and projection space (right) when $d = 2$ and $r = 2$

4.3.2 Quadrilaterals and Hexahedra

Let $\widehat{M} = (-1, 1)^d$ be the reference hypercube with barycentre $\hat{a}_0$ and vertices $\hat{a}_i$, $i = 1, \ldots, 2^d$. Refine $\widehat{M}$ into 2^d congruent cubes $\widehat{T}_i$, $i = 1, \ldots, 2^d$. Let $F_M : \widehat{M} \to M$ be the bijective multilinear reference mapping onto a macro-element $M \in \mathcal{M}_h$. The refinement of $\widehat{M}$ induces a refinement of M into 2^d cells $T \in \mathcal{T}_h$; see Figure 4.2 for the two-dimensional case. Of course, each of the $\widehat{T}_i$, $i = 1, \ldots, 2^d$, can be mapped bijectively onto a unique reference cube $\widehat{T}$ by a linear mapping. The resulting bijective multilinear reference mapping from $\widehat{T}$ onto $T \in \mathcal{T}_h$ will be denoted by F_T.
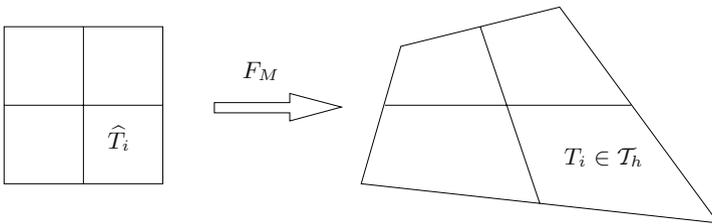
Fig. 4.2. Reference macro-element $\widehat{M}$ (left) and macro-element $M = F_M(\widehat{M}) \in \mathcal{M}_h$ (right) in the two-dimensional case

The projection space D_h can be defined in two ways: either as an image of a space on the reference macro-element $\widehat{M}$ or directly on the macro-element M.

In general this leads to different finite element spaces, so we distinguish between them. We say the projection space D_h is mapped if it is defined on the reference macro-element $\widehat{M}$. If the projection space is defined directly on the macro-element it is said to be unmapped, and this is indicated by an extra superscript "unm". The space Y_h that is used to approximate the velocity components and the pressure is constructed by continuous piecewise polynomials of degree at most r in each variable on the children $\widehat{T}_i$ of the reference macro-element $\widehat{M}$, i.e., by the standard mapped quadrilateral or hexahedral elements.

Projection Spaces Based on Mapped Finite Elements

First consider the standard continuous and discontinuous (mapped) finite element spaces for the approximation and projection spaces, respectively. The approximation space comprises continuous piecewise mapped polynomials of degree at most r in each variable on $\mathcal{T}_h$. The projection space is built from discontinuous piecewise mapped polynomials of degree at most $r - 1$ on $\mathcal{T}_{2h}$. This is summarized as $(Y_h, D_h) = (Q_{r,h}, Q^{\mathrm{disc}}_{r-1,2h})$ where

$$Q_{r,h} := \{v \in H^1(\Omega) \,:\, v|_K \circ F_T \in Q_r(\widehat{T}) \quad \forall T \in \mathcal{T}_h\},$$
$$Q^{\mathrm{disc}}_{r-1,2h} := \{v \in L^2(\Omega) \,:\, v|_M \circ F_M \in Q_{r-1}(\widehat{M}) \quad \forall M \in \mathcal{T}_{2h}\}.$$

Lemma 4.13. *Define the LPS method by setting $(Y_h, D_h) = (Q_{r,h}, Q^{disc}_{r-1,2h})$ with an arbitrary but fixed polynomial degree $r \in \mathbb{N}$. Then on shape-regular meshes there exist interpolation operators satisfying (4.4)-(4.5) and the fluctuation operator satisfies (4.17).*

Alternatively, one can choose smaller projection spaces D_h without losing the approximation property of the fluctuation operator (cf. Lemma 4.6) on families of uniformly-refined meshes. Indeed, let us choose D_h to be

$$P^{\mathrm{disc}}_{r-1,2h} := \{v \in L^2(\Omega) \,:\, v|_M \circ F_M \in P_{r-1}(\widehat{M}) \quad \forall M \in \mathcal{T}_{2h}\}.$$

This choice gives more stabilization in the sense that the stabilizing term vanishes on the smaller subset $P^{\mathrm{disc}}_{r-1,2h} \subset Q^{\mathrm{disc}}_{r-1,2h}$. The existence of the special interpolations still holds without any modification, but to ensure the consistency estimate we have to restrict ourselves to families of uniformly-refined meshes; see [ABF02] for quadrilaterals and [Mat01] for hexahedra. These families are generated by successively refining a given initial mesh.

Lemma 4.14. *Define the LPS method by setting $(Y_h, D_h) = (Q_{r,h}, P^{disc}_{r-1,2h})$ with an arbitrary but fixed polynomial degree $r \in \mathbb{N}$. Then on shape-regular meshes there exist interpolation operators that satisfy (4.4)–(4.5). The fluctuation operator satisfies (4.17) on families of uniformly-refined meshes.*
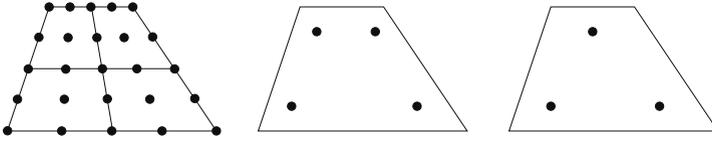
Fig. 4.3. Two variants of choosing the projection space. Degrees of freedom for the approximation space $Q_{r,h}$ (left), the projection space $Q^{\text{disc}}_{r-1,2h}$ (middle), and the projection space $P^{\text{disc}}_{r-1,2h}$ (right) for $d = 2$ and $r = 2$

Projection Spaces Based on Unmapped Finite Elements

Now choose for the projection space D_h the space of discontinuous piecewise unmapped polynomials of degree at most $r - 1$ on $\mathcal{T}_{2h}$; that is, any $q_h \in D_h$ is on each $M \in \mathcal{T}_{2h}$ a polynomial of degree at most $r - 1$. We use this projection space with the approximation space comprising continuous piecewise mapped polynomials of degree at most r in each variable on $\mathcal{T}_h$, i.e.,

$$(Y_h, D_h) = (Q_{r,h}, P^{\text{disc,unm}}_{r-1,2h})$$

where

$$Q_{r,h} := \{v \in H^1(\Omega) \,:\, v|_T \circ F_T \in Q_r(\widehat{T}) \quad \forall T \in \mathcal{T}_h\},$$
$$P^{\text{disc,unm}}_{r-1,2h} := \{v \in L^2(\Omega) \,:\, v|_M \in P_{r-1}(M) \quad \forall M \in \mathcal{T}_{2h}\}.$$

Lemma 4.15. *Define the LPS method by setting $(Y_h, D_h) = (Q_{r,h}, P^{disc}_{r-1,2h})$ with an arbitrary but fixed polynomial degree $r \in \mathbb{N}$. Then on shape-regular meshes there exist interpolation operators that satisfy (4.4)–(4.5) and the fluctuation operator satisfies (4.17).*

4.4 Schemes Based on Enrichment of Approximation Spaces

In the previous section, stabilization was achieved by local projection onto coarser meshes. As a result, for each basis function $\varphi \in Y_h$, the fluctuation $\kappa_h \varphi = \varphi - \pi_h \varphi$ has in general a support larger than the support of φ. This generates a larger stencil in the stiffness matrix, which might not fit into the data structure of a given computer code. In the present section we demonstrate that the key property of LPS – the existence of interpolants with additional orthogonality properties – can also be guaranteed by enriching the approximation space instead of thinning out the projection space. The main benefit of this alternative approach is a smaller stencil of the stiffness matrix compared with the two-level approach of Section 4.3. Both simplicial and hexahedral families of meshes will be considered here. For more details and for the proofs see [MST07].

4.4.1 Simplices

Let $\widehat{T}$ be the reference simplex. Use barycentric coordinates $\hat{\lambda}_i$, $i = 1, \ldots, d+1$, on $\widehat{T}$. Let

$$\hat{b}(\hat{x}) := (d+1)^{d+1} \prod_{i=1}^{d+1} \hat{\lambda}_i(\hat{x})$$

denote the bubble function that takes the value 1 at the barycentre of $\widehat{T}$. Our approximation space will be based on the enriched space

$$P_r^{\text{bubble}}(\widehat{T}) := P_r(\widehat{T}) + \hat{b} \cdot P_{r-1}(\widehat{T})$$

while the projection space is the space of discontinuous piecewise polynomials of degree at most $r-1$ on the same mesh, i.e., $(Y_h, D_h) := (P_{r,h}^{\text{bubble}}, P_{r-1,h}^{\text{disc}})$ where

$$P_{r,h}^{\text{bubble}} := \{ v \in H^1(\Omega) \ : \ v|_T \circ F_T \in P_r^{\text{bubble}}(\widehat{T}) \quad \forall T \in \mathcal{T}_h \},$$

$$P_{r-1,h}^{\text{disc}} := \{ v \in L^2(\Omega) \ : \ v|_T \circ F_T \in P_{r-1}(\widehat{T}) \quad \forall T \in \mathcal{T}_h \}.$$

Lemma 4.16. *Define the LPS method by setting $(Y_h, D_h) = (P_{r,h}^{bubble}, P_{r-1,h}^{disc})$ with an arbitrary but fixed polynomial degree $r \in \mathbb{N}$. Then on shape-regular simplicial meshes there are interpolation operators satisfying (4.4)–(4.5) and the fluctuation operator satisfies (4.17).*

4.4.2 Quadrilaterals and Hexahedra

As in Section 4.3.2 we consider mapped

$$P_{r-1,h}^{\text{disc}} := \{ v \in L^2(\Omega) \ : \ v|_T \circ F_T \in P_{r-1}(\widehat{T}) \quad \forall T \in \mathcal{T}_h \}$$

and unmapped

$$P_{r-1,h}^{\text{disc,unm}} := \{ v \in L^2(\Omega) \ : \ v|_T \in P_{r-1}(T) \quad \forall T \in \mathcal{T}_h \}$$

finite element spaces for the projection space D_h. Note that in order to guarantee the optimal order of the consistency error for the mapped projection space, we have to restrict our attention to families of uniformly-refined quadrilateral/hexahedral meshes – see [ABF02, Mat01]. For unmapped projection spaces the consistency error is of the required order on general shape-regular meshes and one constructs the approximation spaces by (preferably minimal) enrichment of standard finite element spaces.

Projection Spaces Based on Mapped Finite Elements

Let

$$\hat{b}(\hat{x}) = \prod_{i=1}^{d}(1 - \hat{x}_i^2) \in Q_2(\widehat{T})\,, \quad \hat{x} = (\hat{x}_1, \ldots, \hat{x}_d) \in \widehat{T}\,, \quad d = 2, 3\,, \quad (4.25)$$

denote a bubble function associated with the reference cell $\widehat{T} = (-1, 1)^d$. The enriched finite element space is defined on the reference cell by

$$Q_r^{\text{bubble},1}(\widehat{T}) := Q_r(\widehat{T}) \oplus \text{span}\left\{\hat{b}\,\hat{x}_i^{r-1} : i = 1, \ldots, d\right\}$$

and mapped onto the cell $T = F_T(\widehat{T}) \in \mathcal{T}_h$. Thus

$$Q_{r,h}^{\text{bubble},1} := \left\{v \in H^1(\Omega) : v|_T \circ F_T \in Q_r^{\text{bubble},1}(\widehat{T}) \quad \forall T \in \mathcal{T}_h\right\}.$$

Our approximation space comprises continuous piecewise mapped enriched polynomials of degree at most r in each variable while the projection space is the standard space of discontinuous piecewise mapped polynomials of degree at most $r - 1$, i.e., $(Y_h, D_h) := (Q_{r,h}^{\text{bubble},1}, P_{r-1,h}^{\text{disc}})$. In general neither space is polynomial.

Lemma 4.17. *To define the LPS method, set* $(Y_h, D_h) = (Q_{r,h}^{bubble,1}, P_{r-1,h}^{disc})$ *with an arbitrary but fixed polynomial degree* $r \in \mathbb{N}$. *Then on shape-regular meshes there exist interpolation operators that satisfy* (4.4)–(4.5). *The fluctuation operator satisfies* (4.17) *on families of uniformly-refined meshes.*

Remark 4.18. For $r \geq 2$, the space $Q_r^{\text{bubble},1}(\widehat{T})$ has precisely d basis functions more than $Q_r(\widehat{T})$, independently of r.   ♣

To get an impression of the efficiency of the enrichment approach compared with the two-level approach, consider the matrix block that corresponds to one scalar component. We follow [MST07] in comparing asymptotically the numbers of non-zero entries for a decomposition of $\Omega = (0, 1)^d$ into squares/cubes of edge size $1/N$. Since the inner degrees of freedom dominate for high-order elements (i.e., when $r \gg 1$), in the two-dimensional case one has asymptotically $\mathcal{O}(4N^2r^4)$ non-zero entries for the two-level approach whereas the enrichment technique produces only $\mathcal{O}(N^2r^4)$ non-zero entries. In the three-dimensional case these numbers are $\mathcal{O}(8N^3r^6)$ and $\mathcal{O}(N^3r^6)$. This effect is less striking for moderate r. For example, in the case $r = 2$ and $d = 2$, one has asymptotically $\mathcal{O}(144N^2)$ versus $\mathcal{O}(75N^2)$ non-zero entries.

Projection Spaces Based on Unmapped Finite Elements

To relax the assumption that the families of meshes are uniformly refined, we turn to unmapped projection spaces. Choose the space

$$Q_r^{\text{bubble},2}(\widehat{T}) := Q_r(\widehat{T}) + \hat{b} \cdot Q_{r-1}(\widehat{T})$$

with the bubble function $\hat{b}$ of (4.25), and define the enriched space

$$Q_{r,h}^{\text{bubble},2} := \{v \in H^1(\Omega) \; : \; v|_T \circ F_T \in Q_r^{\text{bubble},2}(\widehat{T}) \quad \forall T \in \mathcal{T}_h\}.$$

Now our choice is $(Y_h, D_h) = (Q_{r,h}^{\text{bubble},2}, P_{r-1,h}^{\text{disc,unm}})$.

Lemma 4.19. *Choose $(Y_h, D_h) = (Q_{r,h}^{bubble,2}, P_{r-1,h}^{disc,unm})$ in the LPS method, where the polynomial degree $r \in \mathbb{N}$ is arbitrary but fixed. Then on shape-regular simplicial meshes there exist interpolation operators that satisfy (4.4)–(4.5) and the fluctuation operator satisfies (4.17).*

Remark 4.20. The dimension of the space $Q_{r,h}^{\text{bubble},2}$ is larger than that of $Q_{r,h}^{\text{bubble},1}$. Comparing the dimensions of spaces $Y_h(T)$ and $D_h(T)$, we surmise that the enriched space could be reduced, but the question of constructing interpolants with additional orthogonality properties remains open.   ♣

4.5 Relationship to Subgrid Modelling

The idea of subgrid modelling is due to Guermond [Gue99a] and was first applied to a scalar transport equation. It is based on a scale separation of the underlying finite element space, viz.,

$$Y_h = Y_H \oplus Y_h^H$$

where Y_H represents the space of large scales and Y_h^H the space of small scales. Associated with this scale separation is a suitable projection operator $P_H : Y_h \to Y_H \subset Y_h$ that is the identity mapping on the subspace Y_H. Let $\overline{\kappa}_h := \text{id} - P_H$ denote the fluctuation operator. Assume that the finite element space Y_H is based on a shape-regular decomposition of the domain into cells $M \in \mathcal{M}_h$ of diameter h_M. Then it is proposed [Gue99a, EG04] to add a stabilizing term of the form

$$S(u_h, v_h) = \sum_{M \in \mathcal{M}_h} h_M \big(\nabla \overline{\kappa}_h u_h, \nabla \overline{\kappa}_h v_h\big)_M$$

or

$$S(u_h, v_h) = \sum_{M \in \mathcal{M}_h} h_M \big((b \cdot \nabla)\overline{\kappa}_h u_h, (b \cdot \nabla)\overline{\kappa}_h v_h\big)_M$$

to the standard Galerkin method. These stabilization terms can be interpreted as an artificial diffusion in the streamline direction for the subscales that are represented by Y_h^H. This approach has been developed in different ways;

for an extension to time-dependent convection-diffusion problems see, e.g., [JKL06]. Scale separation also plays an important role in large eddy simulation of turbulent flows – see [Joh06].

Scale separation can be implemented in various ways. In the two-level approach, Y_h and Y_H are standard finite element spaces on different refinement levels (which we indicate by writing $Y_H = Y_{2h}$) and Y_h^H is spanned by those hierarchical basis functions that need to be added to the coarse space Y_{2h} to generate Y_h. An alternative viewpoint is to consider Y_h as a finite element space Y_H that is enriched by a space Y_h^H that contains suitable functions, e.g., higher-order polynomials. Both variants differ from LPS since the stabilization term in the subgrid modelling approach is based on gradients of fluctuations – viz., $\nabla(\mathrm{id} - P_H)u_h$ – whereas the local projection method uses fluctuations of the gradients – viz., $(\mathrm{id} - \pi_h)\nabla u_h$.

In applications, the projection $P_H : Y_h \to Y_{2h}$ in the two-level approach has often been chosen as the global Lagrange interpolant $I_{2h,r}$ that maps into Y_{2h} [BB01, BB06, BBJL07, BR06a, BR06b, Lub06]. This generates a stabilizing term of the form

$$
\begin{aligned}
&\mathcal{S}_h^3\big((\mathbf{u}_h, p_h); (\mathbf{v}_h, q_h)\big) \\
&\quad = \sum_{M \in \mathcal{M}_h} \left(\tau_M(\nabla\overline{\boldsymbol{\kappa}}_h\mathbf{u}_h, \nabla\overline{\boldsymbol{\kappa}}_h\mathbf{v}_h)_M + \alpha_M(\nabla\overline{\kappa}_h p_h, \nabla\overline{\kappa}_h q_h)_M\right)
\end{aligned} \qquad (4.26)
$$

instead of the S_h and S_h^1 of (4.3b) and (4.14). In the following subsections, we study the relationship between the stabilizing terms S_h^1 and S_h^3.

4.5.1 Two-Level Approach with Piecewise Linear Elements

Consider first the case where $\mathcal{T}_h$ is generated from a refinement of a shape-regular triangulation $\mathcal{T}_{2h}$ in $\mathbb{R}^d$ through simplicial subdivision by joining the barycentre to its vertices; see Figure III.3.8 for the cases $d = 2, 3$. Let Y_h and Y_{2h} denote the spaces of continuous piecewise linear finite elements associated with the triangulations $\mathcal{T}_h$ and $\mathcal{T}_{2h}$ respectively.

Lemma 4.21. *Let $d \geq 1$. Let $\pi_{2h,0}$ be the L^2 projection onto the space $P_{0,2h}^{disc}$ of piecewise constant functions and $I_{2h,1} : Y_h \to Y_{2h}$ the Lagrange interpolant into the space $P_{1,2h}$ of continuous piecewise linear functions. Then*

$$
\pi_{2h,0}(\nabla v_h)\big|_M = \nabla I_{2h,1}\big(v_h\big|_M\big) \quad \forall v_h \in P_{1,h}, \forall M \in \mathcal{T}_{2h}.
$$

Hence LPS and subgrid modelling are identical at the discrete level.

Proof. We restrict ourselves to the scalar case since the assertion for the vector-valued case then follows immediately by considering each component separately.

Let $v_h|_M$ be the restriction of an arbitrary function $v_h \in Y_h$ to a macro-simplex $M \in \mathcal{T}_{2h}$. Denote the barycentre of M by a_0 and its vertices by

a_i, $i = 1, \ldots, d+1$. Let the barycentric coordinates on the simplex M be λ_i, $i = 1, \ldots, d+1$, where $\lambda_i(a_i) = 1$ for each i. When M is subdivided into simplices, each having for its vertices the barycentre and $d-1$ vertices of M, denote these new simplices by T_i, $i = 1, \ldots, d+1$, where $a_i \notin T_i$ for each i. Define a continuous piecewise linear function on M by

$$\varphi_0(x) = (d+1)\lambda_i(x) \quad \text{for } x \in T_i, \; i = 1, \ldots, d+1.$$

One can then use the nodal functionals $N_i(v) = v(a_i)$, $i = 0, \ldots, d+1$, to write

$$v_h\big|_M = \sum_{i=1}^{d+1} N_i(v_h)\lambda_i + \widetilde{N}_0(v_h)\varphi_0 \tag{4.27}$$

where

$$\widetilde{N}_0(v) = N_0(v) - \frac{1}{d+1}\sum_{i=1}^{d+1} N_i(v).$$

Since $N_i(\varphi_0) = 0$ for $i = 1, \ldots, d+1$, we have $I_{2h,1}v_h = \sum_{i=1}^{d+1} N_i(v_h)\lambda_i$, whence

$$\nabla I_{2h,1}v_h = \sum_{i=1}^{d+1} N_i(v_h)\nabla\lambda_i.$$

Let ∇_h denote the gradient operator that is applied piecewise. As $\nabla_h v_h$ is constant on each subdomain T_j for $j = 1, \ldots, d+1$, and $|T_j| = |M|/(d+1)$, we compute the L^2 projection onto $P_0(M)$ to be

$$\pi_{2h,0}(\nabla_h v_h) = \frac{1}{d+1}\sum_{j=1}^{d+1} \nabla_h v_h\big|_{T_j}.$$

For $j = 1, \ldots, d+1$, from (4.27) one has

$$\nabla v_h\big|_{T_j} = \sum_{i=1}^{d+1} N_i(v_h)\nabla\lambda_i + (d+1)\widetilde{N}_0(v_h)\nabla\lambda_j,$$

$$\frac{1}{d+1}\sum_{j=1}^{d+1}\nabla v_h\big|_{T_j} = \sum_{i=1}^{d+1} N_i(v_h)\nabla\lambda_i + \widetilde{N}_0(v_h)\nabla\sum_{j=1}^{d+1}\lambda_j = \sum_{i=1}^{d+1} N_i(v_h)\nabla\lambda_i.$$

This equation says that $\pi_{2h,0}(\nabla v_h)|_M = \nabla I_{2h,1}(v_h|_M)$. It follows that the stabilizing terms in the two approaches are identical. $\quad\square$

Remark 4.22. In general

$$\pi_{2h,r-1}\nabla_h v_h\Big|_M \neq \nabla I_{2h,r}v_h \quad \forall v_h \in P_{r,h}, \; r \geq 2,$$

where $\pi_{2h,r-1}$ is the L^2 projection onto the space $P_{r-1,2h}^{\mathrm{disc}}$ of discontinuous piecewise polynomials of degree at most $r-1$ on the coarse mesh $\mathcal{T}_{2h}$ and

$I_{2h,r} : Y_h \rightarrow P_{r,2h}$ is the Lagrange interpolant in the space of continuous piecewise polynomials of degree at most r on the coarse mesh $\mathcal{T}_{2h}$. Similarly, in general on quadrilateral or hexahedral meshes $\mathcal{T}_{2h}$ one has

$$\pi_{2h,r-1}\nabla_h v_h\Big|_M \neq \nabla I_{2h,r} v_h \quad \forall v_h \in Q_{r,h}^d, \; d \geq 2, \; r \geq 1,$$

where $\pi_{2h,r-1}$ is the L^2 projection onto the space $Q_{r-1,2h}^{\mathrm{disc}}$ of discontinuous piecewise polynomials of degree at most $r-1$ in each variable, while $I_{2h,r} : Y_h \rightarrow Q_{r,2h}$ is the Lagrange interpolant in the space of continuous piecewise polynomials of degree at most r in each variable. As an example, consider the case $r = 2$, $d = 1$. For the reference macro-element $\widehat{M} = (-1,+1)$ and the piecewise quadratic function

$$\widehat{v}(\hat{x}) = \begin{cases} 4\hat{x}(1-\hat{x}) & \text{if } 0 \leq \hat{x} \leq 1, \\ 0 & \text{if } -1 \leq \hat{x} < 0, \end{cases}$$

one can see that

$$\hat{\pi}_{2h,1}\hat{\nabla}\hat{v} = -\hat{x} \neq 0 = \hat{\nabla}\hat{I}_{2h,2}\hat{v}.$$

Thus in general subgrid modelling and LPS do not construct identical stabilization terms. But as we shall see later, this does not exclude the possibility of spectral equivalence of the stabilization terms. ♣

4.5.2 Enriched Piecewise Linear Elements

The previous subsection demonstrated that LPS and subgrid modelling employ the same stabilization term in the two-level approach with $Y_h = P_{1,h}$ and $D_h = P_{0,2h}^{\mathrm{disc}}$. We now show that the same is true for enriched piecewise linear elements, i.e., when $Y_h = P_{1,h}^{\mathrm{bubble}}$ and $D_h = P_{0,h}^{\mathrm{disc}}$.

Lemma 4.23. *Let $d \geq 1$. Let $\pi_{h,0}$ be the L^2 projection onto the space $P_{0,h}^{disc}$ of piecewise constant functions and let $I_{h,1} : Y_h \rightarrow P_{1,h}$ be the Lagrange interpolant in the space $P_{1,h}$ of continuous piecewise linear functions. Then*

$$\pi_{h,0}(\nabla v_h)\big|_T = \nabla I_{h,1}\big(v_h\big|_T\big) \quad \forall v_h \in P_{1,h}^{bubble}, \; \forall T \in \mathcal{T}_h.$$

Hence LPS and subgrid modelling are identical at the discrete level.

Proof. For simplicity of notation we present the proof for the scalar case as its extension to the vector-valued case in the space Y_h^d is straightforward. Consider a simplex $T \in \mathcal{T}_h$ with vertices a_i, $i = 1, \ldots, d+1$, barycentre a_0, and barycentric coordinates λ_i, $i = 1, \ldots, d+1$, where $\lambda_i(a_i) = 1$ for each i. The restriction $v_h\big|_T$ of a finite element function $v_h \in Y_h$ to T can be represented through its nodal functionals $N_i(v) = v(a_i)$, $i = 0, \ldots, d+1$, as

$$v_h\big|_T = \sum_{i=1}^{d+1} N_i(v_h)\lambda_i + \widetilde{N}_0(v_h)\,b = I_{h,1}v_h + \widetilde{N}_0(v_h)\,b$$

where

$$\widetilde{N}_0(v) = N_0(v) - \frac{1}{d+1}\sum_{i=1}^{d+1} N_i(v) \quad \text{and} \quad b = (d+1)^{d+1}\prod_{i=1}^{d+1}\lambda_i.$$

Hence

$$\nabla v_h\big|_T = \nabla(I_{h,1}v_h) + N_0(v_h)\,\nabla b.$$

Since $\nabla(I_{h,1}v_h)$ is constant on T, one has $\pi_{h,0}\nabla(I_{h,1}v_h) = \nabla(I_{h,1}v_h)$. Thus it remains only to show that

$$\pi_{h,0}(\nabla b) = \frac{1}{|T|}\int_T \nabla b\,dx = \mathbf{0}.$$

But this identity follows immediately from Gauss's theorem as b vanishes on ∂T. That is, $\pi_{h,0}(\nabla v_h)\big|_T = \nabla I_{h,1}\big(v_h\big|_T\big)$ and it follows that the stabilizing terms in both approaches are identical. $\quad\square$

4.5.3 Spectral Equivalence of the Stabilizing Terms on Simplices

The spectral equivalence of the stabilizing terms S_h^3 given by (4.26) and S_h^1 given by (4.21) will now be shown on simplices. To this end, it is sufficient to prove the existence of positive constants C_3 and C_4 such that

$$C_3\|\kappa_h\nabla w_h\|_{0,M} \le \|\nabla\overline{\kappa}_h w_h\|_{0,M} \le C_4\|\kappa_h\nabla w_h\|_{0,M} \qquad (4.28)$$

for all $w_h \in Y_h$ and $M \in \mathcal{M}_h$.

Consider first the two-level approach.

Lemma 4.24. *Let $(Y_h, D_h) = (P_{r,h}, P_{r-1,2h}^{disc})$. Write $\pi_{2h,r-1}$ for the L^2 projection onto D_h, $\kappa_h = \mathrm{id} - \pi_{2h,r-1}$ and $I_{2h,r}$ for the Lagrange interpolant in $P_{r,2h}$. Set $\overline{\kappa}_h = \mathrm{id} - I_{2h,r}$. Then the stabilizing terms S_h^3 and S_h^1 are spectrally equivalent.*

Proof. For each $M \in \mathcal{T}_{2h}$ let $F_M : \widehat{M} \to M$ be the affine mapping from the reference macro-cell $\widehat{M}$ onto the cell M. Thus $F_M(\hat{x}) = B_M\hat{x} + b_M$ for all $\hat{x} \in \widehat{M}$, where B_M is a $d \times d$ matrix and b_M is a column vector. The L^2 projection $\pi_{2h,r-1}$ and the Lagrange interpolant $I_{2h,r}$ are invariant with respect to affine transformations, i.e., denoting the corresponding operators on the reference cell by $\hat{\pi}$ and $\hat{I}$, one has

$$\widehat{(\pi_{2h,r-1}w)} = \hat{\pi}\hat{w}, \qquad \widehat{I_{2h,r}w} = \hat{I}\hat{w}$$

and the corresponding relations

$$\widehat{\kappa_h \nabla w} = \hat{\kappa} \, \widehat{\nabla w}, \qquad \widehat{\overline{\kappa}_h w} = \widehat{\overline{\kappa}} \, \widehat{\nabla w}$$

for the fluctuation operators. Now the transformation formulas $\widehat{\nabla v} = B_M^{-T} \hat{\nabla} \hat{v}$ and $\hat{\nabla}\hat{v} = B_M^T \widehat{\nabla v}$ [Cia02, Chapter 3.1] yield

$$\|\kappa_h \nabla w\|_{0,M} = |\det B_M|^{1/2} \, \|\widehat{\kappa_h \nabla w}\|_{0,\widehat{M}} = |\det B_M|^{1/2} \, \|\hat{\kappa} B_M^{-T} \hat{\nabla}\hat{w}\|_{0,\widehat{M}}$$
$$\leq |\det B_M|^{1/2} \, \|B_M^{-1}\| \, \|\hat{\kappa}\hat{\nabla}\hat{w}\|_{0,\widehat{M}}, \qquad\qquad (4.29\text{a})$$
$$\|\hat{\nabla}\widehat{\overline{\kappa}}\hat{w}\|_{0,\widehat{M}} = \|\hat{\nabla}\widehat{\overline{\kappa}_h w}\|_{0,\widehat{M}} = |\det B_M|^{-1/2}\|B_M^T \nabla \overline{\kappa}_h w\|_{0,M}$$
$$\leq |\det B_M|^{-1/2}\|B_M\| \, \|\nabla \overline{\kappa}_h w\|_{0,M}, \qquad\qquad (4.29\text{b})$$

where $\|B_M\|$ and $\|B_M^{-1}\|$ are the matrix norms of B_M and B_M^{-1} that are induced by the Euclidean vector norm. For shape-regular meshes one has $\|B_M^{-1}\| \, \|B_M\| \leq C$.

If there is a constant C such that $\|\hat{\kappa}\hat{\nabla}\hat{w}\|_{0,\widehat{M}} \leq C \, \|\hat{\nabla}\widehat{\overline{\kappa}}\hat{w}\|_{0,\widehat{M}}$, then from (4.29) we get

$$\|\kappa_h \nabla w\|_{0,M} \leq C_3^{-1} \, \|\nabla \overline{\kappa}_h w\|_{0,M}$$

which is the left-hand inequality of (4.28). The proof of the right-hand inequality follows from $\|\hat{\nabla}\widehat{\overline{\kappa}}\hat{w}\|_{0,\widehat{M}} \leq C \, \|\hat{\kappa}\hat{\nabla}\hat{w}\|_{0,\widehat{M}}$ by similar arguments.

To derive these hypothesized inequalities on the reference element, consider the mappings

$$\hat{w} \mapsto \|\hat{\kappa}\hat{\nabla}\hat{w}\|_{0,\widehat{M}} \quad \text{and} \quad \hat{w} \mapsto \|\hat{\nabla}\widehat{\overline{\kappa}}\hat{w}\|_{0,\widehat{M}}.$$

Each is a norm on the respective finite-dimensional factor spaces

$$P_r(\widehat{M}) \big/ \{\hat{w} \, : \, \hat{\kappa}\hat{\nabla}\hat{w} = \mathbf{0}\} \qquad \text{and} \qquad P_r(\widehat{M}) \big/ \{\hat{w} \, : \, \hat{\nabla}\widehat{\overline{\kappa}}\hat{w} = \mathbf{0}\}.$$

Suppose that $\hat{\kappa}\hat{\nabla}\hat{w} = \mathbf{0}$. Then

$$\hat{\nabla}\hat{w} = \hat{\pi}\hat{\nabla}\hat{w} \in \left(P_{r-1}(\widehat{M})\right)^d \Rightarrow \hat{w} \in P_r(\widehat{M}) \Rightarrow \hat{I}\hat{w} = \hat{w} \Rightarrow \hat{\nabla}\widehat{\overline{\kappa}}\hat{w} = \mathbf{0}.$$

Conversely, suppose that $\hat{\nabla}\widehat{\overline{\kappa}}\hat{w} = \mathbf{0}$. Recalling that $\hat{w}$ is continuous on $\widehat{M}$, we obtain

$$\hat{w} = \hat{I}\hat{w} + \text{const} \in P_r(\widehat{M}) \Rightarrow \hat{\nabla}\hat{w} = \hat{\nabla}\hat{I}\hat{w} \in \left(P_{r-1}(\widehat{M})\right)^d$$
$$\Rightarrow \hat{\pi}\hat{\nabla}\hat{w} = \hat{\nabla}\hat{w}$$
$$\Rightarrow \hat{\kappa}\hat{\nabla}\hat{w} = \mathbf{0}.$$

Thus the two factor spaces coincide and the desired inequalities follow immediately from the equivalence of norms on finite-dimensional spaces. $\square$

Let us turn to the case of enriched finite element spaces Y_h.

Lemma 4.25. *Choose* $Y_h = P_{r,h}^{bubble}$, *which was defined in Section 4.4.1. Set* $D_h = P_{r-1,h}^{disc}$. *Let* $\pi_{h,r-1}$ *be the* L^2 *projection onto* D_h *and* $I_{h,r}$ *the Lagrange interpolant in* $P_{r,h}$. *Set* $\kappa_h = \mathrm{id}-\pi_{h,r-1}$ *and* $\overline{\kappa}_h = \mathrm{id}-I_{h,r}$. *Then the stabilizing terms* S_h^3 *and* S_h^1 *are spectrally equivalent.*

Proof. By using the affine transformation $F_T : \widehat{T} \to T$ from the reference cell $\widehat{T}$ onto T, one can show – as in the proof of Lemma 4.24 – that it suffices to establish the corresponding estimates on the reference cell. As before, this is done by showing that the mappings

$$\hat{w} \mapsto \|\hat{\kappa}\hat{\nabla}\hat{w}\|_{0,\widehat{T}}, \qquad \hat{w} \mapsto \|\hat{\nabla}\hat{\overline{\kappa}}\hat{w}\|_{0,\widehat{T}}$$

are norms on the corresponding factor spaces

$$P_r(\widehat{M}) \big/ \{\hat{w} \; : \; \hat{\kappa}\hat{\nabla}\hat{w} = \mathbf{0}\} \qquad \text{and} \qquad P_r(\widehat{M}) \big/ \{\hat{w} \; : \; \hat{\nabla}\hat{\overline{\kappa}}\hat{w} = \mathbf{0}\}.$$

Suppose that $\hat{\kappa}\hat{\nabla}\hat{w} = \mathbf{0}$. Then

$$\hat{\nabla}\hat{w} = \hat{\pi}\hat{\nabla}\hat{w} \in \big(P_{r-1}(\widehat{T})\big)^d \Rightarrow \hat{w} \in P_r(\widehat{T}) \Rightarrow \hat{I}\hat{w} = \hat{w} \Rightarrow \hat{\nabla}\hat{\overline{\kappa}}\hat{w} = \mathbf{0}.$$

Conversely, suppose that $\hat{\nabla}\hat{\overline{\kappa}}\hat{w} = \mathbf{0}$. We obtain

$$\hat{w} = \hat{I}\hat{w} + \mathrm{const} \in P_r(\widehat{T}) \Rightarrow \hat{\nabla}\hat{w} = \hat{\nabla}\hat{I}\hat{w} \in \big(P_{r-1}(\widehat{T})\big)^d$$
$$\Rightarrow \hat{\pi}\hat{\nabla}\hat{w} = \hat{\nabla}\hat{w}$$
$$\Rightarrow \hat{\kappa}\hat{\nabla}\hat{w} = \mathbf{0}.$$

Hence there exist two constants C_3 and C_4 such that

$$C_3\|\kappa_h\nabla w_h\|_{0,T} \le \|\nabla\overline{\kappa}_h w_h\|_{0,M} \le C_4\|\kappa_h\nabla w_h\|_{0,T} \qquad \forall w_h \in Y_h, \quad \forall T \in \mathcal{T}_h,$$

and the stabilizing terms S_h^3 and S_h^1 are spectrally equivalent. $\square$

Remark 4.26. For quadrilateral and hexahedral elements one does not have in general the spectral equivalence of the stabilizing terms. For example, consider the case $d = 2$, $r = 1$. In the two-level approach, for the function $\hat{w}(\hat{x}) = \hat{x}_1\hat{x}_2$ on the macro-element $\widehat{M} = (-1,+1)^2$ one has

$$\hat{\nabla}\hat{w} - \hat{\pi}\hat{\nabla}\hat{w} = \hat{\nabla}\hat{w} = (\hat{x}_2, \hat{x}_1)^T,$$

but the Lagrange interpolant $\hat{I}$ in $Q_1(\widehat{M})$ gives

$$\hat{I}\hat{w} = \hat{w} \quad \Rightarrow \quad \hat{\nabla}(\hat{w} - \hat{I}\hat{w}) = (0,0)^T.$$

The situation is the same for enriched approximation spaces Y_h on a reference cell $\widehat{T}$.   ♣

5

Local Projection Method for Inf-Sup Stable Elements

The previous chapter showed that local projection stabilization (LPS) for equal-order interpolation can handle two types of instabilities – that caused by a violation of the discrete inf-sup condition and that due to dominant convection in the case of high Reynolds number. But the flow problem is often only part of a coupled flow-transport problem; in the next chapter we shall see that mass conservation in the transport equation depends on the properties of the discrete velocity and in particular on the satisfaction of the incompressibility constraint. Unfortunately, when LPS is applied with equal-order interpolation, the discrete divergence-free property of the velocity field is disturbed by the term

$$\sum_{M \in \mathcal{M}_h} \alpha_M (\boldsymbol{\kappa}_h \nabla p_h, \boldsymbol{\kappa}_h \nabla q_h)_M$$

that stabilizes the pressure. For inf-sup stable finite element pairs, this pressure stabilization is unnecessary and we are faced only with the instability caused by dominant convection. Thus it is of interest to consider local projection stabilization for inf-sup stable finite elements.

The main objective of this chapter is an analysis of convergence properties of LPS applied to inf-sup stable discretizations of the Oseen problem. We shall restrict our attention to the enrichment variant of LPS and to a stabilizing term that controls separately fluctuations of the derivative in the streamline direction and fluctuations of the divergence. An interesting point is that for inf-sup stable finite element pairs one does not need an H^1-stable interpolation operator with additional orthogonality properties to prove the stability of the discrete problem, unlike the case of equal-order interpolation (Lemma 4.2). As a consequence, one has much more flexibility in choosing the approximation and projection spaces. Most of the known inf-sup stable finite element pairs approximate the velocity components by elements of order r and the pressure by elements of order $r-1$, which yields error estimates of order r, so compared to LPS with equal-order interpolation by elements of order r, half an order

of convergence is lost. Recently, in [MT07], new inf-sup stable finite element pairs have been proposed that approximate both velocity and pressure by elements of order r. Here, in contrast to "classical" equal-order interpolation, the velocity components and the pressure are discretized by different finite elements. We prove that the discrete inf-sup condition holds true for these finite element spaces and derive an error estimate of order $r + 1/2$ uniformly in the viscosity and reaction coefficients. In the case of discontinuous pressure approximations, an additional term controlling the jumps of the pressure over inner cell faces must be added.

5.1 Discretization by Inf-Sup Stable Elements

The Oseen problem is

$$
\begin{aligned}
-\nu \triangle \mathbf{u} + (\mathbf{b} \cdot \nabla)\mathbf{u} + \sigma \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega, \\
\nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega, \\
\mathbf{u} &= \mathbf{0} \quad \text{on } \partial\Omega,
\end{aligned}
$$

where $\Omega \subset \mathbb{R}^d$ is bounded with a Lipschitz-continuous boundary, $\nu > 0$ and $\sigma \geq 0$ are constants, and $\mathbf{b} \in \left(W^{1,\infty}(\Omega)\right)^d$ is a given velocity field for which $\nabla \cdot \mathbf{b} = 0$. Set $\mathbf{V} = \left(H_0^1(\Omega)\right)^d$ and $Q = L_0^2(\Omega)$. Then a weak formulation of this problem is:

Find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that one has

$$
A\big((\mathbf{u}, p); (\mathbf{v}, q)\big) = (\mathbf{f}, \mathbf{v}) \qquad \forall (\mathbf{v}, q) \in \mathbf{V} \times Q \tag{5.1}
$$

where

$$
\begin{aligned}
A\big((\mathbf{u}, p); (\mathbf{v}, q)\big) = \nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + \big((\mathbf{b} \cdot \nabla)\mathbf{u}, \mathbf{v}\big) + \sigma(\mathbf{u}, \mathbf{v}) \\
- (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}).
\end{aligned}
$$

As stated previously in Theorem 1.5, (5.1) has a unique solution for all $\nu > 0$.

We use a family $\{\mathcal{T}_h\}$ of shape-regular decompositions of Ω into d-simplices, quadrilaterals, or hexahedra. The set of all inner element faces $E \not\subset \partial\Omega$ is denoted by $\mathcal{E}_h$. Associate with each face $E \in \mathcal{E}_h$ an arbitrary but fixed unit normal vector n_E, and let T_E be a fixed element from $\mathcal{T}_h$ such that $E \subset \partial T_E$. If $T_1, T_2 \in \mathcal{T}_h$ are two different cells from $\mathcal{T}_h$ that share a common face $E = \partial T_1 \cap \partial T_2$, then the jump of each piecewise smooth function r_h across the face E is defined by

$$
[r_h]_E = (r_h|_{T_1})\big|_E - (r_h|_{T_2})\big|_E
$$

where n_E is directed from T_1 into T_2.

Let $Y_h \subset H_0^1(\Omega)$ be a scalar finite element space of continuous piecewise mapped polynomial functions over $\mathcal{T}_h$. The finite element space $\mathbf{V}_h$ for approximating the velocity field is $\mathbf{V}_h := Y_h^d$. The pressure is discretized using a finite element space $Q_h \subset Q$ of continuous or piecewise continuous functions on $\mathcal{T}_h$. In this chapter we consider inf-sup stable pairs $(\mathbf{V}_h, Q_h)$: assume that there exists a positive constant β_0 such that

$$\inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{(q_h, \nabla \cdot \mathbf{v}_h)}{|\mathbf{v}_h|_1 \, \|q_h\|_0} \geq \beta_0 > 0 \tag{5.2}$$

uniformly in h.

Let $D_h^i(T)$, $i = 1, 2$, be finite-dimensional spaces on each cell $T \in \mathcal{T}_h$ and let $\pi_T^i : L^2(T) \to D_h^i(T)$ be the associated local L^2 projections into $D_h^i(T)$. The global projection spaces D_h^i are defined by

$$D_h^i := \bigoplus_{T \in \mathcal{T}_h} D_h^i(T), \quad i = 1, 2.$$

These spaces are discontinuous with respect to the family $\mathcal{T}_h$. For $i = 1, 2$, the mapping $\pi_h^i : L^2(\Omega) \to D_h^i$ defined by $(\pi_h^i v)|_T := \pi_T^i(v|_T)$ for all $T \in \mathcal{T}_h$ is the L^2 projection into the projection space D_h^i. Associate with each π_h^i the fluctuation operators $\kappa_h^i := \mathrm{id} - \pi_h^i$ where $\mathrm{id} : L^2(\Omega) \to L^2(\Omega)$ is the identity mapping. Note that the case $D_h^i = \{0\}$ is allowed, which means that κ_h^i is then the identity mapping. The operators π_h^i and κ_h^i will be applied component by component to vector-valued and tensor-valued arguments.

The stabilizing term is

$$S_h(\mathbf{u}, \mathbf{v}) := \sum_{T \in \mathcal{T}_h} \Big(\tau_T \big(\kappa_h^1(\mathbf{b} \cdot \nabla) \mathbf{u}, \kappa_h^1(\mathbf{b} \cdot \nabla) \mathbf{v} \big)_T$$
$$+ \gamma_T \big(\kappa_h^2(\nabla \cdot \mathbf{u}), \kappa_h^2(\nabla \cdot \mathbf{v}) \big)_T \Big), \tag{5.3}$$

which controls the fluctuations of the derivatives in the streamline direction and the fluctuations of the divergence. Other stabilization terms are considered in [MT07]. On the product space $\mathbf{V} \times Q$ define the bilinear form

$$A_h\big((\mathbf{u}, p); (\mathbf{v}, q)\big) := \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + \big((\mathbf{b} \cdot \nabla) \mathbf{u}, \mathbf{v}\big) + \sigma(\mathbf{u}, \mathbf{v})$$
$$+ S_h(\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u})$$

and the mesh-dependent norm

$$|||(\mathbf{v}, q)||| := \big(\nu |\mathbf{v}|_1^2 + \sigma \|\mathbf{v}\|_0^2 + (\nu + \sigma)\|q\|_0^2 + S_h(\mathbf{v}, \mathbf{v}) \big)^{1/2}.$$

Then our stabilized discrete problem is

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that one has

$$A_h\big((\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)\big) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h. \tag{5.4}$$

Existence, uniqueness and convergence properties of solutions of (5.4) will be studied in the sections that follow.

5.2 Stability and Consistency

Consider first the solvability of the discrete problem (5.4).

Lemma 5.1. *Let* $\max\{\nu, \sigma, \tau_T, \gamma_T\} \leq C$. *Then there exists a positive constant* β, *which is independent of* ν, σ, *and* h, *such that*

$$\inf_{(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{A_h\big((\mathbf{v}_h, q_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{v}_h, q_h)|||\, |||(\mathbf{w}_h, r_h)|||} \geq \beta > 0.$$

Proof. Let $(\mathbf{v}_h, q_h)$ be an arbitrary element of $\mathbf{V}_h \times Q_h$. Integrating the convection term by parts, one obtains

$$A_h\big((\mathbf{v}_h, q_h); (\mathbf{v}_h, q_h)\big) = \nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + S_h(\mathbf{v}_h, \mathbf{v}_h).$$

The discrete inf-sup condition (5.2) ensures that for each $q_h \in Q_h$ there exists $\mathbf{z}_h = \mathbf{z}_h(q_h) \in \mathbf{V}_h$ such that

$$(\nabla \cdot \mathbf{z}_h, q_h) = -\|q_h\|_0^2 \quad \text{and} \quad \|\mathbf{z}_h\|_1 \leq C_1\|q_h\|_0, \tag{5.5}$$

where C_1 depends only on the inf-sup constant β_0 and the Friedrichs constant for the domain Ω. Hence

$$A_h\big((\mathbf{v}_h, q_h); (\mathbf{z}_h, 0)\big) = \nu(\nabla \mathbf{v}_h, \nabla \mathbf{z}_h) + \big((\mathbf{b} \cdot \nabla)\mathbf{v}_h, \mathbf{z}_h\big) + \sigma(\mathbf{v}_h, \mathbf{z}_h)$$
$$+ S_h(\mathbf{v}_h, \mathbf{z}_h) + \|q_h\|_0^2 \tag{5.6}$$

on using the first property from (5.5). We shall estimate the first four terms of (5.6). Of these, the first and third can be bounded in a standard way: using the hypothesis that $\nu, \sigma \leq C$, one has

$$\big|\nu(\nabla \mathbf{v}_h, \nabla \mathbf{z}_h) + \sigma(\mathbf{v}_h, \mathbf{z}_h)\big| \leq \nu|\mathbf{v}_h|_1\,|\mathbf{z}_h|_1 + \sigma\|\mathbf{v}_h\|_0\,\|\mathbf{z}_h\|_0$$
$$\leq C\big(\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2\big)^{1/2}\|q_h\|_0$$
$$\leq \frac{\|q_h\|_0^2}{6} + C\big(\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2\big),$$

where the second property from (5.5) was invoked. An integration by parts shows that the second term of (5.6) satisfies

$$\big|\big((\mathbf{b} \cdot \nabla)\mathbf{v}_h, \mathbf{z}_h\big)\big| = \big|\big((\mathbf{b} \cdot \nabla)\mathbf{z}_h, \mathbf{v}_h\big)\big| \leq C|\mathbf{z}_h|_1\|\mathbf{v}_h\|_0 \leq \frac{\|q_h\|_0^2}{6} + C\|\mathbf{v}_h\|_0^2$$

where the boundedness of $\mathbf{b}$ and (5.5) were used. It remains to consider the stabilizing term S_h. Since π_h is the L^2 projection onto the discontinuous finite element space D_h, the corresponding fluctuation operator κ_h is locally L^2 stable. Thus the boundedness of the user-chosen parameters τ_T, γ_T and of $\mathbf{b}$ imply that

$$
\left| S_h(\mathbf{v}_h, \mathbf{z}_h) \right| \leq \left(S_h(\mathbf{v}_h, \mathbf{v}_h) \right)^{1/2} \left(S_h(\mathbf{z}_h, \mathbf{z}_h) \right)^{1/2} \leq C \left(S_h(\mathbf{v}_h, \mathbf{v}_h) \right)^{1/2} |\mathbf{z}_h|_1
$$
$$
\leq \frac{\|q_h\|_0^2}{6} + C S_h(\mathbf{v}_h, \mathbf{v}_h).
$$

Combining the above estimates, we obtain

$$
A_h\big((\mathbf{v}_h, q_h); (\mathbf{z}_h, 0)\big) \geq \frac{\|q_h\|_0^2}{2} - C\big[\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + S_h(\mathbf{v}_h, \mathbf{v}_h)\big] - C\|\mathbf{v}_h\|_0^2.
$$

Multiply this inequality by $2(\nu + \sigma)$ then use Friedrichs's inequality to get the bound

$$
2(\nu + \sigma)\|\mathbf{v}_h\|_0^2 \leq C\big(\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2\big);
$$

this yields

$$
A_h\big((\mathbf{v}_h, q_h); 2(\nu + \sigma)(\mathbf{z}_h, 0)\big) \geq (\nu + \sigma)\|q_h\|_0^2
$$
$$
- C_2\big[\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + S_h(\mathbf{v}_h, \mathbf{v}_h)\big]
$$

with a certain constant C_2. For each $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$, define the pair $(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h$ by

$$
(\mathbf{w}_h, r_h) := (\mathbf{v}_h, q_h) + \frac{2(\nu + \sigma)}{1 + C_2}(\mathbf{z}_h, 0).
$$

Then

$$
A_h\big((\mathbf{v}_h, q_h); (\mathbf{w}_h, r_h)\big)
$$
$$
\geq \frac{\nu + \sigma}{1 + C_2}\|q_h\|_0^2 + \left(1 - \frac{C_2}{1 + C_2}\right)\big[\nu|\mathbf{v}_h|_1^2 + \sigma\|\mathbf{v}_h\|_0^2 + S_h(\mathbf{v}_h, \mathbf{v}_h)\big]
$$
$$
\geq \frac{1}{1 + C_2}|||(\mathbf{v}_h, q_h)|||^2.
$$

It remains to show that $|||(\mathbf{w}_h, r_h)||| \leq C|||(\mathbf{v}_h, q_h)|||$. Towards this we have

$$
|||(\mathbf{w}_h, r_h)||| \leq |||(\mathbf{v}_h, q_h)||| + \frac{2(\nu + \sigma)}{1 + C_2}|||(\mathbf{z}_h, 0)|||
$$
$$
\leq |||(\mathbf{v}_h, q_h)||| + \frac{2(\nu + \sigma)}{1 + C_2} C\|\mathbf{z}_h\|_1
$$
$$
\leq |||(\mathbf{v}_h, q_h)||| + C(\nu + \sigma)\|q_h\|_0 \leq C_3|||(\mathbf{v}_h, q_h)|||.
$$

Hence the desired inf-sup condition holds true with $\beta = 1/\big(C_3(1 + C_2)\big)$.   □

Remark 5.2. Lemma 5.1 implies existence and uniqueness of a solution for the discrete problem (5.4), together with a stability bound on that solution. Note that the mapping $w \mapsto \|\kappa_T w\|_{0,T}$ vanishes on the local projection space $D_h(T)$. Thus the stability of the discrete problem increases as the dimension of the projection space decreases, since the norm $||| \cdot |||$ becomes stronger. In other words, we can control the stability of the discrete problem by choosing an appropriate projection space.   ♣

Next we study the consistency error caused by adding the LPS terms to the standard Galerkin discretization. Assume that the fluctuation operator κ_h^1 provides local approximation properties of order s, i.e., that

$$\|\kappa_h^1 w\|_{0,T} \leq Ch_T^s |w|_{s,T} \quad \forall w \in H^s(T),\, \forall T \in \mathcal{T}_h. \tag{5.7}$$

Note that (5.7) is always satisfied for $s = 0$ since $(\kappa_h^1 w)|_T = w|_T - \pi_T^1(w|_T)$ and π_T^1 is the L^2 projection on $D_h^1(T)$. It is fulfilled for $s > 0$ if for example $D_h^1(T) \subset P_{s-1}(T)$; this follows from the Bramble-Hilbert lemma.

Lemma 5.3. *Let $(\mathbf{u}, p) \in \mathbf{V} \times Q$ and $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ be the solutions of (5.1) and (5.4), respectively. Furthermore, assume that $\mathbf{u} \in H^{s+1}(\Omega)^d$ for some integer $s \in [0, r]$. Assume that the fluctuation operator κ_h^1 satisfies assumption (5.7) and $\mathbf{b}|_T \in W^{s,\infty}(T)^d$ with $\max_T \|\mathbf{b}\|_{s,\infty,T} \leq C$. Then for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has*

$$\left| A_h\big((\mathbf{u} - \mathbf{u}_h, p - p_h); (\mathbf{v}_h, q_h)\big) \right| \leq C \left(\sum_{T \in \mathcal{T}_h} \tau_T\, h_T^{2s} \|\mathbf{u}\|_{s+1,T}^2 \right)^{1/2} |||(\mathbf{v}_h, q_h)|||.$$

Proof. Using (5.4) and

$$A_h\big((\mathbf{u}, p); (\mathbf{v}_h, q_h)\big) = S_h(\mathbf{u}, \mathbf{v}_h) + (\mathbf{f}, \mathbf{v}_h) \quad \forall (\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h,$$

we see that only $S_h(\mathbf{u}, \mathbf{v}_h)$ has to be estimated. The definition (5.3) gives

$$\left| S_h(\mathbf{u}, \mathbf{v}_h) \right| \leq \big(S_h(\mathbf{u}, \mathbf{u})\big)^{1/2} \big(S_h(\mathbf{v}_h, \mathbf{v}_h)\big)^{1/2} \leq \big(S_h(\mathbf{u}, \mathbf{u})\big)^{1/2} |||(\mathbf{v}_h, q_h)|||.$$

The boundedness of $\max_T \|\mathbf{b}\|_{s,\infty,T}$ and the properties of the fluctuation operator κ_h^1 now yield

$$S_h(\mathbf{u}, \mathbf{u}) \leq C \sum_{T \in \mathcal{T}_h} h_K^{2s} \tau_T \big|(\mathbf{b} \cdot \nabla)\mathbf{u}\big|_{s,T}^2 \leq C \sum_{T \in \mathcal{T}_h} \tau_T h_T^{2s} \|\mathbf{u}\|_{s+1,T}^2$$

where $\nabla \cdot \mathbf{u} = 0$ has been used. $\square$

5.3 Convergence

To study the order of convergence of our method, we couch the approximation properties of the spaces $\mathbf{V}_h$ and Q_h in terms of the existence of corresponding interpolation operators. First, we consider the usual inf-sup stable pairs $(\mathbf{V}_h, Q_h)$ that approximate the velocity components and the pressure by elements of order r and $r - 1$ respectively. In general, the constant in the error estimate is independent of ν and the mesh size h but may depend on σ. Then we show that under additional assumptions one can construct interpolation

operators that enjoy certain orthogonality properties. These interpolation operators allow us to establish estimates with error constants that are independent of the data ν, σ and h. Finally, we turn to the case of inf-sup stable pairs $(\mathbf{V}_h, Q_h)$ that approximate both the velocity components and the pressure by elements of order r. An example for the lowest-order case ($r = 1$) with continuous pressure approximation will be the Mini-element [ABF84, BF91]. For each case considered we give several examples of approximation spaces $\mathbf{V}_h, Q_h$ and projection spaces D_h^i, $i = 1, 2$ that satisfy all the assumptions of our convergence theory.

5.3.1 Methods of Order r in the Case $\sigma > 0$

Consider inf-sup stable pairs $(\mathbf{V}_h, Q_h)$ of finite element spaces of polynomial order r and $r - 1$ respectively. Assume in this subsection that $r \geq 2$. Assume that interpolation operators $\mathbf{j}_h : \mathbf{V} \cap H^2(\Omega)^d \to \mathbf{V}_h$ and $i_h : Q \cap H^2(\Omega) \to Q_h$ exist such that for all $\mathbf{w} \in H^\ell(T)^d$, $2 \leq \ell \leq r + 1$, one has

$$\|\mathbf{w} - \mathbf{j}_h\mathbf{w}\|_{0,T} + h_T|\mathbf{w} - \mathbf{j}_h\mathbf{w}|_{1,T} \leq Ch_T^\ell\|\mathbf{w}\|_{\ell,T} \quad \forall T \in \mathcal{T}_h, \qquad (5.8a)$$

and for all $q \in H^\ell(T)$, $2 \leq \ell \leq r$,

$$\|q - i_h q\|_{0,T} + h_T|q - i_h q|_{1,T} \leq Ch_T^\ell\|q\|_{\ell,T} \quad \forall T \in \mathcal{T}_h. \qquad (5.8b)$$

Furthermore, let the pressure interpolation i_h satisfy the orthogonality condition

$$(q - i_h q, r_h) = 0 \quad \forall r_h \in D_h^2, \forall q \in Q \cap H^2(\Omega). \qquad (5.8c)$$

Theorem 5.4. *Assume that the spaces $\mathbf{V}_h$, Q_h satisfy (5.2) and (5.8) and the function $\mathbf{b}$ satisfies the regularity assumption of Lemma 5.3. Choose the projection space D_h^1 so that the associated fluctuation operator κ_h^1 fulfils (5.7) for some integer $s \in [0, r]$. Let the user-chosen parameters satisfy $\gamma_T \sim 1$ and $\tau_T \leq Ch_T^{2(r-s)}$ for some positive constant C. Let $(\mathbf{u}, p) \in (\mathbf{V} \cap H^{r+1}(\Omega)^d) \times (Q \cap H^r(\Omega))$ and $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ be the solutions of (5.1) and (5.4). Then for each $\sigma > 0$ there exists a positive constant C_σ, which is independent of ν and h, such that*

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)||| \leq C_\sigma \left[\sum_{T \in \mathcal{T}_h} h_T^{2r} \left(\|\mathbf{u}\|_{r+1,T}^2 + \|p\|_{r,T}^2 \right) \right]^{1/2}. \qquad (5.9)$$

Proof. By Lemma 5.1 one has

$$|||(\mathbf{j}_h\mathbf{u} - \mathbf{u}_h, i_h p - p_h)|||$$

$$\leq \frac{1}{\beta} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{A_h\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}_h, i_h p - p_h); (\mathbf{w}_h, r_h)\big)}{|||(w_h, r_h)|||}$$

$$\leq \frac{1}{\beta} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{A_h\big((\mathbf{u} - \mathbf{u}_h, p - p_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||}$$

$$+ \frac{1}{\beta} \sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{A_h\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}, i_h p - p); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||} .$$

Invoking Lemma 5.3, the consistency error can be bounded:

$$\sup_{(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h} \frac{A_h\big((\mathbf{u} - \mathbf{u}_h, p - p_h); (\mathbf{w}_h, r_h)\big)}{|||(\mathbf{w}_h, r_h)|||} \leq C \left(\sum_{T \in \mathcal{T}_h} \tau_T \, h_T^{2s} \|\mathbf{u}\|_{s+1,T}^2 \right)^{1/2}.$$

The terms in $A_h\big((\mathbf{j}_h\mathbf{u} - \mathbf{u}, i_h p - p); (\mathbf{w}_h, r_h)\big)$ will be estimated individually. For the stabilizing term S_h, one has

$$S_h(\mathbf{j}_h\mathbf{u} - \mathbf{u}, \mathbf{w}_h) \leq \big(S_h(\mathbf{j}_h\mathbf{u} - \mathbf{u}, \mathbf{j}_h\mathbf{u} - \mathbf{u})\big)^{1/2} \big(S_h(\mathbf{w}_h, \mathbf{w}_h)\big)^{1/2}$$

$$\leq C \left[\sum_{T \in \mathcal{T}_h} (\tau_T + \gamma_T) h_T^{2r} \|\mathbf{u}\|_{r+1,T}^2 \right]^{1/2} |||(\mathbf{w}_h, r_h)|||$$

where the L^2 stability of the fluctuation operators κ_h^i, $i = 1, 2$, the boundedness of $\mathbf{b}$, and the interpolation properties of $\mathbf{j}_h$ were used. Furthermore,

$$\big|\nu\big(\nabla(\mathbf{j}_h\mathbf{u} - \mathbf{u}), \nabla\mathbf{w}_h\big) + \sigma(\mathbf{j}_h\mathbf{u} - \mathbf{u}, \mathbf{w}_h)\big|$$

$$\leq \big(\nu|\mathbf{j}_h\mathbf{u} - \mathbf{u}|_1^2 + \sigma\|\mathbf{j}_h\mathbf{u} - \mathbf{u}\|_0^2\big)^{1/2} \big(\nu|\mathbf{w}_h|_1^2 + \sigma\|\mathbf{w}_h\|_0^2\big)^{1/2}$$

$$\leq C \left[\sum_{T \in \mathcal{T}_h} (\nu + \sigma h_T^2) h_T^{2r} \|\mathbf{u}\|_{r+1,T}^2 \right]^{1/2} |||(\mathbf{w}_h, r_h)|||$$

via the Cauchy-Schwarz inequality and the interpolation properties of $\mathbf{j}_h$. Now consider the pressure-related terms. We have

$$\big(r_h, \nabla \cdot (\mathbf{j}_h\mathbf{u} - \mathbf{u})\big) \leq \|r_h\|_0 \|\nabla \cdot (\mathbf{j}_h\mathbf{u} - \mathbf{u})\|_0$$

$$\leq C \left(\sum_{T \in \mathcal{T}_h} \frac{h_T^{2r}}{\nu + \sigma} \|\mathbf{u}\|_{r+1,T}^2 \right)^{1/2} |||(\mathbf{w}_h, r_h)||| \qquad (5.10)$$

and, by (5.8c),

$$(p - i_h p, \nabla \cdot \mathbf{w}_h) = (p - i_h p, \kappa_h^2 \nabla \cdot \mathbf{w}_h)$$

$$\leq C \left(\sum_{T \in \mathcal{T}_h} \gamma_T^{-1} h_T^{2r} \|p\|_{r,T}^2 \right)^{1/2} |||(\mathbf{w}_h, r_h)|||. \qquad (5.11)$$

The convective term is handled by

$$\left| ((\mathbf{b} \cdot \nabla)(\mathbf{j}_h \mathbf{u} - \mathbf{u}), \mathbf{w}_h) \right| \leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2r} \|\mathbf{u}\|_{r+1,T}^2 \right)^{1/2} \|\mathbf{w}_h\|_0$$

$$\leq C \left(\sum_{T \in \mathcal{T}_h} \frac{h_T^{2r}}{\nu + \sigma} \|\mathbf{u}\|_{r+1,T}^2 \right)^{1/2} |||(\mathbf{w}_h, r_h)||| \qquad (5.12)$$

where the boundedness of $\mathbf{b}$ and, in the final inequality, Friedrichs's inequality have been used. Putting together all these estimates and using $\tau_T \leq Ch_T^{2(r-s)}$, $\gamma_T \sim 1$ and $\max\{\nu, \sigma\} \leq C$, we obtain

$$|||(\mathbf{j}_h \mathbf{u} - \mathbf{u}_h, i_h p - p_h)||| \leq C \left[\sum_{T \in \mathcal{T}_h} h_T^{2r} \left(\|\mathbf{u}\|_{r+1,T}^2 + \|p\|_{r,T}^2 \right) \right]^{1/2}.$$

The interpolation properties of $\mathbf{j}_h$, i_h and the upper bounds on τ_T, γ_T yield

$$|||(\mathbf{u} - \mathbf{j}_h \mathbf{u}, p - i_h p)||| \leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2r} \left(\|\mathbf{u}\|_{r+1,T}^2 + \|p\|_{r,T}^2 \right) \right)^{1/2}.$$

Finally, the triangle inequality

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)||| \leq |||(\mathbf{u} - \mathbf{j}_h \mathbf{u}, p - i_h p)||| + |||(\mathbf{j}_h \mathbf{u} - \mathbf{u}_h, i_h p - p_h)|||$$

gives the statement of the theorem. $\square$

Next we give – without attempting to be exhaustive – examples of approximation spaces $\mathbf{V}_h$, Q_h and projection spaces D_h^1, D_h^2 that satisfy all the hypotheses of Theorem 5.4. For a simplex $T \in \mathcal{T}_h$, let $\widehat{T}$ denote the reference unit simplex in $\mathbb{R}^d$. For a quadrilateral/hexahedron T, let $\widehat{T}$ be the reference cube $(-1, 1)^d$. The reference mapping $F_T : \widehat{T} \to T$ is affine for simplices and generally non-affine for quadrilaterals and hexahedra. Let $P_k(\widehat{T})$, $k \geq 0$, denote the space of polynomials with total degree at most k while $Q_k(\widehat{T})$, $k \geq 0$, is the space of polynomials of degree at most k in each variable. For convenience, we set $P_{-k}(\widehat{T}) = Q_{-k}(\widehat{T}) = \{0\}$ for all positive integers k. Furthermore, on the reference simplex $\widehat{T}$ define the spaces

$$P_k^+(\widehat{T}) := P_k(\widehat{T}) + \hat{b} \cdot P_{k-2}(\widehat{T}), \qquad P_k^{++}(\widehat{T}) := P_k(\widehat{T}) + \hat{b} \cdot P_{k-1}(\widehat{T}),$$

where $\hat{b} \in P_{d+1}(\widehat{T})$ is a bubble function that vanishes on the boundary $\partial\widehat{T}$. Set

$$Q_k^+(\widehat{T}) := Q_k(\widehat{T}) + \hat{b} \cdot \text{span} \{x_i^{k-1} : i = 1, \ldots, d\}$$

on the reference cube $\widehat{T}$ where $\hat{b} \in Q_2(\widehat{T})$ is a bubble function that vanishes on $\partial\widehat{T}$. These spaces on the reference cells are used to define mapped finite element spaces. Set

$$P_r^{\mathrm{disc}} = \big\{ v \in L^2(\Omega) \,:\, v|_T \circ F_T \in P_r(\widehat{T}) \,\forall T \in \mathcal{T}_h \big\}, \quad P_r = P_r^{\mathrm{disc}} \cap H^1(\Omega),$$

$$Q_r^{\mathrm{disc}} = \big\{ v \in L^2(\Omega) \,:\, v|_T \circ F_T \in Q_r(\widehat{T}) \,\forall T \in \mathcal{T}_h \big\}, \quad Q_r = Q_r^{\mathrm{disc}} \cap H^1(\Omega),$$

and

$$P_r^+ = \big\{ v \in H^1(\Omega) \,:\, v|_T \circ F_T \in P_r^+(\widehat{T}) \,\forall T \in \mathcal{T}_h \big\},$$

$$P_r^{++} = \big\{ v \in H^1(\Omega) \,:\, v|_T \circ F_T \in P_r^{++}(\widehat{T}) \,\forall T \in \mathcal{T}_h \big\},$$

$$Q_r^+ = \big\{ v \in H^1(\Omega) \,:\, v|_T \circ F_T \in Q_r^+(\widehat{T}) \,\forall T \in \mathcal{T}_h \big\}.$$

For brevity write $\mathbf{V}_h = Q_k$ and $Q_h = P_k$ instead of $\mathbf{V}_h = \big(Q_k \cap H_0^1(\Omega) \big)^d$ and $Q_h = P_k \cap L_0^2(\Omega)$. The mapped spaces P_r^{disc} are used later also on quadrilaterals and hexahedra for the pressure and the projection spaces. While these spaces do not enjoy the usual approximation properties on arbitrary families of meshes, these properties are valid on families of uniformly-refined meshes, which are often used in practice. For details, see [ABF02, Mat01, MS07].

In the construction of pressure interpolations that satisfy (5.8c), the following lemmas will be helpful. We start with continuous pressure approximations and introduce the notation

$$Q_h(T) := \big\{ q_h|_T \,:\, q_h \in Q_h + \mathrm{span}\,\{1\} \big\}, \quad \widetilde{Q}_h(T) := \big\{ q_h \,:\, b_T \cdot q_h \in Q_h(T) \big\},$$

where b_T denotes the mapped bubble function of lowest polynomial degree, i.e., $b_T \in P_{d+1}(T)$ for simplices in $\mathbb{R}^d$ and $b_T \in Q_2(T)$ for quadrilaterals and hexahedra.

Lemma 5.5. *Let the interpolation operator* $i_h^* : Q \cap H^2(\Omega) \to Q_h \subset H^1(\Omega)$ *have the approximation property* (5.8b). *Let the projection spaces* D_h^2 *satisfy* $D_h^2(T) \subset \widetilde{Q}_h(T)$ *for all* $T \in \mathcal{T}_h$. *Then there exists an interpolation operator* $i_h : Q \cap H^2(\Omega) \to Q_h$ *that satisfies the approximation property* (5.8b) *and the orthogonality condition* (5.8c).

Proof. Modify i_h^* by setting $i_h q := i_h^* q + d_h(q)$, with $d_h(q)|_T := b_T \cdot \tilde{d}_T$ where $\tilde{d}_T \in \widetilde{Q}_h(T)$ is defined locally on each $T \in \mathcal{T}_h$ by

$$(d_h(q), r_h)_T = (b_T \cdot \tilde{d}_T, r_h)_T = (q - i_h^* q, r_h)_T \quad \forall r_h \in \widetilde{Q}_h(T). \qquad (5.13)$$

The uniqueness of a solution $\tilde{d}_T \in \widetilde{Q}_h(T)$ follows from the observation that $(d, r) \mapsto (b_T \cdot d, r)_T$ is a weighted L^2 inner product on $\widetilde{Q}_h(T)$. Since the bubble function b_T vanishes on the boundary ∂T of each cell, the interpolant $i_h q := i_h^* q + d_h(q)$ belongs to $Q_h \subset Q \cap H^1(\Omega)$ and locally preserves polynomials of degree at most r. Thus the Bramble-Hilbert lemma implies (5.8b) for simplicial finite elements. In the case of quadrilateral and hexahedral finite elements on uniformly-refined meshes, we appeal to the results of [ABF02, Mat01, MS07]. Furthermore, (5.13) shows that the error $q - i_h q$ is perpendicular to the projection space D_h^2 and the orthogonality property (5.8c) holds true. $\quad\square$

The version of Lemma 5.5 for discontinuous pressure approximations is as follows.

Lemma 5.6. *Let $Q_h = P_{r-1}^{disc}$ or $Q_h = Q_{r-1}^{disc}$, with $D_h^2 \subset Q_h + span\ \{1\}$. Then the L^2 projection $i_h : L^2(\Omega) \to Q_h$ has the approximation property (5.8b) and the orthogonality condition (5.8c). If $\nabla \cdot \mathbf{V}_h \subset Q_h + span\ \{1\}$ then one has $(q - i_h q, \nabla \cdot \mathbf{w}_h) = 0$ for all $\mathbf{w}_h \in \mathbf{V}_h$, independently of the choice of D_h^2.*

Proof. The discontinuity of the pressure space Q_h implies that the L^2 projection is local. Consequently the approximation property (5.8b) follows from the Bramble-Hilbert lemma for simplicial finite elements in the usual way. In the case of quadrilateral and hexahedral finite elements on uniformly-refined meshes, the result is proved in [ABF02, Mat01, MS07]. Furthermore, one has

$$(q - i_h q, r_h) = 0 \quad \forall r_h \in Q_h + span\ \{1\}.$$

Thus for $D_h^2 \subset Q_h + span\ (1)$, we conclude that (5.8c) holds true. If one has $\nabla \cdot \mathbf{V}_h \subset Q_h + span\ \{1\}$, then set $r_h = \nabla \cdot \mathbf{w}_h$ to get $(q - i_h q, \nabla \cdot \mathbf{w}_h) = 0$ for all $\mathbf{w}_h \in \mathbf{V}_h$, independently of the choice of D_h^2. $\square$

We turn now to concrete examples, starting with continuous pressure approximations; see Table 5.1. The assumptions (5.2) and (5.8) are clearly satis-

Table 5.1. Taylor-Hood families of order $r \geq 2$

$\mathbf{V}_h$	Q_h	D_h^1	D_h^2	τ_T	γ_T	s	t	$\|\|\| \cdot \|\|\|$
P_r	P_{r-1}	P_{s-1}^{disc}	P_{t-1}^{disc}	$\mathcal{O}(h_T^{2(r-s)})$	~ 1	$s \leq r$	$t \leq r - d - 1$	$\mathcal{O}(h^r)$
Q_r	Q_{r-1}	Q_{s-1}^{disc}	Q_{t-1}^{disc}	$\mathcal{O}(h_T^{2(r-s)})$	~ 1	$s \leq r$	$t \leq r - 2$	$\mathcal{O}(h^r)$

fied for the Taylor-Hood families on simplices and quadrilaterals/hexahedra. Indeed, the additional orthogonality assumption (5.8c) for the pressure interpolation can be fulfilled by using a sufficiently small projection space D_h^2; in particular the choices $P_{-1}^{disc} = Q_{-1}^{disc} = \{0\}$ always satisfy (5.8c). By Lemma 5.5, the largest possible projection space D_h^2 such that (5.8c) still holds is given by the bubble part $\widetilde{Q}_h(T)$ of P_{r-1} and Q_{r-1} respectively. The bubble part corresponds to P_{r-d-2}^{disc} for simplicial elements and to Q_{r-3}^{disc} for quadrilateral/hexahedral elements. Finally, the fluctuation operator κ_h^1 satisfies assumption (5.7) for all choices of D_h^1 given in Table 5.1.

Consider now examples of inf-sup stable finite element pairs $\mathbf{V}_h, Q_h$ with discontinuous pressure approximations. The inf-sup stability and approximation properties listed in Table 5.2 follow from [CR73, GR86, MT02]. The orthogonality assumption (5.8c) is satisfied for $D_h^2 \subset Q_h + span\ \{1\}$ when using the local L^2 projection as a pressure interpolation; see Lemma 5.6.

Table 5.2. Families of order $r \geq 2$ with discontinuous pressure approximations

$\mathbf{V}_h$	Q_h	D_h^1	D_h^2	τ_T	γ_T	s	t	$||| \cdot |||$
P_r^+	P_{r-1}^{disc}	P_{s-1}^{disc}	P_{t-1}^{disc}	$\mathcal{O}(h_T^{2(r-s)})$	~ 1	$s \leq r$	$t \leq r$	$\mathcal{O}(h^r)$
Q_r	P_{r-1}^{disc}	P_{s-1}^{disc}	P_{t-1}^{disc}	$\mathcal{O}(h_T^{2(r-s)})$	~ 1	$s \leq r$	$t \leq r$	$\mathcal{O}(h^r)$

Remark 5.7. If $D_h^2 \subset Q_h + \text{span}\{1\}$, then the L^2 projection of a discretely divergence-free function $\mathbf{w}_h$ is zero since

$$(\pi_h^2 \nabla \cdot \mathbf{w}_h, r_h) = (\nabla \cdot \mathbf{w}_h, r_h) = 0 \quad \forall r_h \in D_h^2.$$

This is the case for all the families of Table 5.2. As a consequence, the discrete solution $\mathbf{u}_h$ does not depend on the choice of the projection space. Nevertheless the algebraic properties of the discrete system depend on the choice of D_h^2. ♣

Remark 5.8. The enrichment of P_r in the first row of Table 5.2 is needed only to ensure the inf-sup condition on arbitrary shape-regular meshes. If one considers only families of meshes that are generated by dividing a d-simplex into $(d+1)$ simplices in the usual way (using hyperplanes through the barycentre and sets of $d-2$ vertices), then the inf-sup condition holds true for $r \geq d$; see [Qin94, SV85, Zha05]. Thus in this case one can replace P_r^+ by P_r. The pair (P_r, P_{r-1}) is known as the Scott-Vogelius element. ♣

5.3.2 Methods of Order r in the Case $\sigma \geq 0$

A careful inspection of the proof of Theorem 5.4 shows that when $\sigma = 0$, the error constant in (5.9) is no longer uniformly bounded as $\nu \to 0$ owing to the estimates (5.10) and (5.12). We shall see below that one can get error estimates that hold uniformly for all $\sigma \geq 0$ by choosing a special interpolant $\mathbf{j}_h : \mathbf{V} \cap H^2(\Omega)^d \to \mathbf{V}_h$. In this subsection the polynomial degree is $r \geq 2$.

To handle both continuous and discontinuous pressure approximations, modify the discrete problem by introducing the additional stabilizing term

$$J_h(p, q) := \sum_{E \in \mathcal{E}_h} \alpha_E \big\langle [p]_E, [q]_E \big\rangle_E,$$

where the α_E are user-chosen parameters. Define a bilinear form A_h^*, a stabilizing term S_h^* and an associated mesh-dependent norm $||| \cdot |||_*$ by

$$A_h^*\big((\mathbf{u}, p); (\mathbf{v}, q)\big) := A_h\big((\mathbf{u}, p); (\mathbf{v}, q)\big) + J_h(p, q),$$

$$S_h^*\big((\mathbf{u}, p); (\mathbf{v}, q)\big) := S_h(\mathbf{u}, \mathbf{v}) + J_h(p, q),$$

$$|||(\mathbf{v}, q)|||_* := \big[|||(\mathbf{v}, q)|||^2 + J_h(q, q)\big]^{1/2}.$$

Note that this modification does not introduce any additional consistency error since for smooth solutions $p \in H^1(\Omega)$ one has $[p]_E = 0$ on all $E \in \mathcal{E}_h$ where $\mathcal{E}_h$ is the set of all inner faces.

We start from a quasi-local interpolation operator with a discrete divergence property [GS03] then modify it following [MST07] so that the interpolation error becomes orthogonal to the projection space.

Assumption A: There exists an operator $\mathbf{j}_h^* : \mathbf{V} \to \mathbf{V}_h$ satisfying

$$(q_h, \nabla \cdot (\mathbf{w} - \mathbf{j}_h^* \mathbf{w})) = 0 \quad \forall \mathbf{w} \in \mathbf{V}, \, \forall q_h \in Q_h, \tag{5.14a}$$

$$|\mathbf{v} - \mathbf{j}_h^* \mathbf{v}|_{m,T} \le C \, h_T^{\ell-m} |\mathbf{v}|_{\ell,\omega(T)} \quad \forall \mathbf{v} \in \mathbf{V} \cap H^\ell(\Omega)^d, \, \forall T \in \mathcal{T}_h, \tag{5.14b}$$

for $0 \le m \le 1$, $1 \le \ell \le r+1$, where $\omega(T)$ is a neighbourhood of T. Moreover, let the local inf-sup condition

$$\exists \beta_1 > 0 \, \forall h > 0 \, \forall T \in \mathcal{T}_h : \inf_{q_h \in D_h^1(T)} \sup_{v_h \in Y_h(T)} \frac{(v_h, q_h)_T}{\|v_h\|_{0,T} \|q_h\|_{0,T}} \ge \beta_1 > 0 \tag{5.15}$$

be satisfied where $Y_h(T) := \{v_h|_T : v_h \in Y_h, \, v_h = 0 \text{ on } \Omega \backslash T\}$ is the local bubble part of the scalar finite element space Y_h.   ♣

Remark 5.9. The existence of quasi-local interpolation operators $\mathbf{j}_h^*$ satisfying (5.14) has been established for a wide family of pairs $\mathbf{V}_h$, Q_h in [GS03]. As regards (5.15), it is necessary that $Y_h(T)$ – compared with $D_h^1(T)$ – be rich enough. In particular, one must have $\dim Y_h(T) \ge \dim D_h^1(T)$. Examples of spaces Y_h, D_h^1 satisfying (5.15) have been given in Section 4.4.   ♣

Lemma 5.10. *Let Assumption A be satisfied. Then there exists an interpolation operator $\mathbf{j}_h : \mathbf{V} \to \mathbf{V}_h$ with the following orthogonality and approximation properties:*

$$(\mathbf{w} - \mathbf{j}_h \mathbf{w}, q_h) = 0 \quad \forall q_h \in (D_h^1)^d, \, \forall \mathbf{w} \in \mathbf{V}, \tag{5.16a}$$

$$|\mathbf{v} - \mathbf{j}_h \mathbf{v}|_{m,T} \le C \, h_T^{\ell-m} |\mathbf{v}|_{\ell,\omega(T)} \quad \forall \mathbf{v} \in \mathbf{V} \cap H^\ell(\Omega)^d, \, \forall T \in \mathcal{T}_h, \tag{5.16b}$$

for $0 \le m \le 1$, $1 \le \ell \le r+1$. If in addition $\nabla Q_h \subset (D_h^1)^d$, then

$$|(r_h, \nabla \cdot (\mathbf{w} - \mathbf{j}_h \mathbf{w}))| \le C \left(\sum_{E \in \mathcal{E}_h} \alpha_E^{-1} h_{T_E}^{2r+1} |\mathbf{w}|_{r+1,\omega(T_E)}^2 \right)^{1/2} \left(J_h(r_h, r_h) \right)^{1/2} \tag{5.17}$$

for all $r_h \in Q_h$ and all $\mathbf{w} \in \mathbf{V} \cap H^{r+1}(\Omega)^d$.

Proof. Under the hypotheses (5.14b) and (5.15), it is shown in [MST07, Theorem 2.2] that there exists an interpolation operator $\mathbf{j}_h$ satisfying (5.16). It is constructed by setting $\mathbf{j}_h \mathbf{w} := \mathbf{j}_h^* \mathbf{w} + \mathbf{z}_h(\mathbf{w})$ where $\mathbf{z}_h(\mathbf{w})|_T \in \mathbf{V}_h(T) := Y_h(T)^d$ is defined locally by

$$(\mathbf{z}_h(\mathbf{w}), q_h)_T = (\mathbf{w} - \mathbf{j}_h^* \mathbf{w}, q_h)_T \quad \forall q_h \in \left(D_h^1(T)\right)^d$$

which immediately guarantees (5.16a). One can establish the local bound

$$\|\mathbf{z}_h(\mathbf{w})\|_{0,T} \leq \frac{1}{\beta_1} \|\mathbf{w} - \mathbf{j}_h^* \mathbf{w}\|_{0,T}$$

from which (5.16b) follows by invoking (5.14b) and an inverse inequality. It remains to prove that (5.17) holds true. As $\mathbf{j}_h \mathbf{w} = \mathbf{j}_h^* \mathbf{w} + \mathbf{z}_h(\mathbf{w})$, for $r_h \in Q_h$ and $\mathbf{w} \in \mathbf{V} \cap H^{r+1}(\Omega)^d$ one obtains

$$(r_h, \nabla \cdot (\mathbf{w} - \mathbf{j}_h \mathbf{w})) = -(r_h, \nabla \cdot \mathbf{z}_h(\mathbf{w})) = \sum_{T \in \mathcal{T}_h} (\nabla r_h, \mathbf{z}_h(\mathbf{w}))_T$$

$$= \sum_{T \in \mathcal{T}_h} (\nabla r_h, \mathbf{w} - \mathbf{j}_h^* \mathbf{w})_T$$

$$= -\sum_{T \in \mathcal{T}_h} (r_h, \nabla \cdot (\mathbf{w} - \mathbf{j}_h^* \mathbf{w}))_T$$

$$+ \sum_{E \in \mathcal{E}_h} \left\langle [r_h]_E, (\mathbf{w} - \mathbf{j}_h^* \mathbf{w}) \cdot n_E \right\rangle_E.$$

This calculation used (5.14a), $\mathbf{z}_h(\mathbf{w}) = 0$ on ∂T for all $T \in \mathcal{T}_h$, $\nabla(r_h|_T) \in \left(D_h^1(T)\right)^d$ and $\mathbf{w} - \mathbf{j}_h^* \mathbf{w} = 0$ on $\partial \Omega$. The first term on the right-hand side vanishes because of (5.14a). For $E \in \mathcal{E}_h \cap \partial T_E$, the scaled trace inequality

$$\|\mathbf{v}\|_{0,E} \leq C \left(h_{T_E}^{-1/2} \|\mathbf{v}\|_{0,T_E} + h_{T_E}^{1/2} |\mathbf{v}|_{1,T_E}\right) \qquad \forall \mathbf{v} \in H^1(T_E)$$

yields

$$\|\mathbf{w} - \mathbf{j}_h^* \mathbf{w}\|_{0,E} \leq C \left(h_{T_E}^{-1/2} h_{T_E}^{r+1} |\mathbf{w}|_{r+1,\omega(T_E)} + h_{T_E}^{1/2} h_{T_E}^r |\mathbf{w}|_{r+1,\omega(T_E)}\right)$$

$$\leq C h_{T_E}^{r+1/2} |\mathbf{w}|_{r+1,\omega(T_E)}$$

by applying (5.14b). The estimate (5.17) now follows by using the Cauchy-Schwarz inequality.   □

Remark 5.11. The bound (5.17) implies that the special interpolant $\mathbf{j}_h$ yields a discrete divergence property for continuous pressure approximations since $J_h(r_h, r_h) = 0$ for $r_h \in H^1(\Omega)$. A simple example of spaces that satisfy Assumption A is the "extended Mini-element family" given by $\mathbf{V}_h = P_r^{++}$, $Q_h = P_r$, and $D_h^1 = P_r^{\text{disc}}$.   ♣

Theorem 5.12. *Assume that the spaces $\mathbf{V}_h$, Q_h satisfy (5.2), (5.8) and Assumption A. Let the function $\mathbf{b}$ satisfy the regularity assumption of Lemma 5.3. Let the projection space D_h^1 be such that the associated fluctuation operator κ_h^1 satisfies (5.7) with $s \in \{r-1, r\}$. Let the user-chosen parameters satisfy*

$\gamma_T \sim 1$ and $\alpha_E \sim h_E$. Assume that $\tau_T \sim h_T^2$ for $s = r - 1$ and $\bar{C} h_T^2 \leq \tau_T \leq \bar{\bar{C}}$ for $s = r$. Let $(\mathbf{u}, p) \in (\mathbf{V} \cap H^{r+1}(\Omega)^d) \times (Q \cap H^r(\Omega))$ be the solution of (5.1) and $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ the solution of (5.4) with A_h replaced by A_h^*. Then there exists a positive constant C, which is independent of ν, σ and h, such that

$$|||(\mathbf{u} - \mathbf{u}_h, p - p_h)|||_* \leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2r} (\|\mathbf{u}\|_{r+1,T}^2 + \|p\|_{r,T}^2) \right)^{1/2}.$$

Proof. The proof of Lemma 5.1 is still valid for A_h^* since

$$A_h^* ((\mathbf{v}_h, q_h); (\mathbf{z}_h, 0)) = A_h ((\mathbf{v}_h, q_h); (\mathbf{z}_h, 0)).$$

Now our argument follows the proof of Theorem 5.4; here we mention only the necessary changes. The treatment of the additional term that appears only for discontinuous pressure approximations is standard:

$$|J_h(i_h p - p, r_h)| \leq (J_h(i_h p - p, i_h p - p))^{1/2} (J_h(r_h, r_h))^{1/2}$$

$$\leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2r} \|p\|_{r,T}^2 \right)^{1/2} |||(w_h, r_h)|||_*,$$

where we used $h_E \sim h_T$ for $E \subset \partial T$ and the same ideas as in the proof of Lemma 5.10 to estimate the interpolation error on cell boundaries. Replacements for the inequalities (5.10) and (5.12) are still needed; using (5.17) and $\alpha_E \sim h_E$, one gets

$$|(r_h, \nabla \cdot (\mathbf{u} - \mathbf{j}_h \mathbf{u}))| \leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2r} \|\mathbf{u}\|_{r+1,T}^2 \right)^{1/2} |||(\mathbf{w}_h, r_h)|||_*$$

for all $\sigma \geq 0$. Since the velocity interpolant has the additional orthogonality property (5.16a) relative to D_h^1, one can estimate the convection term after an integration by parts via

$$|((\mathbf{b} \cdot \nabla)(\mathbf{j}_h \mathbf{u} - \mathbf{u}), \mathbf{w}_h)|$$
$$= |(\mathbf{j}_h \mathbf{u} - \mathbf{u}, (\mathbf{b} \cdot \nabla)\mathbf{w}_h)| = |(\mathbf{j}_h \mathbf{u} - \mathbf{u}, \kappa_h^1(\mathbf{b} \cdot \nabla)\mathbf{w}_h)|$$
$$\leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2(r+1)} \tau_T^{-1} \|\mathbf{u}\|_{r+1,T}^2 \right)^{1/2} |||(\mathbf{w}_h, r_h)|||_*. \quad (5.18)$$

The statement of the theorem follows with $\tau_T^{-1} \leq \bar{C}^{-1} h_T^{-2}$. $\quad\square$

Examples that satisfy all hypotheses of Theorem 5.12 are given in Tables 5.3 and 5.4. Since the pairs (P_r, P_{r-1}) and (Q_r, P_{r-1}^{disc}) satisfy the inf-sup condition (5.2), the enriched-velocity versions of these pairs in Tables 5.3

Table 5.3. Families of order $r \geq 2$ with continuous pressure approximations

$\mathbf{V}_h$	Q_h	D_h^1	D_h^2	τ_T	γ_T	t	$\|\|\| \cdot \|\|\|$
P_r^+	P_{r-1}	P_{r-2}^{disc}	P_{t-1}^{disc}	$\sim h_T^2$	~ 1	$t \leq r-1-d$	$\mathcal{O}(h^r)$
P_r^{++}	P_{r-1}	P_{r-1}^{disc}	P_{t-1}^{disc}	$\bar{C}h_T^2 \leq \tau_T \leq \bar{\bar{C}}$	~ 1	$t \leq r-1-d$	$\mathcal{O}(h^r)$
Q_r	Q_{r-1}	Q_{r-2}^{disc}	Q_{t-1}^{disc}	$\sim h_T^2$	~ 1	$t \leq r-2$	$\mathcal{O}(h^r)$

and 5.4 satisfy (5.2) also. The enrichments have been chosen large enough to satisfy the inf-sup condition (5.15) of Assumption A, which implies the orthogonality property (5.16a) of the velocity interpolation for the given projection space D_h^1. See [GMT08, MST07] for a proof of (5.15). Finally, the largest possible projection space D_h^2 for continuous pressure approximations is the bubble part of the pressure space Q_h.

Table 5.4. Families of order $r \geq 2$ with discontinuous pressures and the modified stabilization term S_h^*

$\mathbf{V}_h$	Q_h	D_h^1	D_h^2	τ_T	γ_T	α_E	t	$\|\|\| \cdot \|\|\|_*$
Q_r	P_{r-1}^{disc}	Q_{r-2}^{disc}	P_{t-1}^{disc}	$\sim h_T^2$	~ 1	$\sim h_E$	$t \leq r$	$\mathcal{O}(h^r)$
Q_r^+	P_{r-1}^{disc}	P_{r-1}^{disc}	P_{t-1}^{disc}	$\bar{C}h_T^2 \leq \tau_T \leq \bar{\bar{C}}$	~ 1	$\sim h_E$	$t \leq r$	$\mathcal{O}(h^r)$

5.3.3 Methods of Order $r + 1/2$

For equal-order interpolations with $\mathbf{V}_h = (Y_h \cap H_0^1(\Omega))^d$ and $Q_h = Y_h \cap Q$, error estimates of order $\mathcal{O}((\nu^{1/2} + h^{1/2})h^r)$ were established in Section 4.4. Unfortunately, these pairs of finite elements are not inf-sup stable and an additional pressure stabilization (pressure-stabilized Petrov-Galerkin or PSPG; see [TMRS92]) was necessary. A careful reading of the proof of Theorem 5.4 shows that the critical term limiting the convergence order to r is $(p - i_h p, \nabla \cdot \mathbf{w}_h)$, which was estimated in (5.11). Thus an improved approximation of the pressure is the key to getting an improved error estimate. In this subsection we consider inf-sup stable pairs $(\mathbf{V}_h, Q_h)$ of finite element spaces that approximate both velocity and pressure by elements of order r.

Consider the two families of spaces listed in Table 5.5.

Theorem 5.13. *Assume that the spaces $\mathbf{V}_h$, Q_h, D_h^1, D_h^2 and the parameters τ_T, γ_T, α_E are chosen according to Table 5.5. Let the function $\mathbf{b}$ satisfy the regularity assumption of Lemma 5.3. Let $(\mathbf{u}, p) \in (\mathbf{V} \cap H^{r+1}(\Omega)^d) \times (Q \cap H^{r+1}(\Omega))$ and $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ be the solutions of (5.1) and (5.4) respectively, where A_h has been replaced by A_h^*. Then there exists a positive*

Table 5.5. Families of order $r + 1/2$ and the modified stabilization term S_h^*

$\mathbf{V}_h$	Q_h	D_h^1	D_h^2	τ_T	γ_T	α_E	r	t	$\|\|\| \cdot \|\|\|_*$
P_r^{++}	P_r	P_{r-1}^{disc}	P_{t-1}^{disc}	$\sim h_T$	$\sim h_T$	$= 0$	$r \geq 1$	$t \leq r - d$	$\mathcal{O}(h^{r+1/2})$
Q_r^+	P_r^{disc}	P_{r-1}^{disc}	P_{t-1}^{disc}	$\sim h_T$	$\sim h_T$	~ 1	$r \geq 2$	$t \leq r + 1$	$\mathcal{O}(h^{r+1/2})$

constant C, which is independent of ν, σ, and h, such that

$$\|\|(\mathbf{u} - \mathbf{u}_h, p - p_h)\|\|_* \leq C \left[\sum_{T \in \mathcal{T}_h} (\nu + h_T) h_T^{2r} \left(\|\mathbf{u}\|_{r+1,T}^2 + \|p\|_{r+1,T}^2 \right) \right]^{1/2}.$$

Proof. The inf-sup stability condition (5.2) is proved in [MT07]. Furthermore, the choice $D_h^1 = P_{r-1}^{\mathrm{disc}}$ guarantees (5.7) with $s = r$ and the consistency error becomes $\mathcal{O}(h^{r+1/2})$ for $\tau_T \leq \bar{C} h_T$.

Assumption A is satisfied for the pairs $(P_r^{++}, P_{r-1}^{\mathrm{disc}})$ and $(Q_r^+, P_{r-1}^{\mathrm{disc}})$; see [MST07]. Therefore we can use the improved estimate (5.18) for the convection term. Moreover, upper bounds for the sizes of the projection spaces D_h^2 follow from the size of the bubble parts of the pressure spaces P_r on simplices and P_r^{disc} on quadrilaterals and hexahedra. The choice $D_h^2 = P_{t-1}^{\mathrm{disc}}$ allows us to apply Lemmas 5.5 and 5.6; thus all the conditions (5.8) are satisfied.

As the pressure space is either P_r or P_r^{disc}, we have the following improved interpolation error estimate:

$$\|q - i_h q\|_{0,T} + h_T |q - i_h q|_{1,T} \leq C h_T^\ell \|q\|_{\ell,T} \quad \forall q \in H^\ell(T), \, 2 \leq \ell \leq r + 1,$$

for all $T \in \mathcal{T}_h$. Consequently the bound (5.12) can be sharpened:

$$(p - i_h p, \nabla \cdot \mathbf{w}_h) = (p - i_h p, \kappa_h^2 \nabla \cdot \mathbf{w}_h)$$

$$\leq C \left(\sum_{T \in \mathcal{T}_h} \gamma_T^{-1} h_T^{2r+2} \|p\|_{r+1,T}^2 \right)^{1/2} \|\|(\mathbf{w}_h, r_h)\|\|_*$$

$$\leq C \left(\sum_{T \in \mathcal{T}_h} h_T^{2r+1} \|p\|_{r+1,T}^2 \right)^{1/2} \|\|(\mathbf{w}_h, r_h)\|\|_*,$$

where $\gamma_T \sim h_T$ was used. The desired result follows. $\square$

Examples of elements that in the convection-dominated case ($\nu < h_T$) converge with order $r + 1/2$ are shown in Figures 5.1 and 5.2.
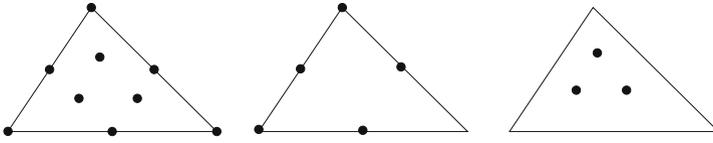
Fig. 5.1. Triangular elements of order 5/2: $\mathbf{V}_h$, Q_h, and D_h^1 (from left to right)
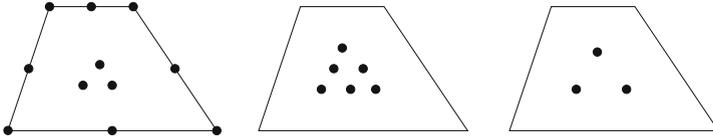
Fig. 5.2. Quadrilateral elements of order 5/2: $\mathbf{V}_h$, Q_h, and D_h^1 (from left to right)

6

Mass Conservation for Coupled Flow-Transport Problems

In this chapter we examine mass-conservation properties of finite element discretizations of coupled flow-transport problems. The system under consideration is described by the unsteady incompressible Navier-Stokes equations and a time-dependent transport equation; see [GS00a, GS00b, Hir88, Hir90] for models where this combination arises. The incompressibility constraint implies that global mass is conserved in the weak solution of the transport equation. Since the discretized velocity only satisfies a discrete incompressibility constraint, global mass is in general conserved only approximately in the numerical scheme. We shall investigate conditions under which discrete global mass conservation can be guaranteed.

6.1 A Model Problem

We consider the simplest case of a coupled flow-transport problem in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. The system is described by the unsteady incompressible Navier-Stokes equations

$$\mathbf{u}_t - \nu \triangle \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega \times (0, T], \tag{6.1a}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \times (0, T], \tag{6.1b}$$

$$\mathbf{u} = \mathbf{u}_b \quad \text{on } \partial\Omega \times (0, T], \tag{6.1c}$$

$$\mathbf{u}(0) = \mathbf{u}_0 \quad \text{in } \Omega, \tag{6.1d}$$

and the time-dependent transport equation

$$c_t - \varepsilon \triangle c + \mathbf{u} \cdot \nabla c = g \quad \text{in } \Omega \times (0, T], \tag{6.2a}$$

$$(c\mathbf{u} - \varepsilon \nabla c) \cdot \mathbf{n} = c_I \, \mathbf{u} \cdot \mathbf{n} \quad \text{on } \Gamma_- \times (0, T], \tag{6.2b}$$

$$\varepsilon \nabla c \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_* \times (0, T], \tag{6.2c}$$

$$c(0) = c_0 \quad \text{in } \Omega. \tag{6.2d}$$

Here $\mathbf{u}$ and p denote the velocity and the pressure of the fluid, ν and ε are small positive numbers, and $T > 0$ is the final time. The boundary $\partial\Omega$ is divided between the inflow boundary $\Gamma_- := \{x \in \partial\Omega : \mathbf{u} \cdot \mathbf{n} < 0\}$ and the remaining part of the boundary $\Gamma_* := \partial\Omega \setminus \Gamma_-$, where $\mathbf{n}$ is the outward-pointing unit normal. Furthermore, c is the concentration of a species transported with the flow field and c_I its concentration at the inflow boundary Γ_-. We assume that the given velocity field $\mathbf{u}_b$ on the boundary $\partial\Omega$ is the restriction of a divergence-free function that is also denoted by $\mathbf{u}_b$. The initial velocity $\mathbf{u}_0$ satisfies the incompressibility constraint $\nabla \cdot \mathbf{u}_0 = 0$.

Various discretization methods for the unsteady incompressible Navier-Stokes equations and the transport equation have been developed in previous chapters for the realistic and important cases where $\nu \ll 1$ and $\varepsilon \ll 1$. Here we shall study the mass conservation of the discretized transport equation when using stabilized schemes. For simplicity of notation we confine our attention to the semi-discretization in space of the problems (6.1) and (6.2). The results can be extended to the fully discretized problems by using discontinuous Galerkin methods in time.

6.2 Continuous and Discrete Mass Conservation

Set $W := H^1(\Omega)$. Let $(\cdot, \cdot)$ and $\langle\cdot, \cdot\rangle_\Gamma$ denote the L^2 inner products on Ω and Γ respectively. A weak formulation of the transport problem (6.2) is:

Find $c(t) \in W$ such that for all $\varphi \in W$ one has $(c(0) - c_0, \varphi) = 0$ and

$$\frac{d}{dt}(c, \varphi) + \varepsilon(\nabla c, \nabla\varphi) + \mathbf{u} \cdot \nabla c, \varphi) - \langle c\, \mathbf{u} \cdot \mathbf{n}, \varphi\rangle_{\Gamma_-}$$
$$= (g, \varphi) - \langle c_I\, \mathbf{u} \cdot \mathbf{n}, \varphi\rangle_{\Gamma_-}. \qquad (6.3)$$

On setting $\varphi \equiv 1$ and using the incompressibility constraint $\nabla \cdot \mathbf{u} = 0$, one derives from (6.3) the global mass conservation property

$$\frac{d}{dt}\int_\Omega c\, dx + \int_{\Gamma_-} c_I\, \mathbf{u} \cdot \mathbf{n}\, d\gamma + \int_{\Gamma_*} c\, \mathbf{u} \cdot \mathbf{n}\, d\gamma = \int_\Omega g\, dx. \qquad (6.4)$$

Assume that the domain Ω is polyhedral and is subdivided into simplicial elements K by a family $\{\mathcal{T}_h\}_{h>0}$ of shape-regular triangulations of Ω. Let $W_h \subset W$ be some finite element space for approximating the concentration c. Then the standard Galerkin discretization of (6.3) is:

Find $c_h(t) \in W_h$ such that for all $\varphi_h \in W_h$ one has $(c_h(0) - c_0, \varphi_h) = 0$ and

$$\frac{d}{dt}(c_h, \varphi_h) + \varepsilon(\nabla c_h, \nabla\varphi_h) + (\mathbf{u}_h \cdot \nabla c_h, \varphi_h) - \langle c_h\, \mathbf{u}_h \cdot \mathbf{n}, \varphi_h\rangle_{\Gamma_-}$$
$$= (g, \varphi_h) - \langle c_I\, \mathbf{u}_h \cdot \mathbf{n}, \varphi_h\rangle_{\Gamma_-}. \qquad (6.5)$$

Note that the divergence-free vector field $\mathbf{u}$ of (6.3) has been replaced in (6.5) by some approximation $\mathbf{u}_h$ that is in general discontinuous. On setting $\varphi_h = 1$ and integrating by parts over each element, one arrives at

$$\frac{d}{dt}\int_\Omega c_h\,dx + \int_{\Gamma_-} c_I \mathbf{u}_h \cdot \mathbf{n}\,d\gamma + \int_{\Gamma_*} c_h \mathbf{u}_h \cdot \mathbf{n}\,d\gamma = \int_\Omega g\,dx + m_h(c_h, \mathbf{u}_h). \quad (6.6)$$

In this equation, compared with the global mass conservation of (6.4) on the continuous level, the additional term

$$m_h(c_h, \mathbf{u}_h) := \sum_{K \in \mathcal{T}_h} (c_h, \nabla \cdot \mathbf{u}_h)_K + \sum_{E \in \mathcal{E}_h} \langle c_h, [\mathbf{u}_h]_E \cdot \mathbf{n}_E\rangle_E \quad (6.7)$$

appears. Here $\mathcal{E}_h$ is the set of inner faces E in $\mathcal{T}_h$, and $(\cdot,\cdot)_K$ and $\langle\cdot,\cdot\rangle_E$ are the L^2 inner products on K and E respectively. With each $E \in \mathcal{E}_h$ we associate an arbitrary but fixed unit normal $\mathbf{n}_E$ and define the jump of a quantity ψ across the common face E of the two adjacent elements K and $\widetilde{K}$ by

$$[\psi]_E := (\psi|_{\widetilde{K}})\big|_E - (\psi|_K)\big|_E$$

where $\mathbf{n}_E$ is an outward-pointing unit normal to ∂K. The discrete counterpart of the global mass conservation equation (6.4) is equation (6.6) with $m_h(c_h, \mathbf{u}_h) = 0$. In Section 6.4 we shall discuss conditions under which $m_h(c_h, \mathbf{u}_h)$ vanishes.

Discrete mass conservation is not guaranteed by the standard Galerkin discretization unless $m_h(c_h, \mathbf{u}_h) = 0$. What happens if one uses stabilized schemes to solve (6.2)? Let us consider such schemes of the following type:

Find $c_h(t) \in W_h$ such that for all $\varphi_h \in W_h$ one has $c_h(0) - c_0, \varphi_h) = 0$ and

$$\frac{d}{dt}(c_h, \varphi_h) + \varepsilon(\nabla c_h, \nabla \varphi_h) + (\mathbf{u}_h \cdot \nabla c_h, \varphi_h) - \langle c_h\,\mathbf{u}_h \cdot \mathbf{n}, \varphi_h\rangle_{\Gamma_-}$$
$$+ S_h(c_h, \varphi_h) = (g, \varphi_h) - \langle c_I\,\mathbf{u}_h \cdot \mathbf{n}, \varphi_h\rangle_{\Gamma_-}.$$

In the streamline diffusion method (SDFEM) of Section III.4.3, weighted residuals of the strong form of the differential equation are added. That is, one has

$$S_{SD}(c_h, \varphi_h) := \sum_{K \in \mathcal{T}_h} \delta_K(c_{h,t} - \varepsilon\triangle c_h + \mathbf{u}_h \cdot \nabla c_h - g, \mathbf{u}_h \cdot \nabla \varphi_h)_K$$

with user-chosen SD parameters δ_K.

The subgrid modelling method of Section 4.5 considers a subspace of resolvable scales W_H in the approximating space W_h, together with a projector $P_H : W_h \to W_H$. The non-resolvable scales are stabilized by adding

$$S_{SGS}(c_h, \varphi_h) := \sum_{K \in \mathcal{T}_h} \tau_K\big(\nabla(\mathrm{id} - P_H)c_h, \nabla(\mathrm{id} - P_H)\varphi_h\big)_K$$

to the standard Galerkin method (6.5), where the τ_K are user-chosen parameters.

Finally, recall the local projection stabilization (LPS) of Section III.3.3.1. This relies on a local projection operator $\widetilde{P} : W_h \to D_h$ into a proper subspace of discontinuous finite elements. The stabilizing term added is

$$S_{LPS}(c_h, \varphi_h) := \sum_{K \in \mathcal{T}_h} \tau_K \big((\mathrm{id} - \widetilde{P})\nabla c_h, (\mathrm{id} - \widetilde{P})\nabla \varphi_h\big)_K$$

with user-chosen parameters τ_K.

In all these cases the stabilizing terms vanish if $\varphi_h \equiv 1$. Hence global mass conservation on the discrete level will be guaranteed for both the standard Galerkin and stabilized methods if the additional term $m_h(c_h, \mathbf{u}_h)$ in (6.6) vanishes.

6.3 Approximated Incompressible Flows

Set $\mathbf{V} = H_0^1(\Omega)^d$, $M = L^2(\Omega)$, and $Q = \{q \in M : (q, 1) = 0\}$. Then a weak formulation of the unsteady incompressible Navier-Stokes problem (6.1) is:

Find $\big(\mathbf{u}(t), p(t)\big) \in (\mathbf{u}_b + \mathbf{V}) \times Q$ such that

$$(\mathbf{u}(0) - \mathbf{u}_0, \mathbf{v}) = 0 \qquad \forall \mathbf{v} \in M^d, \quad (6.8a)$$

$$\frac{d}{dt}(\mathbf{u}, \mathbf{v}) + \nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + \big((\mathbf{u} \cdot \nabla)\mathbf{u}, \mathbf{v}\big) - (\nabla \cdot \mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{V}, \quad (6.8b)$$

$$(\nabla \cdot \mathbf{u}, q) = 0 \qquad \forall q \in Q. \quad (6.8c)$$

Our assumption that $\mathbf{u}_b$ is the restriction of a divergence-free function yields

$$(\nabla \cdot \mathbf{u}, 1) = \langle \mathbf{u} \cdot \mathbf{n}, 1 \rangle_{\partial \Omega} = \langle \mathbf{u}_b \cdot \mathbf{n}, 1 \rangle_{\partial \Omega} = (\nabla \cdot \mathbf{u}_b, 1) = 0. \qquad (6.9)$$

This, combined with (6.8c), implies that $(\nabla \cdot \mathbf{u}, q) = 0$ for all $q \in L^2(\Omega)$.

Consider first the inf-sup stable discretizations of the problem (6.8) that were discussed in Chapters 2 and 5. Let $\mathbf{V}_h \subset \mathbf{V}$, $M_h \subset M$ and $Q_h = M_h \cap Q$ be finite element spaces such that the inf-sup condition

$$\inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{(\nabla \cdot \mathbf{v}_h, q_h)}{|\mathbf{v}_h|_1 \, \|q_h\|_0} \geq \beta \qquad (6.10)$$

is satisfied with a positive constant β that is independent of the mesh size parameter h. Using the discrete spaces $\mathbf{V}_h$ and M_h, the standard Galerkin discretization of (6.8) is:

Find $\big(\mathbf{u}_h(t), p_h(t)\big) \in (\mathbf{u}_{b,h} + \mathbf{V}_h) \times M_h$ such that

$$(\mathbf{u}_h(0) - \mathbf{u}_0, \mathbf{v}_h) = 0 \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \qquad (6.11a)$$

$$\frac{d}{dt}(\mathbf{u}_h, \mathbf{v}_h) + \nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h)$$
$$+ \big((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \mathbf{v}_h\big) - (\nabla \cdot \mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \qquad (6.11b)$$
$$(\nabla \cdot \mathbf{u}_h, q_h) = 0 \qquad \forall q_h \in Q_h, \qquad (6.11c)$$

where $\mathbf{u}_{b,h}$ is some approximation of $\mathbf{u}_b$ for which $\langle \mathbf{u}_{b,h} \cdot \mathbf{n}, 1 \rangle_{\partial\Omega} = 0$. As a consequence,

$$(\nabla \cdot \mathbf{u}_h, 1) = \sum_{K \in \mathcal{T}_h} (\nabla \cdot \mathbf{u}_h, 1)_K = \langle \mathbf{u}_{b,h} \cdot \mathbf{n}, 1 \rangle_{\partial\Omega} - \sum_{E \in \mathcal{E}_h} \langle [\mathbf{u}_h]_E \cdot \mathbf{n}_E, 1 \rangle_E$$
$$= - \sum_{E \in \mathcal{E}_h} \langle [\mathbf{u}_h]_E \cdot \mathbf{n}_E, 1 \rangle_E .$$

Thus if the normal components of $\mathbf{u}_h$ are continuous over the cell faces E, we get the discrete analogue of (6.9), viz., $(\nabla \cdot \mathbf{u}_h, q_h) = 0$ for all $q_h \in M_h$.

When discretizing the Navier-Stokes problem by inf-sup stable finite elements, one has to make a fundamental choice between continuous and discontinuous pressure approximations. The relation (6.11c) says that the incompressibility constraint (6.1b) is satisfied only in an approximate sense. Nevertheless, if discontinuous pressure approximations are used, then mass conservation in the fluid is satisfied more locally since functions with support within one element can be used as test functions.

Now we turn to schemes that are designed to stabilize the twin effects of dominant convection and instabilities caused by finite element pairs $(\mathbf{V}_h, Q_h)$ that fail to satisfy (6.10) – see Chapters 3 and 4. In particular, equal-order interpolation of velocity and pressure that is stabilized by the local projection method (LPS) or by the streamline diffusion method (SDFEM) will be our focus. A common feature of these stabilization methods is an additional stabilizing term in the discrete mass balance equation (6.11c) of the Navier-Stokes system, but this produces an additional error in the mass conservation of the transport equation. Indeed, in the LPS method, (6.11c) is replaced by

$$(\nabla \cdot \mathbf{u}_h, q_h) + \sum_{M \in \mathcal{M}_h} \alpha_M (\kappa_h \nabla p_h, \kappa_h \nabla q_h)_M = 0 \quad \forall q_h \in Q_h,$$

where $\kappa_h = \mathrm{id} - \pi_h$ is the fluctuation operator defined via the local projection $\pi_h : L^2(M)^d \to D_h(M)^d$. Here $D_h(M)$ is a finite element space associated with the family of macro-triangulations $\{\mathcal{M}_h\}_{h>0}$ of Ω into macro-cells and the α_M are user-chosen parameters. Note that in the case of enriched approximation spaces it is possible to have $\mathcal{M}_h = \mathcal{T}_h$. For the SDFEM the discrete mass balance condition (6.11c) is modified to

$$\sum_{K \in \mathcal{T}_h} \tau_K \big((\mathbf{u}_h)_t - \nu \Delta \mathbf{u}_h + (\mathbf{u}_h \cdot \nabla) \mathbf{u}_h + \nabla p_h - \mathbf{f}, \nabla q_h \big)_K$$

$$+ \sum_{E \in \mathcal{E}_h} \alpha_E \big\langle [p_h]_E, [q_h]_E \big\rangle_E + (\nabla \cdot \mathbf{u}_h, q_h) = 0 \quad \forall q_h \in Q_h,$$

which is even more complicated than that for the LPS scheme.

To avoid the discretization error caused by these additional terms in the discrete mass balance equation, one should try to separate the treatments of the two instability phenomena: dominant convection and unstable finite element pairs for approximating velocity and pressure. Such separation techniques have been considered, e.g., by [BH06, FJMT07, GLOS05]. In the following we restrict ourselves to the solution of the Navier-Stokes equations by inf-sup stable conforming finite element pairs and stabilization methods that do not modify the discrete mass balance (6.11c). In this case, the computed velocity field $\mathbf{u}_h(t) \in \mathbf{u}_{b,h} + \mathbf{V}_h$ lies in $H^1(\Omega)^d$ and is discretely divergence-free in the sense that

$$(\nabla \cdot \mathbf{u}_h, q_h) = 0 \quad \forall q_h \in M_h.$$

While (6.11c) implies that this relation is valid for all $q_h \in Q_h \subset M_h$, the choice of approximation $\mathbf{u}_{b,h}$ of the boundary data $\mathbf{u}_b$ guarantees its satisfaction for all $q_h \in M_h$.

6.4 Mass-Conservative Methods

In Section 6.2 we saw that the mass of a species transported with the flow is conserved on the discrete level if and only if the term $m_h(c_h, \mathbf{u}_h)$ of (6.7) vanishes. Recall that on the continuous level the term $m_h(c_h, \mathbf{u})$ vanished owing to the incompressibility condition $\nabla \cdot \mathbf{u} = 0$ and $[\mathbf{u}]_E = 0$ for all inner faces $E \in \mathcal{E}_h$. In the following subsections several approaches that ensure $m_h(c_h, \mathbf{u}_h) = 0$ will be examined.

6.4.1 Higher-Order Flow Approximation

Let us assume that the transport equation is solved by a method of order $r \geq 1$, i.e., the approximation error in space satisfies

$$\inf_{\varphi_h \in W_h} |c - \varphi_h|_m \leq C \, h^{r+1-m} |c|_{r+1}$$

for all $c \in H^{r+1}(\Omega)$ and $0 \leq m \leq r + 1$. One example is the space W_h of continuous piecewise polynomials of degree at most r.

Consider first conforming finite element discretizations of the Navier-Stokes equations. Then the sum over inner faces in $m_h(c_h, \mathbf{u}_h)$ vanishes since $[\mathbf{u}_h]_E = 0$ for all $E \in \mathcal{E}_h$. One also observes that the sum over all cells K in $m_h(c_h, \mathbf{u}_h)$ vanishes if c_h belongs to the approximation space M_h for the

pressure, which is true if $W_h \subset M_h$. Thus mass conservation of the species transported with the flow can be achieved if one chooses $M_h = W_h$, e.g., the space P_r of continuous piecewise polynomials of degree at most r. Then the velocity space has to be rich enough to satisfy the inf-sup condition (6.10). One possible choice would be the vector-valued version of the space P_{r+1} of continuous piecewise polynomials of degree at most $r+1$. The pair (P_{r+1}^d, P_r) is called the Taylor-Hood element and is known to be inf-sup stable [GR86]. We obtain a discretization of $(\mathbf{u}_h, p_h, c_h)$ in $P_{r+1}^d \times P_r \times P_r$. Figure 6.1 shows the relevant degrees of freedoms when $d = 2$ and $r = 2$. That is, the Navier-
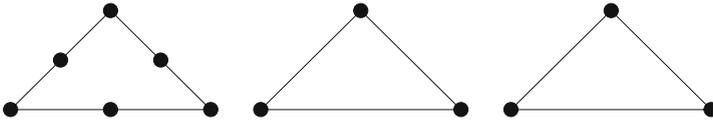
Fig. 6.1. Approximation spaces for velocity, pressure and concentration in the two-dimensional case when second-order Taylor-Hood elements for the flow problem are combined with piecewise linear elements for the transport equation

Stokes problem is discretized by a method which is of order $r+1$ whereas the transport equation is approximated by a lower-order method that is of order r. One therefore expects an error estimate of the form

$$\|\mathbf{u} - \mathbf{u}_h\|_1 + \|p - p_h\|_0 + \|c - c_h\|_1 \leq C\left(h^{r+1} + h^{r+1} + h^r\right)$$

which is suboptimal with respect to the flow problem. Another inf-sup stable example for the two-dimensional case is displayed in Figure 6.2 and turns out to be also mass conservative. Here the pressure and the concentration are
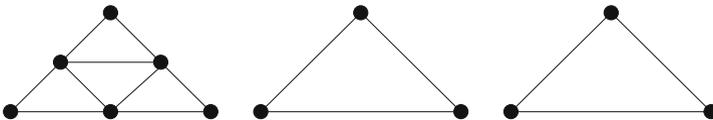
Fig. 6.2. First-order approximation spaces for the flow and the transport problem in the two-dimensional case: piecewise linear elements for pressure and concentration with piecewise linear elements on the next finer mesh level for the velocity

discretized by continuous piecewise linear functions on the triangulation $\mathcal{T}_h$ while each velocity component is approximated by continuous piecewise linear functions on the next refinement level $\mathcal{T}_{h/2}$. Thus the discretization $(\mathbf{u}_h, p_h, c_h)$ lies in $(4P_1)^2 \times P_1 \times P_1$. Although the number of degrees of freedom is the same as for $P_2^2 \times P_1 \times P_1$, the solution $(\mathbf{u}, p)$ of the Navier-Stokes equation is

now approximated only to first order. For flows at higher Reynolds number, the method could be combined with the upwind technique of Section III.3.1.

For quadrilateral or hexahedral elements, mass-conservative methods can be derived in a similar way. This technique of using higher-order approximations for the flow problem to get mass-conservative schemes works for both continuous and discontinuous pressure approximations.

With nonconforming finite element discretizations of the Navier-Stokes equations on simplices, more care is needed as now one may not have $[\mathbf{u}_h]_E = 0$ over the inner faces $E \in \mathcal{E}_h$. Nevertheless a careful investigation of the consistency error [MT05] shows that for a method of order $r + 1$ one needs the velocity to satisfy

$$\big\langle [\mathbf{u}_h]_E, r_h \big\rangle_E = 0 \quad \forall r_h \in P_r(E)^d, \ \forall E \in \mathcal{E}_h.$$

This is just sufficient to guarantee that the sum over inner faces in $m_h(c_h, \mathbf{u}_h)$ vanishes; see (6.7). If the discretization is completed by discontinuous elements of order r for the pressure and continuous elements of order r for the concentration, then in (6.7) one has $m_h(c_h, \mathbf{u}_h) = 0$. Figure 6.3 indicates a variant of this method for the case $r = 3$. For nonconforming finite element
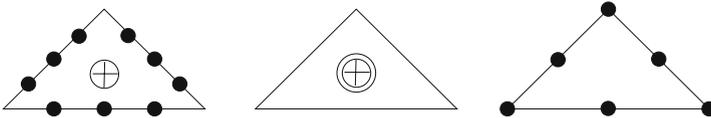
Fig. 6.3. A combined non-conforming/conforming discretization of the coupled flow-transport problem: velocity approximated by enriched nonconforming P_3^2, pressure by discontinuous P_2^{disc} and concentration by continuous P_2

discretizations of the Navier-Stokes equations on quadrilateral or hexahedral cells see [Mat07].

In summary, using a discretization of the flow problem that is one order higher produces global mass conservation of the transport equation. But from a practical point of view this technique is unattractive since it is too costly to discretize the Navier-Stokes equations by an order $r + 1$ method when the transport equation is discretized using an order r method, as the combined method will be only order r.

6.4.2 Post-Processing of the Discrete Velocity

An alternative way of ensuring exact mass balance on the discrete level is to replace the discrete velocity solution $\mathbf{u}_h$ by a different discrete function $\mathbf{w}_h$ that is close to $\mathbf{u}_h$. This approach was proposed in [CKS05b] for the local discontinuous Galerkin method applied to flow problems and we now describe

it in detail. Instead of the standard Galerkin formulation (6.5) for the weak form of the transport equation, one solves the problem

Find $c_h(t) \in W_h$ such that for all $\varphi_h \in W_h$ one has $(c_h(0) - c_0, \varphi_h) = 0$ and

$$
\begin{aligned}
\frac{d}{dt}(c_h, \varphi_h) + \varepsilon(\nabla c_h, \nabla \varphi_h) + (\mathbf{w}_h \cdot \nabla c_h, \varphi_h) & \\
- \langle c_h \, \mathbf{w}_h \cdot \mathbf{n}, \varphi_h \rangle_{\Gamma_-} = (g, \varphi_h) - \langle c_I \, \mathbf{w}_h \cdot \mathbf{n}, \varphi_h \rangle_{\Gamma_-}, & \quad (6.12)
\end{aligned}
$$

where $\mathbf{w}_h$ will be specified below.

Let the Navier-Stokes equations be discretized by the inf-sup stable finite element pair $\left(\left(P_r^{\text{bubble}} \right)^d, P_{r-1}^{\text{disc}} \right)$ that comprises the velocity space of continuous piecewise polynomials of degree at most r enriched with certain bubble functions and the pressure space of discontinuous piecewise polynomials of degree at most $r - 1$; see [GR86]. This stable pair is illustrated in Figure 6.4
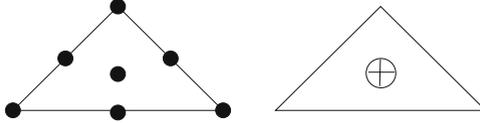
Fig. 6.4. Inf-sup stable finite element pair for the flow problem in the two-dimensional case: velocity approximated by $\left(P_2^{\text{bubble}} \right)^2$ and pressure by P_1^{disc}

for the case $d = 2$ and $r = 2$. The function $\mathbf{w}_h \in \left(P_r^{\text{disc}} \right)^d$ that acts as an approximate velocity field in the transport equation is constructed by post-processing. To this end, define on each element $K \in \mathcal{T}_h$ the vector-valued local interpolation operator $\Pi_K : H^1(K)^d \to P_r(K)^d$ by

$$
\begin{aligned}
\langle (\Pi_K \mathbf{v}) \cdot \mathbf{n}_K, \varphi \rangle_E &= \langle \mathbf{v} \cdot \mathbf{n}_K, \varphi \rangle_E & \forall E \subset \partial K, \, \varphi \in P_r(E), & \quad (6.13a) \\
(\Pi_K \mathbf{v}, \nabla \varphi) &= (\mathbf{v}, \nabla \varphi) & \forall \varphi \in P_{r-1}(K), & \quad (6.13b) \\
(\Pi_K \mathbf{v}, \boldsymbol{\psi}) &= (\mathbf{v}, \boldsymbol{\psi}) & \forall \boldsymbol{\psi} \in \boldsymbol{\Psi}_r(K), & \quad (6.13c)
\end{aligned}
$$

where

$$
\boldsymbol{\Psi}_r(K) := \left\{ \boldsymbol{\psi} \in L^2(K)^d \, : \, (DF_K^T \boldsymbol{\psi}) \circ F_K \in \widehat{\boldsymbol{\Psi}}_r \right\}
$$

with

$$
\widehat{\boldsymbol{\Psi}}_r := \left\{ \hat{\boldsymbol{\psi}} \in P_r(\widehat{K})^d \, : \, \nabla \cdot \hat{\boldsymbol{\psi}} = 0 \text{ in } \widehat{K}, \, \hat{\boldsymbol{\psi}} \cdot \mathbf{n}_{\widehat{K}} = 0 \text{ on } \partial \widehat{K} \right\}.
$$

In the above formulas, we used the reference transformation $F_K : \widehat{K} \to K$ which is a bijective mapping from the reference cell $\widehat{K}$ onto the cell K. Furthermore, $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi} \circ F_K$. The space $P_r(K)^d$ has dimension

$$\dim \left(P_r(K)^d \right) = d \begin{pmatrix} r + d \\ d \end{pmatrix}$$

and we have to fix the degrees of freedom by (6.13). Conditions (6.13a) and (6.13b) give

$$(d+1) \begin{pmatrix} r + d - 1 \\ d - 1 \end{pmatrix} + \begin{pmatrix} r + d - 1 \\ d \end{pmatrix} - 1$$

$$= \begin{cases} (r^2 + 7r + 4)/2 & \text{if } d = 2, \\ (r^3 + 15r^2 + 38r + 18)/6 & \text{if } d = 3 \end{cases}$$

linear equations since the gradient of the constant function in $P_{r-1}(K)$ is zero. In the case $d = 2$ the space $\widehat{\boldsymbol{\Psi}}_r$ is characterized quite simply [BF91] as

$$\widehat{\boldsymbol{\Psi}}_r = \left\{ \hat{\psi} = \ \text{curl} \ \left(\hat{\lambda}_1 \hat{\lambda}_2 \hat{\lambda}_3 \varphi \right) \ : \varphi \in P_{r-2} \right\}$$

where $\hat{\lambda}_1 \hat{\lambda}_2 \hat{\lambda}_3$ is the cubic bubble function that vanishes on the boundary of the reference cell $\widehat{K}$. Hence $\dim\{\widehat{\boldsymbol{\Psi}}_r\} = r(r-1)/2$ in the case $d = 2$. It is more delicate to determine the dimension of the space $\{\widehat{\boldsymbol{\Psi}}_r\}$ when $d = 3$, but it turns out that for $d = 2$ and 3 the number of equations in (6.13) equals the number of degrees of freedom. See [BF91] for a proof of the P_r^d-unisolvence of the degrees of freedom.

An interpolation operator $\mathbb{P}$ was introduced by [CKS05b] in a general framework of local discontinuous Galerkin methods; in (6.13) we have adapted this operator to our situation and called it Π_K. The interpolation operator Π_K satisfies

$$\Pi_K \mathbf{v} = \mathbf{v} \quad \forall \mathbf{v} \in P_r(K)^d$$

owing to the P_r^d-unisolvence of its degrees of freedom. One can show by direct computation that

$$\Pi_K \mathbf{v} = S_K \mathbb{P}^{BDM} S_K^{-1} \mathbf{v}$$

where $S_K : H^1(\widehat{K})^d \to H^1(K)^d$ is the Piola mapping defined by

$$(S_K \hat{\mathbf{v}})(x) = (\det DF_K)^{-1} DF_K \ \hat{\mathbf{v}} \big(F_K^{-1}(x) \big)$$

and $\mathbb{P}^{BDM}$ is the BDM projection studied in [BDM85]. The equivalence of norms in finite-dimensional spaces implies the stability of the local interpolation operator Π_K on $P_r(K)^d$, i.e.,

$$\|\Pi_K \mathbf{v}\|_{1,K} \leq C \|\mathbf{v}\|_{1,K} \quad \forall v \in P_r(K)^d, \ K \in \mathcal{T}_h.$$

The local interpolation operators Π_K can be assembled to form a global interpolation operator Π_h by setting

$$(\Pi_h \mathbf{v})|_K := \Pi_K (\mathbf{v}|_K) \quad \forall K \in \mathcal{T}_h.$$

In general the function $\Pi_h \mathbf{v}$ will not belong to $H^1(\Omega)^d$ but to the space $\left(P_r^{\mathrm{disc}}\right)^d$ of discontinuous piecewise polynomials of degree at most r in each component.

We show that the post-processed solution $\Pi_h \mathbf{u}_h$ is piecewise divergence-free. Starting from the incompressibility constraint (6.11c), one uses the conditions (6.13a) and (6.13b) in the definition of Π_K to obtain

$$0 = (\nabla \cdot \mathbf{u}_h, q_h)_K = -(\mathbf{u}_h, \nabla q_h)_K + \langle \mathbf{u}_h \cdot \mathbf{n}_K, q_h \rangle_{\partial K}$$
$$= -(\Pi_K \mathbf{u}_h, \nabla q_h)_K + \langle (\Pi_K \mathbf{u}_h) \cdot \mathbf{n}_K, q_h \rangle_{\partial K}$$
$$= (\nabla \cdot \Pi_K \mathbf{u}_h, q_h)_K,$$

since $\nabla q_h \in P_{r-1}(K)^d$ and $q_h|_E \in P_r(E)$ for all faces $E \subset \partial K$.

Furthermore, observe that the function α defined piecewise by

$$\alpha|_K := \nabla \cdot \Pi_K \mathbf{u}_h$$

belongs to Q_h. Since $\alpha|_K \in P_{r-1}(K)$, we have to show only that α has zero integral mean over Ω. Indeed, one has

$$\int_\Omega \alpha \, dx = \sum_{K \in \mathcal{T}_h} \int_K \nabla \cdot \Pi_K \mathbf{u}_h \, dx = \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\Pi_K \mathbf{u}_h) \cdot \mathbf{n}_K \, d\gamma$$
$$= \sum_{K \in \mathcal{T}_h} \mathbf{u}_h \cdot \mathbf{n}_K \, d\gamma = \int_{\partial \Omega} \mathbf{u}_h \cdot \mathbf{n} \, d\gamma = 0$$

where the condition (6.13a) was invoked.

Thus α can be used as a pressure test function in (6.11c). We thence obtain

$$0 = (\nabla \cdot \Pi_h \mathbf{u}_h, \alpha) = \sum_{K \in \mathcal{T}_h} (\nabla \cdot \Pi_h \mathbf{u}_h, \nabla \cdot \Pi_h \mathbf{u}_h)_K$$

so $\nabla \cdot \Pi_h \mathbf{u}_h|_K \equiv 0$, i.e., the post-processed velocity solution is piecewise divergence-free.

The modified convection field in (6.12) is chosen to be $\mathbf{w}_h := \Pi_h \mathbf{u}_h$. This ensures that the first term in $m_h(c_h, \mathbf{w}_h)$ vanishes. Moreover, the normal component of $\mathbf{w}_h$ has no jumps across inner faces because of condition (6.13a) in the definition of Π_K. For, given any $\varphi \in P_r(E)$, one has

$$\langle [\Pi_h \mathbf{u}_h]_E \cdot \mathbf{n}_E, \varphi \rangle_E = \langle \Pi_{\widetilde{K}} \mathbf{u}_h \cdot \mathbf{n}_E, \varphi \rangle_E - \langle \Pi_K \mathbf{u}_h \cdot \mathbf{n}_E, \varphi \rangle_E$$
$$= \langle \mathbf{u}_h \cdot \mathbf{n}_E, \varphi \rangle_E - \langle \mathbf{u}_h \cdot \mathbf{n}_E, \varphi \rangle_E = 0$$

where K and $\widetilde{K}$ are the two elements adjacent to E. As $[\Pi_h \mathbf{u}_h]_E \in P_r(E)^d$, we conclude that $[\Pi_h \mathbf{u}_h]_E = 0$. That is, the second term of $m_h(c_h, \mathbf{w}_h)$ also vanishes.

As regards the approximation order of $\mathbf{w}_h = \Pi_h \mathbf{u}_h$, in the broken $H^1(\Omega)^d$ norm one gets

$$\|\mathbf{u} - \Pi_h \mathbf{u}_h\|_{1,h} \leq \|\mathbf{u} - I_h \mathbf{u}\|_{1,h} + \|\Pi_h(I_h \mathbf{u}) - \Pi_h \mathbf{u}_h\|_{1,h}$$
$$\leq Ch^r |\mathbf{u}|_{r+1} + C(\|I_h \mathbf{u} - \mathbf{u}\|_{1,h} + \|\mathbf{u} - \mathbf{u}_h\|_{1,h})$$
$$\leq Ch^r |\mathbf{u}|_{r+1} + C(\mathbf{u}, p)h^r,$$

where I_h is the standard P_r interpolation operator for vector-valued functions and $\mathbf{u}_h$ is the discrete velocity field computed by the $\left((P_r^{\text{bubble}})^d, P_{r-1}^{\text{disc}}\right)$ element. The above calculation uses the triangle inequality, the approximation property of the P_r interpolation operator I_h, the stability property of the post-processing operator Π_h on the discrete space of piecewise polynomials of degree at most r, and the approximation property of the numerical solution $\mathbf{u}_h$ of the Navier-Stokes equations.

Finally, let us consider in more detail the construction of the post-processing operator Π_h in the case $r = d = 2$. Suppose that the Navier-Stokes
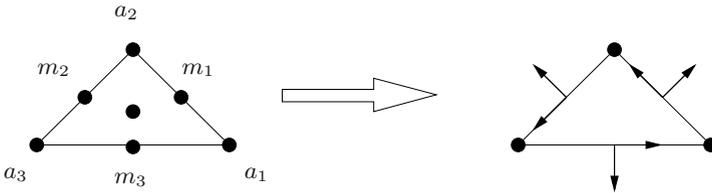
Fig. 6.5. Post-processing operator $\Pi_h : \left(P_2^{\text{bubble}}\right)^2 \to \left(P_2^{\text{disc}}\right)^2$

equations have been solved using the inf-sup stable pair $\left((P_2^{\text{bubble}})^2, P_1^{\text{disc}}\right)$. Consider a cell $K \in \mathcal{T}_h$ with vertices a_i, $i = 1, 2, 3$, midpoints m_i of the edges $E_i = \overline{a_i a_{i+1}}$ for $i = 1, 2, 3$ (where $a_4 = a_1$), and barycentric coordinates λ_i associated with a_i for $i = 1, 2, 3$; see Figure 6.5. Set $\lambda_4 = \lambda_1$ for later use. Then the velocity on K can be written in the form

$$\mathbf{u}_h = \sum_{i=1}^{7} N_i(\mathbf{u}_h)\varphi_i$$

where the local scalar basis functions are

$$\varphi_i = \lambda_i(2\lambda_i - 1), \qquad\qquad i = 1, 2, 3,$$
$$\varphi_i = 4\lambda_{i-3}\lambda_{i-2} - 20\lambda_1\lambda_2\lambda_3, \qquad i = 4, 5, 6,$$
$$\varphi_7 = 60\lambda_1\lambda_2\lambda_3,$$

and the nodal functionals are

$$N_i(v) = v(a_i), \qquad\qquad i = 1, \ldots, 3,$$
$$N_i(v) = v(m_{i-3}), \qquad\qquad i = 4, \ldots, 6,$$
$$N_7(v) = |K|^{-1} \int_K v(x)\, dx.$$

For vector-valued functions we use the convention that the nodal functionals are applied component by component, i.e., applying a nodal functional to a vector-valued function will result in a vector. The above representation of $\mathbf{u}_h$ is valid since $N_i(\varphi_j) = \delta_{ij}$, $i, j = 1, \ldots, 7$. Now the post-processing operator Π_K leaves unchanged every function in $\left(P_2(K)\right)^2$, and consequently one gets

$$\Pi_K \mathbf{u}_h = \sum_{i=1}^{6} N_i(\mathbf{u}_h)\psi_i + \Pi_K\left(\widetilde{N}_7(\mathbf{u}_h)\varphi_7\right)$$

with

$$\widetilde{N}_7(\mathbf{w}) = \left(N_7(\mathbf{w}) - \frac{N_4(\mathbf{w}) + N_5(\mathbf{w}) + N_6(\mathbf{w})}{3}\right).$$

Here $\{\psi_i, i = 1, \ldots, 6\}$ is the standard nodal basis of the scalar space $P_2(K)$ which is defined by

$$\begin{aligned}\psi_i &= \lambda_i(2\lambda_i - 1), & i &= 1, 2, 3, \\ \psi_i &= 4\lambda_{i-3}\lambda_{i-2}, & i &= 4, 5, 6.\end{aligned}$$

Let $\mathbf{n}_i$, $i = 1, 2, 3$, denote the outward-pointing unit normal on the edge E_i. Now $\varphi_7 \equiv 0$ on E_i, $i = 1, 2, 3$, so from (6.13a) one sees that for any vector $\mathbf{a} \in \mathbb{R}^2$ one has

$$\left\langle \Pi_K(\mathbf{a}\,\varphi_7) \cdot \mathbf{n}_i, \varphi\right\rangle_{E_i} = 0 \quad \forall \varphi \in P_2(E_i), \, i = 1, 2, 3. \tag{6.14}$$

From the formula for $\Pi_K \mathbf{u}_h$ above, we have only to evaluate $\Pi_K(\mathbf{a}\varphi_7)$ with $\mathbf{a} = \widetilde{N}_7(\mathbf{u}_h)$. Since $\Pi_K(\mathbf{a}\,\varphi_7) \cdot \mathbf{n}_i|_{E_i} \in P_2(E_i)$, $i = 1, 2, 3$, these functions may be used as test functions φ in (6.14); it follows that

$$\Pi_K(\mathbf{a}\,\varphi_7) \cdot \mathbf{n}_i|_{E_i} \equiv 0, \quad i = 1, 2, 3.$$

At each vertex $\mathbf{a}_i$, $i = 1, 2, 3$, two of the three equations yield

$$\Pi_K(\mathbf{a}\,\varphi_7)(\mathbf{a}_i) = 0, \quad i = 1, 2, 3.$$

Moreover, at the midpoint $\mathbf{m}_i$ of edge E_i one has

$$\Pi_K(\mathbf{a}\,\varphi_7)(\mathbf{m}_i) \cdot \mathbf{n}_i = 0, \quad i = 1, 2, 3.$$

Consequently $\Pi_K(\mathbf{a}\,\varphi_7)$ has only tangential components along the edges and can be written as

$$\Pi_K(\mathbf{a}\,\varphi_7) = \sum_{i=1}^{3} w_i \psi_{i+3}\tau_i$$

where τ_i is the unit tangent vector along the edge E_i and the w_i, $i = 1, 2, 3$, are given by $w_i = \Pi_K(\mathbf{a}\,\varphi_7)(\mathbf{m}_i) \cdot \tau_i$. This implies that only the tangential component will change when $\mathbf{u}_h$ is replaced by the divergence-free function

$\mathbf{w}_h = \Pi_h \mathbf{u}_h$. Since $\nabla P_1(K) = \mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ with $\mathbf{e}_1 = (1,0)^T$ and $\mathbf{e}_2 = (0,1)^T$, we obtain from (6.13b) the two relations

$$\frac{1}{3} \sum_{i=1}^{3} w_i (\boldsymbol{\tau}_i \cdot \mathbf{e}_j) = \mathbf{a} \cdot \mathbf{e}_j, \quad j = 1, 2.$$

The remaining equation needed to determine the 3 unknowns w_i, $i = 1, 2, 3$, follows from (6.13c). That is, for each cell one must solve one linear $3{\times}3$ system to compute the post-processed solution, which will be globally continuous at all vertices with continuous normal fluxes at the midpoints of the edges.

Compared with the method of Section 6.4.1, we solve both the transport and the Navier-Stokes equations with a method of order r. In this sense, the method of postprocessing the discrete velocity is well balanced. Moreover, mass conservation on the discrete level is guaranteed.

6.4.3 Scott-Vogelius Elements

Now we discuss finite element discretizations of the Navier-Stokes equations that guarantee that the discrete velocity solution $\mathbf{u}_h$ is piecewise divergence-free without any post-processing. To this end, let us consider discretizations with $(P_r^d, P_{r-1}^{\mathrm{disc}})$ elements, i.e., continuous piecewise polynomials of degree at most r for the velocity approximation and discontinuous piecewise polynomials of degree at most $r - 1$ for the pressure. For the two-dimensional case with $r \geq 4$, this finite element pair is inf-sup stable when special meshes that exhibit so-called singular vertices are excluded [SV85]. In $\mathbb{R}^3$ the characterization of those meshes that have to be excluded to guarantee inf-sup stability is more delicate, but it is a much simpler task to find a family of meshes on which this element is stable. Recently, in [Zha05], inf-sup stability has been shown on a certain type of macro-element mesh provided that the polynomial degree r is at least as large as the space dimension d. We start in the two-dimensional
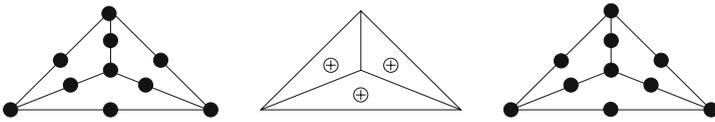
Fig. 6.6. Scott-Vogelius element for $r = d = 2$: continuous piecewise quadratic approximations of velocity and concentration together with discontinuous piecewise linear approximation of the pressure

case with a shape-regular decomposition of the domain into macro-triangles and perform one refinement step by joining the barycentre of each macro-element to its vertices. On the resulting mesh the pair $(P_r^2, P_{r-1}^{\mathrm{disc}})$ is inf-sup stable provided that $r \geq 2$. Similarly, in three dimensions one starts with a

shape-regular decomposition of the domain into macro-tetrahedra. Each of these is divided into 4 tetrahedra by performing one refinement step; each "child" tetrahedron has for its vertices three vertices and the barycentre of the "parent" tetrahedron. For $r \geq 3$ the pair $(P_r^3, P_{r-1}^{\mathrm{disc}})$ is inf-sup stable on such a family of meshes.

Although a restriction on the mesh is needed, this pair is attractive since a discretely divergence-free function is divergence-free. For, owing to the discrete spaces used, the divergence of each discrete velocity field belongs to the pressure space. Hence the discrete mass balance equation (6.11c) yields

$$0 = (\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{u}_h),$$

i.e., the discrete velocity solution $\mathbf{u}_h$ is divergence-free in the L^2 sense. Of course this yields $m_h(c_h, \mathbf{u}_h) = 0$ in (6.7) since the continuous velocity approximation u_h has no jumps.

Moreover, when solving the Navier-Stokes equations for higher Reynolds numbers one can exploit the macro-triangulation by using it as the coarser grid in the two-level local projection stabilization of Chapter 3. Thus let $\mathcal{T}_{2h}$ denote the macro-triangulation and $\mathcal{T}_h$ the triangulation after the refinement step described above. Then the approximation spaces for velocity, pressure, and concentration live on the finer triangulation $\mathcal{T}_h$ and will be $(P_r^d, P_{r-1}^{\mathrm{disc}}, P_r)$, while the projection space P_{r-1}^{disc} is defined on the coarser triangulation $\mathcal{T}_{2h}$.

7

Adaptive Error Control

The derivation of reliable and efficient *a posteriori error estimates* for the Navier-Stokes equations is an important consideration in computational fluid dynamics. On perusing existing *a priori* error estimates for the Navier-Stokes equations, one notices fundamental differences from and fresh difficulties compared with the estimates for diffusion-dominated and convection-dominated elliptic problems in Chapter II.3.6. Some new obstacles that appear are

- a smallness condition on the Reynolds number $\mathrm{Re} = 1/\nu$ to ensure uniqueness of the solution (Chapter 1)
- a well-posedness assumption for a linearized problem with an *a priori* error estimate that depends strongly on an unknown stability constant (Theorem 3.14).

As we saw in Example 3.16, this stability constant can grow exponentially in the Reynolds number Re. One should be aware that such a property will restrict considerably the quantitative value of our estimates. On the other hand, in the special case of a no-flow problem (see Remark 2.10), one has uniqueness of the solution for all Reynolds numbers and error estimates with a right-hand side that is a polynomial function of Re.

We now sketch the steps followed in deriving quantitative error estimates for the stationary Navier-Stokes equation in the following weak formulation, which is easily seen to be equivalent to (1.2):

Find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\nu(\nabla \mathbf{u}, \nabla \mathbf{v}) + n(\mathbf{u}, \mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}) = (\mathbf{f}, \mathbf{v}). \qquad (7.1)$$

Here, as in (1.2), $\mathbf{V} := H_0^1(\Omega)^d$ and $Q := L_0^2(\Omega)$.

In Chapter III.3.6 four different types of error estimators were discussed. Here we shall concentrate on two of them: residual estimators and goal-oriented estimators, which are often used in practical computations. In particular we consider the adaptive control of the global L^2 norm of the velocity field as an example for a residual estimator, and a goal-oriented estimator

to control the drag or the lift coefficients in flows around an obstacle will be examined. The main steps in such a technique are:

- error representation via a linearized continuous dual problem
- use of the projection property (Galerkin orthogonality)
- interpolation error estimates for the dual solution
- strong stability for the continuous dual problem.

In order not to overload the presentation with technical details, we shall first study a standard finite element discretization with finite element spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$ that satisfy the Babuška-Brezzi stability condition. Later we give some remarks concerning the use of stabilized discretization. Our discrete problem is:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$ one has

$$\nu(\nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + n(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot \mathbf{u}_h) = (\mathbf{f}, \mathbf{v}_h). \quad (7.2)$$

In what follows, let $(\mathbf{u}, p) \in \mathbf{V} \times Q$ and $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ be the solutions of the continuous problem (7.1) and the discrete problem (7.2). We introduce the following linearized dual problem:

Find $(\mathbf{w}, r) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\nu(\nabla \mathbf{v}, \nabla \mathbf{w}) - n(\mathbf{u}, \mathbf{w}, \mathbf{v}) + n(\mathbf{v}, \mathbf{u}_h, \mathbf{w})$$
$$+ (r, \nabla \cdot \mathbf{v}) - (q, \nabla \cdot \mathbf{w}) = (\mathbf{g}, \mathbf{v}). \quad (7.3)$$

Let $(\boldsymbol{\varphi}, \chi) \in \mathbf{V} \times Q$ denote the solution of (7.3) for the particular right-hand side $\mathbf{g} = \mathbf{u} - \mathbf{u}_h$. Then, setting $\mathbf{v} = \mathbf{u} - \mathbf{u}_h \in \mathbf{V}$ and $q = p - p_h$ in (7.3) and integrating by parts, one obtains the error representation

$$\|\mathbf{u} - \mathbf{u}_h\|_0^2 = \nu(\nabla(\mathbf{u} - \mathbf{u}_h), \nabla \boldsymbol{\varphi}) - n(\mathbf{u}, \boldsymbol{\varphi}, \mathbf{u} - \mathbf{u}_h) + n(\mathbf{u} - \mathbf{u}_h, \mathbf{u}_h, \boldsymbol{\varphi})$$
$$+ (\chi, \nabla \cdot (\mathbf{u} - \mathbf{u}_h)) - (p - p_h, \nabla \cdot \boldsymbol{\varphi}),$$
$$= \nu(\nabla(\mathbf{u} - \mathbf{u}_h), \nabla \boldsymbol{\varphi}) + n(\mathbf{u}, \mathbf{u}, \boldsymbol{\varphi}) - n(\mathbf{u}_h, \mathbf{u}_h, \boldsymbol{\varphi})$$
$$- (p - p_h, \nabla \cdot \boldsymbol{\varphi}) + (\chi, \nabla \cdot (\mathbf{u} - \mathbf{u}_h)). \quad (7.4)$$

As (7.1) holds for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$, one can subtract (7.2) from (7.1); this yields the projection property

$$\nu(\nabla(\mathbf{u} - \mathbf{u}_h), \nabla \mathbf{v}_h) + n(\mathbf{u}, \mathbf{u}, \mathbf{v}_h) - n(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h)$$
$$- (p - p_h, \nabla \cdot \mathbf{v}_h) + (q_h, \nabla \cdot (\mathbf{u} - \mathbf{u}_h)) = 0 \quad (7.5)$$

for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$.

Now choose $\mathbf{v}_h := i_h \boldsymbol{\varphi}$ and $q_h := j_h \chi$ to be interpolants (Scott-Zhang or Clément) to the solution $(\boldsymbol{\varphi}, \chi)$ of the dual problem (7.3) that satisfy

$$\|\boldsymbol{\varphi} - i_h \boldsymbol{\varphi}\|_{0,T} \leq C h_T^2 |\boldsymbol{\varphi}|_{2,\omega(T)} \qquad \text{for all } \boldsymbol{\varphi} \in H^2(\omega(T))^d,$$
$$\|\boldsymbol{\varphi} - i_h \boldsymbol{\varphi}\|_{0,E} \leq C h_T^{3/2} |\boldsymbol{\varphi}|_{2,\omega(T)} \qquad \text{for all } E \subset \partial T, \ \boldsymbol{\varphi} \in H^2(\omega(T))^d,$$
$$\|\chi - j_h \chi\|_{0,T} \leq C h_T^1 |\chi|_{1,\omega(T)} \qquad \text{for all } \chi \in H^1(\omega(T))^d.$$

Using (7.5) one can rewrite (7.4) as

$$
\begin{aligned}
\|\mathbf{u} - \mathbf{u}_h\|_0^2 = {} & \nu(\nabla(\mathbf{u} - \mathbf{u}_h), \nabla(\boldsymbol{\varphi} - i_h\boldsymbol{\varphi})) \\
& + n(\mathbf{u}, \mathbf{u}, \boldsymbol{\varphi} - i_h\boldsymbol{\varphi}) - n(\mathbf{u}_h, \mathbf{u}_h, \boldsymbol{\varphi} - i_h\boldsymbol{\varphi}) \\
& - (p - p_h, \nabla \cdot (\boldsymbol{\varphi} - i_h\boldsymbol{\varphi})) + (\chi - j_h\chi, \nabla \cdot (\mathbf{u} - \mathbf{u}_h)).
\end{aligned}
$$

Integrating by parts on each element, this becomes

$$
\begin{aligned}
\|\mathbf{u} - \mathbf{u}_h\|_0^2 = {} & \sum_{T \in \mathcal{T}_h} (\mathbf{f} + \nu\Delta\mathbf{u}_h - (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \nabla p_h, \boldsymbol{\varphi} - i_h\boldsymbol{\varphi})_T \\
& + \sum_{E \in \mathcal{E}_h} \left(\left[p_h\mathbf{n}_E - \nu\frac{\partial\mathbf{u}_h}{\partial\mathbf{n}_E} \right]_E , \boldsymbol{\varphi} - i_h\boldsymbol{\varphi} \right)_E \\
& - \sum_{T \in \mathcal{T}_h} (\nabla \cdot \mathbf{u}_h, \chi - j_h\chi)_T.
\end{aligned}
$$

Now invoke the approximation properties of the interpolants to get

$$
\begin{aligned}
\|\mathbf{u} - \mathbf{u}_h\|_0^2 \leq {} & C_1 \sum_{T \in \mathcal{T}_h} h_T^2 \|\mathbf{f} + \nu\Delta\mathbf{u}_h - (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \nabla p_h\|_{0,T} \, |\boldsymbol{\varphi}|_{2,\omega(T)} \\
& + C_2 \sum_{E \in \mathcal{E}_h} h_E^{3/2} \left\| \left[p_h\mathbf{n}_E - \nu\frac{\partial\mathbf{u}_h}{\partial\mathbf{n}_E} \right]_E \right\|_{0,E} \, |\boldsymbol{\varphi}|_{2,\omega(T)} \\
& + C_3 \sum_{T \in \mathcal{T}_h} h_T \|\nabla \cdot \mathbf{u}_h\|_{0,T} \, |\chi|_{1,\omega(T)}.
\end{aligned}
$$

Assume that the linearized dual problem (7.3) satisfies the strong stability estimate

$$
\nu|\mathbf{w}|_2 + |r|_1 \leq S(\mathbf{u}, \mathbf{u}_h) \, \|\mathbf{g}\|_0 \quad \text{for all } \mathbf{g} \in \mathbf{L}^2(\Omega).
$$

Then, recalling that $(\boldsymbol{\varphi}, \chi)$ is the solution of (7.3) for $\mathbf{g} = \mathbf{u} - \mathbf{u}_h$, we see that

$$
\|\mathbf{u} - \mathbf{u}_h\|_0 \leq S(\mathbf{u}, \mathbf{u}_h) \, C \, (\eta_1 + \eta_2 + \eta_3), \tag{7.6a}
$$

where

$$
C := \max\{C_1, 3C_2, C_3\},
$$

$$
\eta_1^2 := \sum_{T \in \mathcal{T}_h} \frac{h_T^4}{\nu^2} \|\mathbf{f} + \nu\Delta\mathbf{u}_h - (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \nabla p_h\|_{0,T}^2, \tag{7.6b}
$$

$$
\eta_2^2 := \sum_{E \in \mathcal{E}_h} \frac{h_E^3}{\nu^2} \left\| \left[p_h\mathbf{n}_E - \nu\frac{\partial\mathbf{u}_h}{\partial\mathbf{n}_E} \right]_E \right\|_{0,E}^2, \tag{7.6c}
$$

$$
\eta_3^2 := \sum_{T \in \mathcal{T}_h} h_T^2 \|\nabla \cdot \mathbf{u}_h\|_{0,T}^2. \tag{7.6d}
$$

If in (7.6a) we can bound (either theoretically or numerically) the stability factor $S(\mathbf{u}, \mathbf{u}_h)$ by a quantity of moderate size, then $\eta_1 + \eta_2 + \eta_3$ is a suitable error indicator. Thus the crucial question is the size of $S(\mathbf{u}, \mathbf{u}_h)$. In [JR94] it is proved that $S(\mathbf{u}, \mathbf{u}_h) = \mathcal{O}(\mathrm{Re}) = \mathcal{O}(1/\nu)$ for a model problem of nearly parallel streamwise constant pipe flow; this estimate appears to be sharp in terms of its dependence on Re. But in the general case, $S(\mathbf{u}, \mathbf{u}_h)$ has been estimated only from computational results [Joh95, HJ04].

Remark 7.1. Verfürth [Ver89] derives a similar *a posteriori error estimate* in the energy norm $|(\mathbf{v}, q)| := (|\mathbf{v}|_1^2 + \|q\|_0^2)^{1/2}$ using a Mini-element discretization (piecewise linear functions enriched by bubbles for the velocity approximation, and piecewise linear functions yielding a continuous pressure approximation) of the Stokes problem. This estimator can be extended to the case of the Navier-Stokes equations (at least for small Reynolds numbers) and to other discretizations. The main differences between estimators for the energy and L^2 norms are the different scalings of the local residuals in (7.6b) and (7.6d) and of the jumps in (7.6c) that control discontinuities in the pressure and in the normal derivative of the velocity. ♣

Now we turn to the DWR (dual weighted residual) method that was discussed in Section III.3.6.1; this technique is important when controlling quantities like drag and lift coefficients in flows around obstacles. To be more precise, consider the configuration of Figure 7.1. Assume the no-slip condition
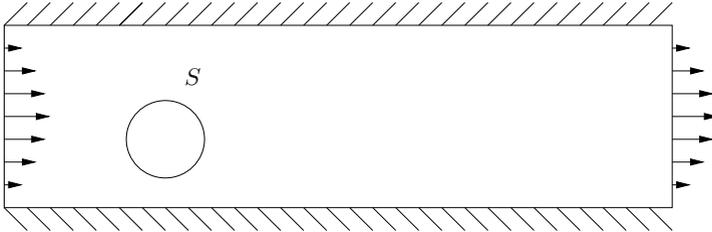
Fig. 7.1. Channel flow around an obstacle

$\mathbf{u} = \mathbf{0}$ at the walls and parabolic in-flow and out-flow velocity profiles. The drag and lift coefficients are defined by the functional

$$J(\mathbf{u}, p) := \frac{2}{\bar{U}^2 D} \int_S \mathbf{n}^t \big(\nu (\nabla \mathbf{u} + \nabla \mathbf{u}^t) - p\, \mathbb{I} \big) \, \mathbf{e} \, d\gamma,$$

where $\mathbf{e}$ denotes the (column) vector in the downwind and crosswind directions of flow, respectively, and S is the surface of the obstacle, D its diameter, $\bar{U}$ the reference velocity, and $\mathbb{I}$ the identity tensor.

The derivation of a goal-oriented estimator resembles the approach for estimating the global L^2 norm of the error. One replaces the linearized dual problem (7.3) by the following problem:

Find $(\mathbf{w}, r) \in \mathbf{V} \times Q$ such that for all $(\mathbf{v}, q) \in \mathbf{V} \times Q$ one has

$$\nu(\nabla \mathbf{v}, \nabla \mathbf{w}) - n(\mathbf{u}, \mathbf{w}, \mathbf{v}) + n(\mathbf{v}, \mathbf{u}_h, \mathbf{w})$$
$$+ (r, \nabla \cdot \mathbf{v}) - (q, \nabla \cdot \mathbf{w}) = J(\mathbf{v}, q). \qquad (7.7)$$

Let $(\boldsymbol{\varphi}, \chi) \in \mathbf{V} \times Q$ be the solution of (7.7) and let $(\boldsymbol{\varphi}_h, \chi_h) \in \mathbf{V}_h \times Q_h$ be arbitrary. Then one can deduce the error representation

$$J(\mathbf{u}, p) - J(\mathbf{u}_h, p_h) = \sum_{T \in \mathcal{T}_h} (\mathbf{f} + \nu \Delta \mathbf{u}_h - (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \nabla p_h, \boldsymbol{\varphi} - \boldsymbol{\varphi}_h)_T$$

$$+ \sum_{E \in \mathcal{E}_h} \left(\left[p_h \mathbf{n}_E - \nu \frac{\partial \mathbf{u}_h}{\partial \mathbf{n}_E} \right]_E, \boldsymbol{\varphi} - \boldsymbol{\varphi}_h \right)_E$$

$$- \sum_{T \in \mathcal{T}_h} (\nabla \cdot \mathbf{u}_h, \chi - \chi_h)_T.$$

A residual-type error estimator follows by setting $\mathbf{v}_h = i_h \mathbf{v}_h$ and $\chi_h = j_h \chi$, then invoking the approximation properties of the interpolants and the strong stability of the linearized dual problem. On the other hand, the idea of the dual weighted residual approach is to use instead the estimate

$$|J(\mathbf{u}, p) - J(\mathbf{u}_h, p_h)| \le \sum_{T \in \mathcal{T}_h} \|r_{1,T}\|_{0,T} \|\boldsymbol{\varphi} - \boldsymbol{\varphi}_h\|_{0,T}$$

$$+ \sum_{E \in \mathcal{E}_h} \|r_E\|_{0,E} \|\boldsymbol{\varphi} - \boldsymbol{\varphi}_h\|_{0,E} + \sum_{T \in \mathcal{T}_h} \|r_{2,T}\|_{0,T} \|\chi - \chi_h\|_{0,T}$$

and to approximate the weights $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_h\|_{0,T}$, $\|\boldsymbol{\varphi} - \boldsymbol{\varphi}_h\|_{0,E}$, and $\|\chi - \chi_h\|_{0,T}$ by solving the linearized dual problem. In this formula, for the element and edge residuals we used the notation

$$r_{1,T} := \mathbf{f} + \nu \Delta \mathbf{u}_h - (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \nabla p_h,$$

$$r_E := \left[p_h \mathbf{n}_E - \nu \frac{\partial \mathbf{u}_h}{\partial \mathbf{n}_E} \right]_E,$$

$$r_{2,T} := \nabla \cdot \mathbf{u}_h.$$

Different ways of approximating the weights are discussed in [Bec00, BR01, BR03].

So far, we have considered only the standard Galerkin discretization of the Navier-Stokes equation without any stabilization. If stabilization is needed there are two possibilities: one can add a stabilization term to the linearized dual problem or one can start from the stabilized formulation and construct an associated dual problem. In general, the formation of the dual problem and stabilization do not commute [BR03]. But symmetric stabilization terms such as in local projection stabilization (LPS) or the continuous interior penalty approach (CIP) change this situation; see [BV07]. Adaptive error control for the discretization of flow problems remains an attractive field for further research.

References

[AAS07]   M. I. Asensio, B. Ayuso, and G. Sangalli. Coupling stabilized finite element methods with finite difference time integration for advection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 196:3475–3491, 2007.

[AB99]   M. Ainsworth and I. Babuska. Reliable and robust a posteriori error estimation for singularly perturbed reaction-diffusion problems. *SIAM J. Numer. Anal.*, 36:331–353, 1999.

[ABCM02] D. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39:1749–1779, 2002.

[ABF84]   D. N. Arnold, F. Brezzi, and M. Fortin. A stable finite element for the Stokes equation. *Calcolo*, 21:337–344, 1984.

[ABF02]   D.N. Arnold, D. Boffi, and R.S. Falk. Approximation by quadrilateral finite elements. *Math. Comp.*, 71(239):909–922, 2002.

[ABR07]   R. Araya, E. Behrens, and R. Rodriguez. Error estimators for advection-diffusion-reaction problems based on the solution of local problems. *J. Comput. Appl. Math.*, 206:440–453, 2007.

[AC85]   O. Axelsson and G. F. Carey. On the numerical solution of two-point singularly perturbed boundary value problems. *Comput. Methods Appl. Mech. Engrg.*, 50(3):217–229, 1985.

[AC92]   M. Ainsworth and A. Craig. A posteriori error estimators in the finite element method. *Numer. Math.*, 60:429–463, 1992.

[AD92]   T. Apel and M. Dobrowolski. Anisotropic interpolation with applications to the finite element method. *Computing*, 47:277–293, 1992.

[AD01]   M. Ainsworth and W. Doerfler. Fundamental sytems of numerical schemes for linear convection-diffusion equations and their relationship to accuracy. *Computing*, 66:199–229, 2001.

[Ada78]   R. A. Adams. *Sobolev spaces*. Academic press, San Diego, 1978.

[ADN59]   S. Agmon, A. Douglis, and N. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. *Comm. Pure Appl. Math.*, 12:623–727, 1959.

[AEB07]   L. E. Alaoui, A. Ern, and E. Burman. A priori and a posteriori analysis of non-conforming finite elements with face penalty for advection-diffusion equations. *IMA J. Numer. Anal.*, 27:151–171, 2007.

[AF90]   D. C. Arney and J. E. Flaherty. An adaptive mesh-moving and local re-
            finement method for time-dependent partial differential equations. *ACM
            Trans. Math. Software*, 16:48–71, 1990.

[AFMW92]   S. Adjerid, J. E. Flaherty, P. K. Moore, and Y. J. Wang. High-order
            adaptive methods for parabolic systems. *Physica D*, 60:94–111, 1992.

[AH03]   A. R. Ansari and A. F. Hegarty. Numerical solution of a convection
            diffusion problem with Robin boundary conditions. *J. Comput. Appl.
            Math.*, 156(1):221–238, 2003.

[AK90]   O. Axelsson and L. Kolotilina. Monotonicity and discretization error
            estimates. *SIAM J. Numer. Anal.*, 27(6):1591–1611, 1990.

[AK96]   V. B. Andreev and N. V. Kopteva. Investigation of difference schemes
            with an approximation of the first derivative by a central difference
            relation. *Comput. Math. Math. Phys.*, 36:1065–1078, 1996.

[AK98]   V. B. Andreev and N. V. Kopteva. On the convergence, uniform with re-
            spect to a small parameter, of monotone three-point difference schemes.
            Differential Equations, 34:921–929, 1998.

[AK08]   V. B. Andreev and N. Kopteva. Pointwise approximation of corner
            singularities for a singularly perturbed reaction-diffusion equation in an
            L-shaped domain. *Math. Comp.*, 2008. Electronic publication Febr 19.

[AKK74]   L. R. Abrahamsson, H. B. Keller, and H. O. Kreiss. Difference approx-
            imations for singular perturbations of systems of ordinary differential
            equations. *Numer. Math.*, 22:367–391, 1974.

[AL90]   O. Axelsson and W. Layton. Defect correction methods for convection-
            dominated convection-diffusion equations. *RAIRO J. Numer. Anal.*,
            24:423–455, 1990.

[AL96]   T. Apel and G. Lube. Anisotropic mesh refinement in stabilized Galerkin
            methods. *Numer. Math.*, 74:261–282, 1996.

[Ale81]   M. W. Aleksejewskij. Higher order difference schemes for singularly per-
            turbed boundary layer problems (in Russian). *Diff. Equations*, 17:1171–
            1183, 1981.

[Alh07]   K. Alhumaizi. Flux-limiting solution techniques for simulation of
            reaction-diffusion-convection system. *Commun. Nonlinear Sci. Numer.
            Simul.*, 12:953–965, 2007.

[And]   V. B. Andreev. Uniform grid approximation for nonsmooth solutions of
            a mixed problem for a singularly perturbed reaction-diffusion problem
            in a square region. Preprint.

[And01]   V. B. Andreev. The Green function and a posteriori estimates of so-
            lutions of monotone three-point singularly perturbed finite difference
            schemes. *Differential Equations*, 37:923–933, 2001.

[And02]   V. B. Andreev. A priori estimates for solutions of singularly perturbed
            two-point boundary value problems. *Mat. Model.*, 14(5):5–16, 2002. Sec-
            ond International Conference OFEA'2001 "Optimization of Finite Ele-
            ment Approximation, Splines and Wavelets" (Russian) (St. Petersburg,
            2001).

[And06]   V. B. Andreev. On the accuracy of grid approximations to non-
            smooth solutions of a singularly perturbed reaction-diffusion equation
            in a square. *Differential Equations*, 42:895–906, 2006.

[Ang91a]   L. Angermann. A modified error estimator of Babuška-Rheinboldt's
            type for singularly perturbed elliptic problems. In H.-G. Roos, A. Fel-
            genhauer, and L. Angermann, editors, *Numerical methods in singularly*

perturbed problems. Proceedings of an International Workshop, pages 1–12. TU Dresden, 1991.

[Ang91b] Lutz Angermann. Numerical solution of second-order elliptic equations on plane domains. *RAIRO Modél. Math. Anal. Numér.*, 25:169–191, 1991.

[Ang92] L. Angermann. An a posteriori estimation of elliptic boundary value problems by means of upwind fem. *IMA J. Num. Anal.*, 12:201–215, 1992.

[Ang93] L. Angermann. Addendum to the paper: "Numerical solution of second-order elliptic equations on plane domains" [RAIRO Modél. Math. Anal. Numér. **25** (1991), no. 2, 169–191]. *RAIRO Modél. Math. Anal. Numér.*, 27(1):1–7, 1993.

[Ang95a] L. Angermann. Balanced a posteriori error estimates for finite volume type discretizations of convection-dominated elliptic problems. *Computing*, 55(4):305–323, 1995.

[Ang95b] L. Angermann. Error estimates for the finite-element solution of an elliptic singularly perturbed problem. *IMA J. Num. Anal.*, 15:161–196, 1995.

[Ang95c] L. Angermann. An upwind scheme of finite volume type with reduced crosswind diffusion. Bericht 165, Institut für Angewandte Mathematik, Universität Erlangen-Nürnberg, 1995.

[Ang00] L. Angermann. Error analysis of upwind-discretizations for the steady-state incompressible Navier-Stokes equations. *Advances in Computational Mathematics*, 13:167–198, 2000.

[AO82] L. Abrahamsson and S. Osher. Monotone difference schemes for singularly perturbed problems. *SIAM J. Numer. Anal.*, 19:979–992, 1982.

[AO00] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis.* John Wiley, New York, 2000.

[Ape99] T. Apel. *Anisotropic Finite Elements: Local Estimates and Applications.* Advances in Numerical Mathematics. B. G. Teubner, Stuttgart, 1999.

[APS05] R. Araya, A. H. Poza, and E. P. Stephan. A hierarchical a posteriori error estimate for an advection-diffusion-reaction problem. *Math. Models and Meth. in Appl. Sci.*, 15:1119–1139, 2005.

[ARS04] M. I. Asensio, A. Russo, and G. Sangalli. The residual-free bubble numerical method with quadratic elements. *Math. Models Methods Appl. Sci.*, 14(5):641–661, 2004.

[AS55] D. N. Allen and R. V. Southwell. Relaxation methods applied to determine motion, in two dimensions, of a viscous fluid past a fixed cylinder. *J. Mech. Appl. Math.*, 8:129–145, 1955.

[AS95] V. B. Andreev and I. A. Savin. On the convergence, uniform with respect to the small parameter, of A. A. Samarskiĭ's monotone scheme and its modifications. *Comput. Math. Math. Phys.*, 35:581–591, 1995.

[AS96] V. B. Andreev and I. A. Savin. On the computation of a boundary flux with uniform accuracy with respect to a small parameter. *Comput. Math. Math. Phys.*, 36:1687–1692, 1996.

[Azz80] A. Azzam. On differentiability properties of solutions of elliptic differential equations. *J. Math. Anal. Appl.*, 75:431–440, 1980.

[BA72] I. Babuška and A. K. Aziz. Survey lectures on the mathematical foundation of the finite element method. In A.K. Aziz, editor, *The mathematical*

foundation of the finite element method with applications to partial differential equations, pages 1–362. Academic Press, New York, 1972.

[Bai91]   M. J. Baines. An analysis of the moving finite element method. *SIAM J. Numer. Anal.*, 28:1323–1349, 1991.

[Bai94]   M. J. Baines. *Moving finite elements.* Oxford University Press, Oxford, 1994.

[Bak69]   A. S. Bakhvalov. On the optimization of methods for solving boundary value problems with boundary layers (in Russian). *Zh. Vychisl. Mat. i Mat. Fis.*, 9:841–859, 1969.

[Ban98]   R. E. Bank. *PLTMG: a software package for solving elliptic partial differential equations. Users' Guide 8.0.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[BB01]   R. Becker and M. Braack. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo*, 38(4):173–199, 2001.

[BB04]   R. Becker and M. Braack. A two-level stabilization scheme for the Navier-Stokes equations. In M. Feistauer et al., editor, *Numerical mathematics and advanced applications*, pages 123–130, Berlin, 2004. Springer-Verlag.

[BB06]   M. Braack and E. Burman. Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.*, 43:2544–2566, 2006.

[BBB73]   C. Bardos, D. Brezis, and H. Brezis. Perturbations singulieres et prolongements maximaux d'operateurs positives. *Archive Rat. Mech. Anal.*, 53:69–100, 1973.

[BBF93]   C. Baiocchi, F. Brezzi, and L.P. Franca. Virtual bubbles and Galerkin-least-squares type methods (Ga.L.S.). *Comput. Methods Appl. Mech. Engrg.*, 105:125–141, 1993.

[BBFS90]   R. E. Bank, J. F. Bürgler, W. Fichtner, and R. K. Smith. Some upwinding techniques for finite element approximations of convection-diffusion equations. *Numer. Math.*, 58:185–202, 1990.

[BBJL07]   M. Braack, E. Burman, V. John, and G. Lube. Stabilized finite element methods for the generalized Oseen problem. *Comput. Methods Appl. Mech. Engrg.*, 196(4–6):853–866, 2007.

[BC04]   S. Berrone and C. Canuto. Multilevel a posteriori analysis for reaction-convection-diffusion problems. *Appl. Numer. Math.*, 50:371–394, 2004.

[BCG87]   C. Bardos, V. Cea, and P. Grisvard. Error estimates related to singular perturbations of first-order equations and systems. *Comput. Math. Appl.*, 13:801–829, 1987.

[BCGJ07]   B. Bujanda, C. Clavero, J. L. Gracia, and J. C. Jorge. A high order uniformly convergent alternating direction scheme for time dependent reaction-diffusion singularly perturbed problems. *Numer. Math.*, 107:1–25, 2007.

[BCO94]   I. Babuška, G. Caloz, and J. E. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.*, 31:945–981, 1994.

[BDdV04]   P. Binev, W. Dahmen, and R. de Vore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97:219–268, 2004.

[BDM85]   F. Brezzi, J. Jr. Douglas, and L. D. Marini. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.*, 47:217–235, 1985.

[BE02]   E. Burman and A. Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation. *Comput. Methods Appl. Mech. Engrg.*, 191(35):3833–3855, 2002.

[BE05]   E. Burman and A. Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: Discrete maximum principle and convergence. *Math. Comp.*, 74(252):1637–1652, 2005.

[BE07]   E. Burman and A. Ern. Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations. *Math. Comp.*, 76:1119–1140, 2007.

[Bec00]   R. Becker. An optimal-control approach to a posteriori error estimation for finite element discretizations of the Navier–Stokes equations. *East-West J. Numer. Math.*, 9:257–274, 2000.

[BF91]   F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, 1991.

[BF07]   E. Burman and M. A. Fernández. Continuous interior penalty finite element method for the time-dependent Navier-Stokes equations: space discretization and convergence. *Numer. Math.*, 107:39–77, 2007.

[BFH06]   E. Burman, M. A. Fernández, and P. Hansbo. Continuous interior penalty finite element method for Oseen's equations. *SIAM J. Numer. Anal.*, 44:1248–1274, 2006.

[BGL07]   E. Burman, J. Guzmán, and D. Leykekhman. Weighted error estimates of the continuous interior penalty method for singularly perturbed problems. Technical report, EPFL Lausanne, 2007.

[BGS04]   P. B. Bochev, M. D. Gunzburger, and J. N. Shadid. Stability of the SUPG finite element method for transient advection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 193:2301–2323, 2004.

[BH82]   A. N. Brooks and T. J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32:199–259, 1982.

[BH04]   E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193:1437–1453, 2004.

[BH06]   E. Burman and P. Hansbo. A stabilized non-conforming finite element method for incompressible flow. *Comput. Methods Appl. Mech. Engrg.*, 195(23-24):2881–2899, 2006.

[BH07]   I. Boglaev and M. Hardy. Uniform convergence of a weighted average scheme for a nonlinear reaction-diffusion problem. *J. Comput. Appl. Math.*, 200:705–721, 2007.

[BHJJ06] M. J. Baines, M. E. Hubbard, P. K. Jimack, and A. C. Jones. Scale-invariant moving finite elements for nonlinear partial differential equations in two dimensions. *Appl. Numer. Math.*, 56:230–252, 2006.

[BHK84]   A. E. Berger, H. Han, and R. B. Kellogg. A priori estimates and analysis of a numerical method for a turning point problem. *Math. Comp.*, 42:465–492, 1984.

[BHM+99] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli. A priori error analysis of residual-free bubbles for advection-diffusion problems. *SIAM J. Numer. Anal.*, 36(6):1933–1948, 1999.

[BHMS03] F. Brezzi, G. Hauke, L. D. Marini, and G. Sangalli. Link-cutting bubbles for the stabilization of convection-diffusion-reaction problems. *Math. Models Meth. Appl. Sci.*, 13(3):445–461, 2003.

[BK01] J. Brandts and M. Křížek. History and future of superconvergence in three-dimensional finite element methods. In *Finite element methods (Jyväskylä, 2000)*, volume 15 of *GAKUTO Internat. Ser. Math. Sci. Appl.*, pages 22–33. Gakkōtosho, Tokyo, 2001.

[BK02] M. Bause and P. Knabner. Uniform error analysis for Lagrange-Galerkin approximations of convection-dominated problems. *SIAM J. Numer. Anal.*, 39:1954–1984, 2002.

[BKS98] C. J. Budd, G. P. Koomullil, and A. M. Stuart. On the solution of convection-diffusion boundary value problems using equidistributed grids. *SIAM J. Sci. Comput.*, 20:591–618, 1998.

[BL08] M. Braack and G. Lube. Finite elements with local projection stabilization for incompressible flow problems. Technical report, University Göttingen, 2008.

[BLR86] H. Blum, Q. Lin, and R. Rannacher. Asymptotic error epansion and Richardson extrapolation for linear finite elements. *Numer. Math.*, 49:11–37, 1986.

[BM80] J. W. Barrett and K. W. Morton. Optimal finite element solutions to diffusion convection problems in one dimension. *Int. J. Numer. Meth. Engng.*, 15:1457–1474, 1980.

[BM00] G. Beckett and J. A. Mackenzie. Convergence analysis of finite difference approximations on equidistributed grids to a singularly perturbed boundary value problem. *Appl. Numer. Math.*, 35(2):87–109, 2000.

[BMR98] F. Brezzi, L. D. Marini, and A. Russo. Applications of pseudo–residual–free bubbles to the stabilization of convection–diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 166:51–63, 1998.

[BMR05] F. Brezzi, L. D. Marini, and A. Russo. On the choice of stabilizing subgrid for convection–diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 194:127–148, 2005.

[BMRS01] G. Beckett, J.A. Mackenzie, A. Ramage, and D. M. Sloan. On the numerical solution of one-dimensional PDEs using adaptive methods based on equidistribution. *J. Comput. Phys.*, 167:372–392, 2001.

[BMRS02] G. Beckett, J. A. Mackenzie, A. Ramage, and D. M. Sloan. Computational solution of two-dimensional unsteady PDEs using moving mesh methods. *J. Comput. Phys.*, 182:478–495, 2002.

[BMS00] F. Brezzi, D. Marini, and E. Süli. Residual-free bubbles for advection-diffusion problems: the general error analysis. *Numer. Math.*, 85(1):31–47, 2000.

[BMS04] F. Brezzi, D. Marini, and A. Süli. Discontinuous Galerkin methods for first order hyperbolic problems. *Math. Models a. Meth. in Appl. Sci.*, 14:1893–1903, 2004.

[BNV06a] A. Bermúdez, M. Nogueiras, and C. Vázquez. Numerical analysis of convection-diffusion-reaction problems with higher order characteristics/finite elements. I. Time discretization. *SIAM J. Numer. Anal.*, 44:1829–1853, 2006.

[BNV06b] A. Bermúdez, M. Nogueiras, and C. Vázquez. Numerical analysis of convection-diffusion-reaction problems with higher order characteristics/finite elements. II. Fully discretized scheme and quadrature formulas. *SIAM J. Numer. Anal.*, 44:1854–1876, 2006.

[BO99] C. Baumann and J. Oden. A discontinuous hp-finite element method for convection-diffusion problems. *Comput. Meth. Appl. Mech. Engrg.*, 175:311–341, 1999.

[Bob67] L. Bobisud. Second-order linear parabolic equations with a small parameter. *Arch. Rational Mech. Anal.*, 27:385–397, 1967.

[Bog84] I. P. Boglaev. An approximate solution of a nonlinear boundary value problem with a small parameter multiplying the highest derivative. *Zh. Vychisl. Mat. i Mat. Fiz.*, 24:1649–1656, 1758, 1984.

[Bog01] I. Boglaev. Domain decomposition for a singularly perturbed parabolic problem with a convection-dominated term. *J. Comput. Appl. Math.*, 134:283–299, 2001.

[Bog05] I. Boglaev. Monotone Schwarz iterates for a semilinear parabolic convection-diffusion problem. *J. Comput. Appl. Math.*, 183:191–209, 2005.

[Bog06] I. Boglaev. Domain decomposition for a parabolic convection-diffusion problem. *Numer. Methods Partial Diff. Eqs.*, 22:1361–1378, 2006.

[Boh81] E. Bohl. *Finite Modelle gewöhnlicher Randwertaufgaben*. Teubner, Stuttgart, 1981.

[BPS01] J. H. Bramble, J. E. Pasciak, and O. Steinbach. On the stability of the L^2 projection in $H^1(\Omega)$. *Math. Comp.*, 71(237):147–156, 2001.

[BR78] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computation. *SIAM J. Numer. Anal.*, 15:736–754, 1978.

[BR84] K. Böhmer and R. Rannacher. *Defect Correction Methods: Theory and Applications*. Springer-Verlag, Berlin, 1984.

[BR94] F. Brezzi and A. Russo. Choosing bubbles for advection-diffusion problems. *Mathematical Models and Methods in Applied Sciences*, 4:571–587, 1994.

[BR01] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.

[BR03] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.

[BR06a] M. Braack and T. Richter. Solutions of 3D Navier-Stokes benchmark problems with adaptive finite elements. *Computers & Fluids*, 35(4):372–392, 2006.

[BR06b] M. Braack and T. Richter. Stabilized finite elements for 3D reactive flows. *Int. J. Numer. Methods Fluids*, 51(9-10):981–999, 2006.

[BR07] M. Braack and T. Richter. Solving multidimensional reactive flow problems with adaptive finite elements. In R. Rannacher W. Jäger and J. Warnatz, editors, *Reactive Flows, Diffusion and Transport*, pages 93–112. Springer-Verlag, 2007.

[Bra] M. Braack. A stabilized finite element scheme for the Navier-Stokes equations on anisotropic meshes. *M2AN Math. Model. Numer. Anal.* (to appear).

[BRR80]   F. Brezzi, J. Rappaz, and P. Raviart. Finite dimensional approximations of non-linear problems I. branches of non-singular solutions. *Numer. Math.*, 38:1–25, 1980.

[BS89]   I. Boglaev and W. Sirotkin. A numerical algorithm for solving singularly perturbed problems, arising in the modelling of semiconductor structures. Report Russian Academy of Science, Tsernogolovka, 1989.

[BS90]   I. Boglaev and W. Sirotkin. On the numerical solution on nonequidistant meshes of some quasilinear singularly perturbed problems. *Zh. Vychisl. Mat. i Mat. Fis.*, 30:680–696, 1990.

[BS97]   P. Balland and E. Süli. Analysis of the cell-vertex finite volume method for hyperbolic problems with variable coefficients. *SIAM J. Numer. Anal.*, 34(3):1127–1151, 1997.

[BSC$^+$80] A. E. Berger, J. M. Solomon, M. Ciment, S. H. Leventhal, and B. C. Weinberg. Generalized OCI schemes for boundary layer problems. *Math. Comp.*, 151:695–731, 1980.

[BSC81]   A. E. Berger, J. M. Solomon, and M. Ciment. An analysis of a uniformly accurate difference method for a singular perturbation problem. *Math. Comp.*, 37:79–94, 1981.

[BSU94]   I. Babuška, T. Strouboulis, and C. S. Upadhyay. A model study of the quality of a posteriori estimators for linear elliptic problems. *Comput. Methods Appl. Mech. Eng.*, 114:307–378, 1994.

[Bur05]   E. Burman. A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty. *SIAM J. Numer. Anal.*, 43(5):2012–2033, 2005.

[Bur07]   E. Burman. Interior penalty variational multiscale method for the incompressible Navier-Stokes equation: monitoring artificial dissipation. *Comput. Methods Appl. Mech. Engrg.*, 196(41-44):4045–4058, 2007.

[But75]   V. F. Butuzov. On asymptotics of solutions of singularly perturbed equations of elliptic type in the rectangle. *Diff. Equations*, 11:780–787, 1975.

[BV07]   R. Becker and B. Vexler. Optimal control of the convection-diffusion equation using stabilized finite element methods. *Numer. Math.*, 106:349–367, 2007.

[BW90]   R. E. Bank and B. D. Welfert. A posteriori error estimates for the Stokes equations. *Comput. Methods Appl. Mech. Eng.*, 87:323–340, 1990.

[BX03]   R. E. Bank and J. Xu. Asymptotically exact a posteriori error estimators. I. Grids with superconvergence. *SIAM J. Numer. Anal.*, 41:2294–2312, 2003.

[BY91]   A. Brandt and L. Yavneh. Inadequacy of some first-order upwind difference schemes for some recirculating flows. *J. Comput. Phys.*, 93:128–143, 1991.

[CA92]   M. C. Curran and M. B. Allen. Domain-decomposition approach to local grid refinement in finite element collocation. *Numer. Methods Partial Diff. Equat.*, 8:341–355, 1992.

[Cas02]   P. Castillo. Performance of discontinuous Galerkin methods for elliptic PDE's. *J. Sci. Comput.*, 24:524–547, 2002.

[CB02a]   C. Carstensen and S. Bartels. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids I. *Math. Comp.*, 71:945–969, 2002.

[CB02b]   C. Carstensen and S. Bartels. Each averaging technique yields reliable a posteriori error control in FEM on unstructered grids II: Higher order FEM. *Math. Comp.*, 71:971–994, 2002.

[CB02c]   R. Codina and J. Blasco. Analysis of a stabilized finite element approximation of the transient convection-diffusion-reaction equation using orthogonal subscales. *Comput. Vis. Sci.*, 4:167–174, 2002.

[CC07]   F. Celiker and B. Cockburn. Superconverence of the numerical traces of discontinuous Galerkin and hybridized methods for convection-diffusion problems in one space dimension. *Math. Comp.*, 76:67–96, 2007.

[CCSS02]   P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab. Optimal a priori error estimates for the *hp*-version of the local discontinuous Galerkin method for convection-diffusion problems. *Math. Comp.*, 71:455–478, 2002.

[CG04]   C. Clavero and J. L. Gracia. HODIE finite difference schemes on generalized Shishkin meshes. In *Proceedings of the 10th International Congress on Computational and Applied Mathematics (ICCAM-2002)*, volume 164/165, pages 195–206, 2004.

[CG05]   C. Clavero and J. L. Gracia. High order methods for elliptic and time dependent reaction-diffusion singularly perturbed problems. *Appl. Math. Comput.*, 168:1109–1127, 2005.

[CGJ05]   C. Clavero, J. L. Gracia, and J. C. Jorge. High-order numerical methods for one-dimensional parabolic singularly perturbed problems with regular layers. *Numer. Methods Partial Differential Equations*, 21:148–169, 2005.

[CGJ06a]   A. Cangiani, E.H. Georgoulis, and M. Jensen. Continuous and discontinuous finite element methods for convection-diffusion problems: a comparison. Report MA 06-011, University of Leicester, 2006.

[CGJ06b]   C. Clavero, J. L. Gracia, and J. C. Jorge. A uniformly convergent alternating direction HODIE finite difference scheme for 2D time-dependent convection-diffusion problems. *IMA J. Numer. Anal.*, 26:155–172, 2006.

[CGL99]   C. Clavero, J. L. Gracia, and F. Lisbona. High order methods on Shishkin meshes for singular perturbation problems of convection-diffusion type. *Numer. Algorithms*, 22(1):73–97, 1999.

[CGLS02]   C. Clavero, J. L. Gracia, F. Lisbona, and G. I. Shishkin. A robust method of improved order for convection-diffusion problems in a domain with characteristic boundaries. *ZAMM*, 82:631–647, 2002.

[CGMZ76]   I. Christie, D. F. Griffiths, A. R. Mitchell, and O. C. Zienkiewicz. Finite element methods for second order differential equations with significant first derivatives. *Int. J. Numer. Methods Eng.*, 10:1389–1396, 1976.

[CGO05]   C. Clavero, J. L. Gracia, and E. O'Riordan. A parameter robust numerical method for a two dimensional reaction-diffusion problem. *Math. Comp.*, 74:1743–1758, 2005.

[CH84]   K. W. Chang and F. A. Howes. *Nonlinear singular perturbation phenomena: theory and application.* Springer-Verlag, Berlin, 1984.

[Cia02]   P.G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam].

[CJL93]   C. Clavero, J. C. Jorge, and F. Lisbona. Uniformly convergent schemes for singular perturbation problems combining alternating directions and

exponential fitting techniques. In J.J.H. Miller, editor, *Applications of Advanced Computational Methods for Boundary and Interior Layers*, pages 33–52. Boole Press, Dublin, 1993.

[CJLS98]   C. Clavero, J. C. Jorge, F. Lisbona, and G. I. Shishkin. A fractional step method on a special mesh for the resolution of multidimensional evolutionary convection-diffusion problems. *Appl. Numer. Math.*, 27:211–231, 1998.

[CJLS00]   C. Clavero, J. C. Jorge, F. Lisbona, and G. I. Shishkin. An alternating direction scheme on a nonuniform mesh for reaction-diffusion parabolic problems. *IMA J. Numer. Anal.*, 20:263–280, 2000.

[CKS04]   B. Cockburn, G. Kanschat, and D. Schötzau. The local discontinuous Galerkin method for the Oseen equations. *Math. Comp.*, 73:569–593, 2004.

[CKS05a]   B. Cockburn, G. Kanschat, and D. Schötzau. The local discontinuous Galerkin method for linearized incompressible fluid flow: a review. *Comput. & Fluids*, 34:491–506, 2005.

[CKS05b]   B. Cockburn, G. Kanschat, and D. Schötzau. A locally conservative LDG method for the incompressible Navier-Stokes equations. *Math. Comp.*, 74:1067–1095, 2005.

[CKS07]   B. Cockburn, G. Kanschat, and D. Schötzau. A note on discontinuous Galerkin divergence-free solutions of the Navier-Stokes equations. *J. Sci. Comput.*, 31(1-2):61–73, 2007.

[CL93]   C. Clavero and F. Lisbona. Uniformly convergent finite difference methods for singular perturbation problems with turning points. *Numer. Algorithms*, 4:339–359, 1993.

[Clé75]   Ph. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numer.*, 9:77–84, 1975.

[CLGD06]   F. Courty, D. Leservoisier, P. L. George, and A. Deivieux. Continuous metrics and mesh optimization. *Appl. Numer. Math.*, 56:117–145, 2006.

[CLM95]   C. Clavero, F. Lisbona, and J. H. Miller. Uniform convergence of arbitrary order on nonuniform meshes for a singularly perturbed boundary value problem. *J. Comput. Appl. Math.*, 59:155–171, 1995.

[CLMM94]   Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick. First-order system least squares for second order partial differential equations. *SIAM J. Numer. Math.*, 31:1785–1799, 1994.

[CM93]   A. J. Chorin and J. E. Marsden. *A mathematical introduction to fluid mechanics.* Springer-Verlag, Berlin, 1993.

[CMOS01]   C. Clavero, J. H. Miller, E. O'Riordan, and G. I. Shishkin. Numerical experiments for advection-diffusion problems in a channel with a with a 180 bend. *Appl. Num. Math.*, 118:243–256, 2001.

[Coc03]   B. Cockburn. Discontinuous Galerkin methods. *ZAMM*, 83:731–754, 2003.

[Cod93]   R. Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Comput. Methods Appl. Mech. Engrg.*, 110(3-4):325–342, 1993.

[Cod98]   R. Codina. Comparison of some finite element methods for solving the diffusion-convection-reaction equation. *Comput. Methods Appl. Mech. Engrg.*, 156:185–210, 1998.

[CR73]   M. Crouzeix and P. A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO Numer. Anal.*, 3:33–76, 1973.

[CRHE90] M. A. Celia, T. F. Russell, I. Herrera, and R. Ewing. An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation. *Adv. Water Resour.*, 13:187–206, 1990.

[CS99]   R. Codina and O. Soto. Finite element implementation of two-equation and algebraic stress turbulence models for steady incompressible flows. *Int. J. Numer. Meth. Fluids*, 90:309–343, 1999.

[CS07]   A. Cangiani and E. Süli. The residual-free-bubble finite element method on anisotropic partitions. *SIAM J. Numer. Anal.*, 45(4):1654–1678 (electronic), 2007.

[CSB05]   P. Causin, R. Sacco, and C.L. Bottasso. Flux-upwind stabilization of the discontinuous Petrov-Galerkin formulation with Lagrange multipliers for advection-diffusion problems. *M2AN*, 39:1087–1114, 2005.

[CSX07]   L. Chen, P. Sun, and J. Xu. Optimal anisotropic meshes for minimizing interpolation errors in L^p-norm. *Math. Comp.*, 76:179–204, 2007.

[CX05]   L. Chen and J. Xu. An optimal streamline diffusion finite element method for a singularly perturbed problem. In *Recent advances in adaptive computation*, volume 383 of *Contemp. Math.*, pages 191–201. Amer. Math. Soc., Providence, RI, 2005.

[CX08]   L. Chen and J. Xu. Stability and accuracy of adapted finite element methods for singularly perturbed problems. *Numer. Math.*, 109:167–191, 2008.

[Dal95]   J. Dalík. A finite difference method for a two-dimensional convection-diffusion problem with dominating convection. In *Proc. Scient. Comm. Fac. Civil Eng.*, 1, pages 5–44. Brno Univ., 1995.

[dB74]   C. de Boor. Good approximation by splines with variable knots. II. In *Conference on the Numerical Solution of Differential Equations (Univ. Dundee, Dundee, 1973)*, pages 12–20. Lecture Notes in Math., Vol. 363. Springer, Berlin, 1974.

[DD74]   J. Douglas and T. Dupont. Galerkin approximations for the two point boundary value problem using continuous piecewise polynomial spaces. *Numer. Math.*, 22:99–109, 1974.

[Den74]   J. E. Dendy, Jr. Two methods of Galerkin type achieving optimum L^2 rates of convergence for first order hyperbolics. *SIAM J. Numer. Anal.*, 11:637–653, 1974.

[DF04]   V. Dolejsi and J. Felcman. Anisotropic mesh adaptation for numerical solution of boundary value problems. *Num. Meth. for Part. Diff. Equ.*, 20:576–608, 2004.

[DFS02]   V. Dolejsi, M. Feistauer, and C. Schwab. A finite volume discontinuous Galerkin scheme for nonlinear convection-diffusion problems. *Calcolo*, 39:1–40, 2002.

[dG76]   P. P. N. de Groen. *Singularly perturbed differential operators of second order.* Math. Center, Amsterdam, 1976.

[dG81]   P. P. N. de Groen. A finite element method with a large mesh-width for a stiff two-point boundary value problem. *J. Comput. Appl. Math.*, 7:3–15, 1981.

[dGH79]   P. P. N. de Groen and P. W. Hemker. Error bounds for exponentially fitted Galerkin methods to stiff boundary value problems. In *Numerical and Asymptotical Methods for Singular Perturbation Problems*, pages 217–249. Academic Press, 1979.

[DGT94]   O. Dorok, W. Grambow, and L. Tobiska. Aspects of the finite element discretizations for solving the Boussinesq approximation of the Navier-Stokes equations. In F.-K. Hebeker, R. Rannacher, and G. Wittum, editors, *Numerical methods for the Navier-Stokes equations. Proceedings of the Internat. Workshop*, pages 50–61, Heidelberg, 1994. Vieweg-Verlag.

[DHP99]   J. Douglas, Jr., C.-S. Huang, and F. Pereira. The modified method of characteristics with adjusted advection. *Numer. Math.*, 83:353–369, 1999.

[DKV06]   W. Dahmen, A. Kunoth, and J. Vorloeper. Convergence of adaptive wavelet methods for goal-oriented error estimation. In *Numerical mathematics and advanced applications*, pages 39–61. Springer, Berlin, 2006.

[DL02]   T. F. Dupont and Y. Liu. Symmetric error estimates for moving mesh Galerkin methods for advection-diffusion equations. *SIAM J. Numer. Anal.*, 40:914–927, 2002.

[DL06]   R. G. Durán and A. L. Lombardi. Finite element approximation of convection diffusion problems using graded meshes. *Appl. Numer. Math.*, 56:1314–1325, 2006.

[DMP07]   L. Dede, S. Micheletti, and S. Perotto. Anisotropic error control for environmental applications. *Appl. Numer. Math.*, 2007. Published online August 7.

[DMR91]   R. Duran, M. A. Muschietti, and R. Rodriguez. On the asymptotic exactness of error estimators for linear triangular elements. *Numer. Math.*, 59:107–127, 1991.

[DMR92]   R. Duran, M. A. Muschietti, and R. Rodriguez. Asymptotically exact error estimators for rectangular finite elements. *SIAM J. Numer. Anal.*, 29:78–88, 1992.

[DMS80]   E. P. Doolan, J. J. H. Miller, and W. H. A. Schilders. *Uniform numerical methods for problems with initial and boundary layers.* Boole Press, Dublin, 1980.

[Doe78]   E. J. Doedel. The construction of finite-difference approximations to ordinary differential equations. *SIAM J. Numer. Anal.*, 15:450–465, 1978.

[Doe96]   W. Doerfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33:1106–1124, 1996.

[Doe98]   W. Doerfler. Uniformly convergent finite element methods for singularly perturbed convection-diffusion equations. Habilitation, Univ. Freiburg, 1998.

[Doe99a] W. Doerfler. Uniform apriori estimates for singularly perturbed elliptic equations in multi-dimensions. *SIAM J. Numer. Anal.*, 36:1878–1900, 1999.

[Doe99b] W. Doerfler. Uniform error estimates for an exponentially fitted finite element method for singularly perturbed elliptic equations. *SIAM J. Numer. Anal.*, 36:1709–1738, 1999.

[Dor95]   O. Dorok. *Eine stabilisierte finite Elemente Methode zur Lösung der Boussinesq-Approximation der Navier-Stokes-Gleichungen.* PhD thesis, Univ. Magdeburg, 1995.

[DP01]   C. Dawson and J. Proft. A priori error estimates for interior penalty versions of the local discontinuous Galerkin method applied to transport equations. *Numer. Methods Partial Differential Equations*, 17:545–564, 2001.

[DR90]   M. Dobrowolski and H.-G. Roos. Stability estimates for singularly perturbed elliptic boundary value problems. Report Math. 07-19-90, TU Dresden, 1990.

[DR92]   R. Duran and R. Rodriguez. On the asymptotic exactness of Bank-Weiser's estimator. *Numer. Math.*, 62:297–303, 1992.

[DR95]   J. Dalík and H. Růžičková. An explicit modified method of characteristics for the one-dimensional nonstationary convection-diffusion problem with dominating convection. *Appl. Math.*, 40:367–380, 1995.

[DR97]   M. Dobrowolski and H.-G. Roos. A priori estimates for the solution of convection-diffusion problems and interpolation on Shishkin meshes. *Z. Anal. Anwendungen*, 16:1001–1012, 1997.

[DRH98]   J. Donea, B. Roig, and A. Huerta. *High-order accurate time-stepping schemes for convection-diffusion problems*, volume 42 of *Monograph Series*. International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 1998.

[DW89]   J. J. Douglas and J. Wang. An absolutely stabilized finite element for the Stokes problem. *Math. Comp.*, 52:495–508, 1989.

[EAE06]   L. El Alaoui and A. Ern. Nonconforming finite element methods with subgrid viscosity applied to advection-diffusion-reaction equations. *Numer. Methods Partial Differential Equations*, 22:1106–1126, 2006.

[Eck73]   W. Eckhaus. *Matched asymptotic expansions and singular perturbations*. North-Holland, Amsterdam, 1973.

[Eck79]   W. Eckhaus. *Asymptotic analysis of singular pertubations*. North-Holland, Amsterdam, 1979.

[EEHJ95] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to adaptive methods for differential equations. In *Acta Numerica, 1995*, pages 105–158. Cambridge Univ. Press, Cambridge, 1995.

[EG04]   A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[EJ91]   K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems I: a linear model problem. *SIAM J. Numer. Anal.*, 28:43–77, 1991.

[EJ93a]   K. Eriksson and C. Johnson. Adaptive streamline diffusion finite element methods for convection-diffusion problems. *Math. Comp.*, 60(201):167–188, 1993.

[EJ93b]   K. Eriksson and C. Johnson. Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems. *Math. Comp.*, 60:167–188, 1993.

[Eme70]   K. V. Emel'janov. On a difference scheme for a differential equation with a small parameter affecting the highest derivative (in Russian). *Num. Meth. Mech. Cont. Media*, 1:20–30, 1970.

[Eme73]   K. V. Emel'janov. A difference scheme for a three dimensional elliptic equation with a small parameter multiplying the highest derivative. In *Boundary value problems for equations of mathematical physics*, pages 30–42. USSR Academy of Sciences, Sverdlovsk, 1973.

[Eme82]   K. V. Emeljanov. The construction of difference schemes for linear sin-
           gularly perturbed boundary value problems (in Russian). *Proceedings
           Soviet Academy of Science*, 262:1052–1055, 1982.

[EP05]   A. Ern and J. Proft. A posteriori discontinuous Galerkin error estimates
           for transient convection-diffusion equations. *Appl. Math. Lett.*, 18:833–
           841, 2005.

[ESW05]   H. Elman, D. Silvester, and A. Wathen. *Finite elements and fast itera-
           tive solvers*. Numerical mathematics and scientific computation. Oxford
           University Press, Oxford, 2005.

[EW01]   R. E. Ewing and H. Wang. A summary of numerical methods for time-
           dependent advection-dominated partial differential equations. *J. Comp.
           Appl. Math.*, 128:423–445, 2001.

[Ewi83]   R. E. Ewing, editor. *The mathematics of reservoir simulation*, volume 1
           of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, PA, 1983.

[Far83]   P. A. Farrell. *Uniformly convergent difference schemes for singularly
           perturbed turning and non-turning point problems*. PhD thesis, Trinity
           College Dublin, 1983.

[Far88]   P. A. Farrell. Sufficient conditions for the uniform convergence of differ-
           ence schemes for a singularly perturbed turning point problem. *SIAM
           J. Numer. Anal.*, 25:618–643, 1988.

[FdC89]   L. P. Franca and E. G. Dutra do Carmo. The Galerkin-least-squares
           method. *Comput. Methods Appl. Mech. Eng.*, 74:41–54, 1989.

[Fei93]   M. Feistauer. *Mathematical methods in fluid dynamics*, volume 67 of *Pit-
           man Monographs and Surveys in Pure and Applied Mathematics*. Long-
           man Scientific & Technical, Harlow, 1993.

[Fel84]   A. Felgenhauer. Application of a generalized maximum principle to esti-
           mate corner layers in the n-dimensional case. In J. Förste, editor, *Singu-
           larly perturbed differential equations and applications*, number 03/84 in
           Report R-Mech., pages 1–8. Akademie der Wissenschaften, Berlin, 1984.

[Fel94]   A. Felgenhauer. A new analysis of an L-spline finite element method
           for singularly perturbed two-point boundary value problems. Technical
           Report 06, Bergakademie Freiberg, 1994.

[FF92]   L. P. Franca and S. L. Frey. Stabilized finite element methods: II. The
           incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech.
           Engrg.*, 99(2/3):209–233, 1992.

[FFH92]   L. P. Franca, S. L. Frey, and T. J. R. Hughes. Stabilized finite ele-
           ment methods: I. application to the advective-diffusive model. *Comput.
           Methods Appl. Mech. Eng.*, 95:253–276, 1992.

[FFLM95]   M. Feistauer, J. Felcman, and M. Lukáčová-Medvidová. Combined fi-
           nite element–finite volume solution of compressible flow. *J. Comput.
           Appl. Math.*, 63(1-3):179–199, 1995. International Symposium on Math-
           ematical Modelling and Computational Methods Modelling 94 (Prague,
           1994).

[FFLM97]   M. Feistauer, J. Felcman, and M. Lukáčová-Medvidová. On the con-
           vergence of a combined finite volume–finite element method for nonlin-
           ear convection-diffusion problems. *Numer. Methods Partial Differential
           Equations*, 13(2):163–190, 1997.

[FFS03]   M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and com-
           putational methods for compressible flow*. Numerical Mathematics and

Scientific Computation. The Clarendon Press Oxford University Press, Oxford, 2003.

[FG88]   P. A. Farrell and E. C. Gartland. A uniform convergence result for a turning point problem. In *Proceedings of BAIL V*, pages 127–132, Shanghai, 1988.

[FH88]   L. P. Franca and T. J. R. Hughes. Two classes of mixed finite element methods. *Appl. Mech. Engrg.*, 69:89–129, 1988.

[FHM+00] P. A. Farrell, A. F. Hegarty, J. J. Miller, E. O'Riordan, and G. I. Shishkin. *Robust Computational Techniques for Boundary Layers*. Chapman & Hall/CRC, Boca Raton, 2000.

[FHS96a] P. A. Farrell, P. W. Hemker, and G. I. Shishkin. Discrete approximations for singularly perturbed boundary value problems with parabolic layers. I. *J. Comput. Math.*, 14:71–97, 1996.

[FHS96b] P. A. Farrell, P. W. Hemker, and G. I. Shishkin. Discrete approximations for singularly perturbed boundary value problems with parabolic layers. II. *J. Comput. Math.*, 14:183–194, 1996.

[FHS96c] P. A. Farrell, P. W. Hemker, and G. I. Shishkin. Discrete approximations for singularly perturbed boundary value problems with parabolic layers. III. *J. Comput. Math.*, 14:273–290, 1996.

[FHŠ07]   M. Feistauer, J. Hájek, and K. Švadlenka. Space-time discontinuous Galerkin method for solving nonstationary convection-diffusion-reaction problems. *Appl. Math.*, 52:197–233, 2007.

[FJMT07] L. P. Franca, V. John, G. Matthies, and L. Tobiska. An inf-sup stable and residual-free bubble element for the Oseen equations. *SIAM J. Numer. Anal.*, 45(6):2392–2407 (electronic), 2007.

[FL08]   S. Franz and T. Linß. Superconvergence analysis of the Galerkin FEM for a singularly perturbed convection-diffusion problem with characteristic layers. *Numer. Methods Partial Differential Equations*, 24:144–164, 2008.

[FLR]   S. Franz, T. Linß, and H.-G. Roos. Superconvergence analysis of the SDFEM for elliptic problems with characteristic layers. *Appl. Numer. Math.* (to appear).

[FLR01]   A. Fröhner, T. Linß, and H.-G. Roos. Defect correction on Shishkin-type meshes. *Numer. Algorithms*, 26:281–299, 2001.

[FLRS08] S. Franz, T. Linß, H.-G. Roos, and S. Schiller. Uniform superconvergence of a finite element method with edge stabilization for convection-diffusion problems. Technical Report MATH-NM-01-2008, Technical University of Dresden, Dresden, 2008.

[FMM98]   J. M. Fiard, T. A. Manteuffel, and S. F. McCormick. First-order system least squares (fosls) for convection-diffusion problems: numerical results. *SIAM J. Sci. Comput.*, 19(6):1958–1979, 1998.

[FMOS98] P. A. Farrell, J. H. Miller, E. O'Riordan, and G. I. Shishkin. On the non-existence of ε-uniform finite difference methods on uniform meshes for semilinear two-point boundary value problems. *Math. Comp.*, 67:603–617, 1998.

[FMOS01] P. A. Farrell, J. H. Miller, E. O'Riordan, and G. I. Shishkin. Finite difference methods ε-uniform in the maximum norm for quasilinear differential equations with boundary layers. *Comp. Meth. Appl. Math.*, 1:154–172, 2001.

[FMP04]   L. Formaggia, S. Micheletti, and S. Perotto. Anisotropic mesh adaption in computational fluid dynamics: application to the advection-diffusion-reaction and Stokes problems. *Appl. Numer. Math.*, 51:511–533, 2004.

[FN01]   L. P. Franca and A. Neslituurk. On a two-level finite element method for the incompressible Navier-Stokes equations. *Int. J. Numer. Methods Eng.*, 52(4):433–453, 2001.

[FNS98]   L. P. Franca, A. Neslituurk, and M. Stynes. On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method. *Comput. Methods Appl. Mech. Engrg.*, 166:35–49, 1998.

[FOMS01] P. A. Farrell, E. O'Riordan, J. J. H. Miller, and G. I. Shishkin. Parameter-uniform fitted mesh method for quasilinear differential equations with boundary layers. *Comput. Methods Appl. Math.*, 1(2):154–172, 2001.

[For78]   M. Fortin. Résolution numérique des équations de Navier-Stokes par des méthodes d'éléments finis de type mixte. In *Proc. 2nd Int. Symp. Finite Elements in Flow Problems*, S. Margherita Ligure Italy, 1978.

[For88]   B. Fornberg. Generation of finite difference formulas. *Math. Comp.*, 51:702–705, 1988.

[FR92]   R. S. Falk and G. R. Richter. Local error estimates for a finite element method for hyperbolic and convection-diffusion equations. *SIAM J. Numer. Anal.*, 29:730–754, 1992.

[Fra94]   L. P. Franca. Incompressible flows based upon stabilized methods. In F.-K. Hebeker, R. Rannacher, and G. Wittum, editors, *Numerical methods for the Navier-Stokes equations. Proceedings of an Internat. Workshop*, number 47 in Notes on Numerical Fluid Mechanics, pages 89–100, Heidelberg, 1994. Vieweg-Verlag.

[Fri64]   A. Friedman. *Partial differential equations of parabolic type*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1964.

[FRSW99] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen. On parameter choice and iterative convergence for stabilised discretisations of advection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 179:179–195, 1999.

[FŠ04]   M. Feistauer and K. Švadlenka. Discontinuous Galerkin method of lines for solving nonstationary singularly perturbed linear problems. *J. Numer. Math.*, 12:97–117, 2004.

[FSS99]   M. Feistauer, J. Slavík, and P. Stupka. On the convergence of a combined finite volume-finite element method for nonlinear convection-diffusion problems. Explicit schemes. *Numer. Methods Partial Differential Equations*, 15(2):215–235, 1999.

[FT02]   L. P. Franca and L. Tobiska. Stability of the residual free bubble method for bilinear finite elements on rectangular grids. *IMA J. Numer. Anal.*, 22(1):73–87, 2002.

[FTZ08]   S. Franz, L. Tobiska, and H. Zarin. A new approach to recovery of discontinuous Galerkin. Technical report, Otto-von-Guericke Universität, Magdeburg, 2008. (in preparation).

[FVZ90]   R. M. Furzeland, J. G. Verwer, and P. A. Zegeling. A numerical study of three moving-grid methods for one-dimensional partial differential equations which are based on the method of lines. *J. Comput. Phys.*, 89:349–388, 1990.

[Gal94]   G. P. Galdi. *An introduction to the mathematical theory of the Navier–Stokes equations.* Springer-Verlag, 1994.

[Gar87]   E. C. Gartland. Uniform high-order difference schemes for a singularly perturbed two-point boundary value problem. *Math. Comp.*, 48:551–564, 1987.

[Gar88]   E. C. Gartland. Graded-mesh difference schemes for singularly perturbed two-point boundary value problems. *Math. Comp.*, 51:631–657, 1988.

[Gar89]   E. C. Gartland. Strong uniform stability and exact discretizations of a model singular perturbation problem and its finite difference approximations. *Appl. Math. Comput.*, 31:473–485, 1989.

[Gar91]   E. C. Gartland. On the stability of compact discretizations of singularly perturbed differential equations. In H.-G. Roos, A. Felgenhauer, and L. Angermann, editors, *Numerical methods in singularly perturbed problems. Proc.*, pages 63–70. TU Dresden, 1991.

[Gar93]   E. C. Gartland. On the uniform convergence of the Scharfetter-Gummel discretization in one dimension. *SIAM J. Numer. Anal.*, 30:749–758, 1993.

[GdC88]   A. Galeão and E. Dutra do Carmo. A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems. *Comput. Methods Appl. Mech. Engrg.*, 68(1):83–95, 1988.

[Geo05]   E. H. Georgoulis. hp-version interior penalty discontinuous Galerkin finite element methods on anisotropic meshes. Report University of Leicester, Deptm. of Mathematics, 14.02.2005, 2005.

[Geo06]   E. H. Georgoulis. *hp*-version interior penalty discontinuous Galerkin finite element methods on anisotropic meshes. *Int. J. Numer. Anal. Model.*, 3:52–79, 2006.

[GF88]   A.L. Goncharov and I.V. Fryazinov. Difference schemes on a nine-point "cross" pattern for solving the Navier-Stokes equations. *U.S.S.R. Comput. Math. Math. Phys.*, 28:164–172, 1988.

[GFL+83]   H. Goering, A. Felgenhauer, G. Lube, H.-G. Roos, and L. Tobiska. *Singularly perturbed differential equations.* Akademie-Verlag, Berlin, 1983.

[GHH07a]   E. H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods on hp-anisotropic meshes I: a priori analysis. *Intern. J. Comput. Sci. Math.*, 1:221–244, 2007.

[GHH07b]   E. H. Georgoulis, E. Hall, and P. Houston. Discontinuous Galerkin methods on hp-anisotropic meshes II: a posteriori analysis. Report MA 07-10, University of Leicester, 2007.

[GK03]   J. Gopalakrishnan and G. Kanschat. A multilevel discontinuous Galerkin method. *Numerische Mathematik*, 95:527–550, 2003.

[GKO95]   B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods.* Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, 1995. A Wiley-Interscience Publication.

[GL78]   D. F. Griffiths and J. Lorenz. An analysis of the Petrov-Galerkin method applied to a model problem. *Comput. Methods Appl. Mech. Eng.*, 14:39–64, 1978.

[GL07]   J. L. Gracia and F. J. Lisbona. A uniformly convergent scheme for a system of reaction-diffusion equations. *J. Comput. Appl. Math.*, 206:1–16, 2007.

[GLOS05] T. Gelhard, G. Lube, M. A. Olshanskii, and J.-H. Starcke. Stabilized finite element schemes with LBB-stable elements for incompressible flows. *J. Comput. Appl. Math.*, 177:243–267, 2005.

[GMQ06]   J.-L. Guermond, A. Marra, and L. Quartapelle. Subgrid stabilized projection method for 2d unsteady flows at high Reynolds numbers. *Comput. Methods Appl. Mech. Engrg.*, 195:5857–5876, 2006.

[GMSS01] K. Gerdes, J. M. Melenk, C. Schwab, and D. Schötzau. The *hp*-version of the streamline diffusion finite element method in two space dimensions. *Math. Models Methods Appl. Sci.*, 11:301–337, 2001.

[GMT08]   S. Ganesan, G. Matthies, and L. Tobiska. Local projection stabilization of equal order interpolation applied to the Stokes problem. *Math. Comp.*, 2008.

[GP69]   B. F. Gromov and V. S. Petrishchev. On numerical solution to two-dimensional incompressible viscous fluid problems. In *Proc. of All-Union on Numerical Methods of Mechanics of Viscous Fluids*, pages 74–87, Novosibirsk, 1969. Nauka. (in Russian).

[GP83]   M. D. Gunzburger and J. S. Peterson. On conforming finite element methods for the inhomogeneous stationary Navier-Stokes equations. *Numer. Math.*, 42(2):173–194, 1983.

[GR79]   V. Girault and P.-A. Raviart. *Finite element approximation of the Navier–Stokes equations*. Number 749 in Lect. Notes Math. Springer-Verlag, 1979.

[GR82]   V. Girault and P.-A. Raviart. An analysis of upwind schemes for the Navier–Stokes equations. *SIAM J. Numer. Analysis*, 19(2):312–333, 1982.

[GR86]   V. Girault and P.-A. Raviart. *Finite element methods for Navier–Stokes equations*. Springer-Verlag, Berlin, 1986.

[Gri85a] R. Grigorieff. Einige Stabilitätsungleichungen für Differenzenverfahren in nichtkompakter Form. *Numerische Mathematik*, 47:565–576, 1985.

[Gri85b] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman, Boston, 1985.

[GRS90]   W. Grambow, U. Risch, and F. Schieweck. Experiences with the multigrid method applied to high Reynolds number flow. Preprint Math. 6/1990 Univ. Magdeburg, 1990.

[GRS07]   C. Grossmann, H.-G. Roos, and M. Stynes. *Numerical treatment of partial differential equations*. Springer-Verlag, Berlin Heidelberg, 2007.

[GS74]   V. A. Gushchin and V. V. Shchennikov. On a monotone difference scheme of the second order (in Russian). *Z. Vycisl. Mat. i Mat. Fiz.*, 14:789–792, 1974.

[GS93]   W. Guo and M. Stynes. Finite element analysis of exponentially fitted lumped schemes for time-dependent convection-diffusion problems. *Numer. Math.*, 66:347–371, 1993.

[GS94]   W. Guo and M. Stynes. Finite element analysis of an exponentially fitted non-lumped scheme for advection-diffusion equations. *Appl. Numer. Math.*, 15:375–393, 1994.

[GS97]   W. Guo and M. Stynes. Pointwise error estimates for a streamline diffusion scheme on a Shishkin mesh for a convection-diffusion problem. *IMA J. Numer. Anal.*, 17(1):29–59, 1997.

[GS00a]   P. M. Gresho and R. L. Sani. *Incompressible Flow and the finite element method. Advection-diffusion*. Wiley John & Sons, 2000.

[GS00b]   P. M. Gresho and R. L. Sani. *Incompressible Flow and the finite element method. Isothermal Laminar Flow.* Wiley John & Sons, 2000.

[GS02]   M. B. Giles and E. Süli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11:145–236, 2002.

[GS03]   V. Girault and L. R. Scott. A quasi-local interpolation operator preserving the discrete divergence. *Calcolo*, 40(1):1–19, 2003.

[GS04]   J. Grasman and S.-D. Shih. A parabolic singular perturbation problem with an internal layer. *Asymptot. Anal.*, 38:309–318, 2004.

[GT83]   D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order.* Springer-Verlag, Berlin, 1983.

[Gue99a] J.-L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *M2AN Math. Model. Numer. Anal.*, 33:1293–1316, 1999.

[Gue99b] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear contraction semi-groups of class C^0. *Comput. Visual. Sci.*, 2:131–138, 1999.

[Gue01a] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear contraction semi-groups of class C^0 in Hilbert spaces. *Numer. Meth. Part. Diff. Equat.*, 17:1–25, 2001.

[Gue01b] J.-L. Guermond. Subgrid stabilization of Galerkin approximations of linear monotone operators. *IMA J. Numer. Anal.*, 21:165–197, 2001.

[Gün88]   C. Günther. Vergleich verschiedener Differenzenverfahren zur numerischen Lösung der Konvektions-Diffusionsgleichung. Bericht KfK 4439. Kernforschungszentrum Karlsruhe, 1988.

[Gun89]   M. D. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows.* Academic Press, 1989.

[Gün92]   C. Günther. Conservative versions of the locally exact consistent upwind scheme of second order (lecusso-scheme). *Int. J. Numer. Methods Eng.*, 34:793–804, 1992.

[Gun96]   M. D. Gunzburger. Navier-Stokes equations for incompressible flows: finite-element methods. In *Handbook of computational fluid mechanics*, pages 99–157. Academic Press, San Diego, CA, 1996.

[Guo93]   W. Guo. *Uniformly convergent finite element methods for singularly perturbed parabolic partial differential equations.* PhD thesis, University College, Cork, Ireland, 1993.

[Guz06]   J. Guzman. Local analysis of discontinuous Galerkin methods applied to singularly perturbed problems. *J. Numer. Math.*, 14:41–56, 2006.

[GWR04]   V. Gravemeier, W. A. Wall, and E. Ramm. A three-level finite element method for the instationary incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 193:1323–1366, 2004.

[GWR05]   V. Gravemeier, W. A. Wall, and E. Ramm. Large eddy simulation of turbulent incompressible flows by a three-level finite element method. *Int. J. Numer. Meth. Fluids*, 48:1067–1099, 2005.

[Han92]   P. Hansbo. The characteristic streamline diffusion method for convection diffusion problems. *Comput. Methods Appl. Mech. Eng.*, 96:239–253, 1992.

[HB79]   T. J. R. Hughes and A. N. Brooks. A multidimensional upwind scheme with no crosswind diffusion. In T.J.R. Hughes, editor, *Finite Element*

Methods for Convection Dominated Flows, volume 34 of *AMD*. ASME, New York, 1979.

[HBS87] S.-Y. Hahn, J. Bigeon, and J.-C. Sabonnadiere. An 'upwind' finite element method for electromagnetic field problems in moving media. *Int. J. Numer. Methods Eng.*, 24:2071–2086, 1987.

[HD05] G. Hauke and M. H. Doweidar. Fourier analysis of semi-discrete and space-time stabilized methods for the advective-diffusive-reactive equation. I. SUPG. *Comput. Methods Appl. Mech. Engrg.*, 194:45–81, 2005.

[Heg82] A. F. Hegarty. Uniformly convergent finite difference schemes for a two-dimensional singular perturbation problem. In J. J. H. Miller, editor, *Computational and asymptotic methods for boundary and interior layers*, pages 263–268. Boole Press, Dublin, 1982.

[Hei80] J. C. Heinrich. On quadratic elements in finite element solutions of steady-state convection-diffusion equation. *Int. J. Numer. Meth. Engng.*, 15:1041–1052, 1980.

[Hem77] P. W. Hemker. A numerical study of stiff two-point boundary value problems. Mathematical Centre Tracts 80. Amsterdam: Mathematisch Centrum, 1977.

[Hem97] P. W. Hemker. A model singular perturbation problem. *Journal Comp. Appl. Math.*, 76:277–285, 1997.

[Her90] D. Herceg. Uniform fourth order difference scheme for a singularly perturbed problem. *Numerische Mathematik*, 56:675–693, 1990.

[HFB86] T. J. R. Hughes, L. P. Franca, and M. Balestra. A new finite element formulation for computational fluid dynamics. V: Circumventing the Babuška–Brezzi condition: A stable Petrov–Galerkin formulation of the Stokes problem accomodating equal–order interpolations. *Comput. Methods Appl. Mech. Engrg.*, 59:85–99, 1986.

[HFB87] T. J. R. Hughes, L. P. Franca, and M. Balestra. Errata: "A new finite element formulation for computational fluid dynamics. V. Circumventing the Babuška-Brezzi condition: a stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations". *Comput. Methods Appl. Mech. Engrg.*, 62(1):111, 1987.

[HFH89] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII the Galerkin/least-squares method for advective-diffusive equations. *Comput. Methods Appl. Mech. Eng.*, 73:173–189, 1989.

[HFMQ98] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J. B. Quincy. The variational multiscale method – a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 166:3–24, 1998.

[HGG84] A. C. Hindmarsh, P. M. Gresho, and D. F. Griffiths. The stability of explicit Euler time-integration for certain finite difference approximations of the multi-dimensional advection-diffusion equation. *Int. J. Numer. Methods Fluids*, 4:853–897, 1984.

[HGH91] D. F. Hawken, J. J. Gottlieb, and J. S. Hansen. Review of some adaptive node-movement techniques in finite-element and finite-difference solutions of partial differential equations. *J. Comput. Phys.*, 95:254–302, 1991.

[HGH08] P. Houston, E. H. Georgoulis, and E. Hall. Adaptivity and a posteriori error estimation for dG methods on anisotropic meshes. in: Proc. of BAIL 2006, Goettingen, G. Lube, G, Rapin (Eds.), 2008.

[HH92]   I. Harari and T. J. R. Hughes. What are *c* and *h*?: Inequalities for the analysis and design of finite element methods. *Comput. Methods Appl. Mech. Eng.*, 97:157–192, 1992.

[HHG92]   D. F. Hawken, J. S. Hansen, and J. J. Gottlieb. A new finite-difference solution adaptive method. *Phil. Trans. R. Soc. Lond. A*, 341:373–410, 1992.

[HHH00]   E. D. Havik, P. W. Hemker, and W. Hoffmann. Application of the overset grid technique to a model singular perturbation problem. *Computing*, 65:339–356, 2000.

[HHSS03] K. Harriman, P. Houston, B. Senior, and E. Süli. *hp*-version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. In *Recent advances in scientific computing and partial differential equations (Hong Kong, 2002)*, volume 330 of *Contemp. Math.*, pages 89–119. Amer. Math. Soc., Providence, RI, 2003.

[HHZM77] J. C. Heinrich, P. S. Huyakorn, O. C. Zienkiewicz, and A. R. Mitchell. An upwind finite element scheme for two-dimensional convective transport equation. *Int. J. Numer. Meth. Engng.*, 11:131–143, 1977.

[Hir88]   C. Hirsch. *Numerical computation of internal and external flows*, volume 1. Wiley, Chichester, 1988.

[Hir90]   C. Hirsch. *Numerical computation of internal and external flows*, volume 2. Wiley, Chichester, 1990.

[HJ04]   J. Hoffman and C. Johnson. Computability and adaptivity in CFD. In E. Stein, R. de Borst, and T.J.R. Hughes, editors, *Encyclopedia of Computational Mechanics*, volume 3, pages 183–206. John Wiley & Sons, 2004.

[HK82]   H. Han and R. B. Kellogg. A method of enriched subspaces for the numerical solution of a parabolic singular perturbation problem. In J.J.H. Miller, editor, *Computational and asymptotic methods for boundary and interior layers*, pages 46–52. Boole Press, Dublin, 1982.

[HK90]   H. Han and R. B. Kellogg. Differentiability properties of solutions of the equation $-\varepsilon^2 \triangle u + ru = f$ in a square. *SIAM J. Math. Anal.*, 21:394–408, 1990.

[HLP03]   J. M. Hyman, S. Li, and L. R. Petzold. An adaptive moving mesh method with static rezoning for partial differential equations. *Comput. Math. Appl.*, 46:1511–1524, 2003.

[HM81]   W. Höhn and H. D. Mittelmann. Das diskrete Maximumprinzip für finite Elemente höherer Ordnung. *Computing*, 27:145–154, 1981.

[HMM86]   T. J. R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. *Comput. Methods Appl. Engrg.*, 54(3):341–355, 1986.

[HMMR95] X. C. Hu, T. A. Manteuffel, S. McCormick, and T. F. Russell. Accurate discretization for singular perturbations: the one-dimensional case. *SIAM J. Numer. Anal.*, 32:83–109, 1995.

[HMO80]   A. F. Hegarty, J. J. H. Miller, and E. O'Riordan. Uniform second order difference schemes for singular perturbation problems. In J. J. H. Miller, editor, *BAIL II-Proceedings*, pages 301–305, Dublin, 1980. Boole Press.

[HMOS95] A. F. Hegarty, J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. Special meshes for finite difference approximations to an advection-diffusion equation with parabolic layers. *J. Comput. Physics*, 117:47–54, 1995.

[HMOS97] A. F. Hegarty, J. H. Miller, E. O'Riordan, and G. I. Shishkin. Numerical solution of elliptic convection-diffusion problems on fitted meshes. *CWI Quarterly*, 10:239–251, 1997.

[HOS93]   A. F. Hegarty, E. O'Riordan, and M. Stynes. A comparison of uniformly convergent difference schemes for two-dimensional convection-diffusion problems. *J. Comput. Phys.*, 105:24–32, 1993.

[How78]   F. A. Howes. Boundary-interior layer interactions in nonlinear singular perturbation theory. *Memoirs of the Amer. Math. Soc.*, 15, 1978.

[HR82]   J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier–Stokes problem, part I: regularity of solutions and second order error estimates for the spatial discretization. *SIAM J. Numer. Anal.*, 19:275–311, 1982.

[HR86]   J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier–Stokes problem, part II: Stability of solutions and error estimates uniform in time. *SIAM J. Numer. Anal.*, 23:750–777, 1986.

[HR88]   J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier–Stokes problem, part III smoothing property and higher order estimates for spatial discretization. *SIAM J. Numer. Anal.*, 25:489–512, 1988.

[HR90]   J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier–Stokes problem, part IV: Error analysis for second order time discretization. *SIAM J. Numer. Anal.*, 27:353–384, 1990.

[HR93]   R. W. Healy and T. F. Russell. A finite-volume Eulerian-Lagrangian localized adjoint method for solution of the advection-dispersion equation. *Water Resour. Res.*, 29:23992413, 1993.

[HR01]   W. Huang and R. D. Russell. Adaptive mesh movement—the MMPDE approach and its applications. *J. Comput. Appl. Math.*, 128:383–398, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.

[HRR94]   W. Huang, Y. Ren, and R. D. Russell. Moving mesh partial differential equations (mmpdes) based on the equidistribution principle. *SIAM J. Numer. Anal.*, 31:709–731, 1994.

[HRS00]   P. Houston, R. Rannacher, and E. Süli. A posteriori error analysis for stabilised finite element approximations of transport problems. *Comput. Meth. Appl. Mech. Engrg.*, 190:1483–1508, 2000.

[HS88]   T. J. R. Hughes and F. Shakib. Computational aerodynamics and the finite element method. In *Proc. AIAA/AAS Astrodynamics Conf. Reno*, AIAA–88–0031, Jan. 11-15 1988.

[HS90]   P. Hansbo and A. Szepessy. A velocity-pressure streamline diffusion finite element method for the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Eng.*, 84:175–192, 1990.

[HS93]   P. W. Hemker and G. I. Shishkin. Approximation of parabolic PDEs with a discontinuous initial condition. *East-West J. Numer. Math.*, 1:287–302, 1993.

[HS94]   P. W. Hemker and G. I. Shishkin. Discrete approximation of singularly perturbed parabolic pdes with a discontinuous initial condition. *Comp. fluid dyn. J.*, 2:375–392, 1994.

[HS96]   T. J. R. Hughes and J. R. Stewart. A space-time formulation for multiscale phenomena. *J. Comput. Appl. Math.*, 74:217–229, 1996. TICAM Symposium (Austin, TX, 1995).

[HS01a]   P. Houston and E. Süli. Adaptive Lagrange-Galerkin methods for unsteady convection-diffusion problems. *Math. Comp.*, 70:77–106, 2001.

[HS01b]   P. Houston and E. Süli. Stabilised *hp*-finite element approximation of partial differential equations with nonnegative characteristic form. *Computing*, 66:99–119, 2001.

[HS07]   T. J. R. Hughes and G. Sangalli. Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.*, 45:539–557, 2007.

[HSS00]   P. W. Hemker, G. I. Shishkin, and L. P. Shishkina. ε-uniform schemes with high-order time-accuracy for parabolic singular perturbation problems. *IMA J. Numer. Anal.*, 20:99–121, 2000.

[HSS02]   P. Houston, C. Schwab, and E. Süli. Discontinuous hp-finite element methods for advection-diffusion-reaction problems. *SIAM J. Num. Anal.*, 39:2133–2163, 2002.

[HSS03]   P. W. Hemker, G. I. Shishkin, and L. P. Shishkina. Novel direct-correction high-order, in space and time, accurate schemes for parabolic singularly perturbed convection-diffusion problems. *Comput. Methods Appl. Math.*, 3:387–404, 2003.

[HSW07]   P. Houston, D. Schoetzau, and T. P. Wihler. Energy norm a posteriori error estimation of *hp*-adaptive discontinuous Galerkin methods for elliptic problems. *Math. Models Methods Appl. Sci.*, 17:33–62, 2007.

[Hua01]   W. Huang. Practical aspects of formulation and solution of moving mesh partial differential equations. *J. Comput. Phys.*, 171:753–775, 2001.

[Hua06]   W. Huang. Mathematical principles of anisotropic mesh adaption. *Comm. Comput. Phys.*, 1:276–310, 2006.

[Hug95]   T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins od stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127:387–401, 1995.

[HV03]   W. Hundsdorfer and J. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, volume 33 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2003.

[HZ77]   J. C. Heinrich and O. C. Zienkiewicz. Quadratic finite elements schemes for two-dimensional convective transport problems. *Int. J. Numer. Meth. Engng.*, 11:1831–1844, 1977.

[Ike83]   T. Ikeda. *Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena*. North-Holland, 1983.

[Il'69]   A. M. Il'in. A difference scheme for a differential equation with a small parameter affecting the highest derivative (in Russian). *Mat. Zametki*, 6:237–248, 1969.

[INSB96] S. Idelsohn, N. Nigro, M. Storti, and G. Buscaglia. A Petrov-Galerkin formulation for advection-reaction-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 136(1-2):27–46, 1996.

[IW99]   U. Iben and G. Warnecke. A posteriori error estimation for singularly perturbed boundary value problems. Report 23, Univ. Magdeburg, 1999.

[Jak59]   M. Jakob. *Heat transfer*. Wiley, New York, 1959.

[Jim96]   P. K. Jimack. A best approximation property of the moving finite element method. *SIAM J. Numer. Anal.*, 33:2286–2302, 1996.

[JK06a]   V. John and S. Kaya. A finite element variational multiscale method for the Navier-Stokes equations. *SIAM J. Sci. Comput.*, 26:1485–1503, 2006.

[JK06b]   V. John and P. Knobloch. On discontinuity-capturing methods for convection-diffusion equations. In *Numerical mathematics and advanced applications*, pages 336–344. Springer, Berlin, 2006.

[JK07a]   V. John and P. Knobloch. On the performance of SOLD methods for convection-diffusion problems with interior layers. *Int. J. Computing Science and Mathematics*, 1:245–258, 2007.

[JK07b]   V. John and P. Knobloch. Spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: Part I – a review. *Comput. Methods Appl. Mech. Engrg.*, 196:2197–2215, 2007.

[JK07c]   V. John and P. Knobloch. Spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: Part II – analysis for P_1 and Q_1 finite elements. Preprint MATH. 2007/4, Charles University, Faculty of Mathematics and Physics, Prague, 2007.

[JKL06]   V. John, S. Kaya, and W. J. Layton. A two-level variational multiscale method for convection-dominated convection–diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 195:4594–4603, 2006.

[JNP84]   C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Comp. Meth. Appl. Mech. Engrg.*, 45:285–312, 1984.

[Joh87]   C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, Cambridge, 1987.

[Joh95]   C. Johnson. On the computability and error control in CFD. *Intern. J. Numer. Methods Fluids*, 20:777–788, 1995.

[Joh96]   K. Johannsen. Aligned 3d-finite-volumes for convection-diffusion problems. In F. Benkhaldoun and R. Vilsmeier, editors, *Finite volumes for complex applications*, pages 291–300. Hermes, Paris, 1996.

[Joh00]   V. John. A numerical study of a posteriori error estimators for convection-diffusion equations. *Comput. Meth. Appl. Mech. Engrg.*, 190:757–781, 2000.

[Joh06]   V. John. On large eddy simulation and variational multiscale methods in the numerical simulation of turbulent flows. *Appl. Math.*, 51(4):321–353, 2006.

[JR94]   C. Johnson and R. Rannacher. On error control in CFD. In *Numerical methods for the Navier-Stokes equations. Proceedings of an Internat. Workshop*, pages 133–144, Heidelberg, 1994. Vieweg-Verlag.

[JRB95a]   C. Johnson, R. Rannacher, and M. Boman. Numerics and hydrodynamic stability: Toward error control in computational fluid dynamics. *SIAM J. Numer. Anal.*, 32(4):1058–1079, 1995.

[JRB95b]   C. Johnson, R. Rannacher, and M. Boman. On transition to turbulence and error control in CFD. Preprint 95-06 Universität Heidelberg, 1995.

[JS86]   C. Johnson and J. Saranen. Streamline diffusion methods for the incompressible Euler and Navier-Stokes equations. *Math. Comp.*, 47(175):1–18, 1986.

[JSW87]   C. Johnson, A. H. Schatz, and L. B. Wahlbin. Crosswind smear and pointwise errors in the streamline diffusion finite element methods. *Math. Comp.*, 49(179):25–38, 1987.

[Kel71]   H. B. Keller. A new difference scheme for parabolic problems. In B. Hubbard, editor, *Numerical solution of partial differential equations II - SYNSPADE 1970*, pages 327–350. Academic Press, New York, 1971.

[KKN00]   S. Korotov, M. Křížek, and P. Neittaanmäki. Weakend acute type condition for thetrahedral triangulations and the discrete maximum principles. *Math. Comp.*, 70(233):107–119, 2000.

[KL01]   N. Kopteva and T. Linß. Uniform second-order pointwise convergence of a central difference approximation for a quasilinear convection-diffusion problem. *J. Comput. Appl. Math.*, 137:257–267, 2001.

[KL04]   H.-O. Kreiss and J. Lorenz. *Initial-boundary value problems and the Navier-Stokes equations*, volume 47 of *Classics in Appl. Math.* SIAM, Philadelphia, PA, 2004. Reprint of the 1989 edition.

[KLR02]   T. Knopp, G. Lube, and G. Rapin. Stabilized finite element methods with shock capturing for advection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 191(27-28):2997–3013, 2002.

[KLS]   R. B. Kellogg, T. Linß, and M. Stynes. A finite difference method on layer-adapted meshes for an elliptic reaction-diffusion system in two dimensions. *Math. Comp.* (to appear).

[KMS⁺86]   H. O. Kreiss, T. A. Manteuffel, B. Schwartz, B. Wendroff, and A. B. White. Supra-convergent schemes on irregular grids. *Math. Comp.*, 47:537–554, 1986.

[KMS08]   R. B. Kellogg, N. Madden, and M. Stynes. A parameter-robust numerical method for a system of reaction-diffusion equations in two dimensions. *Numer. Methods Partial Differential Equations*, 24:312–334, 2008.

[KN87]   M. Křížek and P. Neittaanmäki. On superconvergence techniques. *Acta Applicandae Mathematicae*, 9:175–198, 1987.

[KNB86]   H. O. Kreiss, N. K. Nichols, and D. R. Brown. Numerical methods for stiff two-point boundary value problems. *SIAM J. Numer. Anal.*, 23:325–368, 1986.

[Kno]   P. Knobloch. On the choice of SUPG parameter on outflow boundary layers. *Adv. Comput. Math.* DOI 10.1007/ss10444-008-9075-6.

[Kno06]   P. Knobloch. Improvements of the Mizukami-Hughes method for convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 196(1-3):579–594, 2006.

[KNZ80]   D. W. Kelly, S. Nakazawa, and O. C. Zienkiewicz. A note on upwinding in finite element approximation to convection-diffusion problems. *Int. J. Numer. Methods Eng.*, 15:1705–1711, 1980.

[Kop96]   N. V. Kopteva. On the convergence, uniform with respect to a small parameter, of a four-point scheme for a one-dimensional stationary convection-diffusion equation. *Differential Equations*, 32:958–964, 1996.

[Kop97]   N. V. Kopteva. On the convergence, uniform with respect to a small parameter, of a scheme with weights for a one-dimensional nonstationary convection-diffusion equation. *Comput. Math. Math. Phys.*, 37:1173–1180, 1997.

[Kop98]   N. V. Kopteva. The two-dimensional Sobolev inequality in the case of an arbitrary grid. *Zh. Vychisl. Mat. Mat. Fiz.*, 38(4):596–599, 1998.

[Kop99]   N. V. Kopteva. On the convergence, uniform with respect to the small parameter, of a scheme with central difference on refined grids. *Comput. Math. Math. Phys.*, 39(10):1594–1610, 1999.

[Kop01a]   N. Kopteva. Maximum norm a posteriori error estimates for a one-dimensional convection-diffusion problem. *SIAM J. Numer. Anal.*, 39:423–441, 2001.

[Kop01b]   N. Kopteva. Uniform pointwise convergence of difference schemes for convection-diffusion problems on layer-adapted meshes. *Computing*, 66:179–197, 2001.

[Kop03]   N. Kopteva. Error expansion for an upwind scheme applied to a two-dimensional convection-diffusion problem. *SIAM J. Numer. Anal.*, 41:1851–1869, 2003.

[Kop04]   N. Kopteva. How accurate is the streamline-diffusion FEM inside characteristic (boundary and interior) layers? *Comput. Methods Appl. Mech. Engrg.*, 193:4875–4889, 2004.

[Kop05]   N. Kopteva. Maximum norm a posteriori error estimates for a 1d singularly perturbed semilinear reaction-diffusion problem. Technical report, University of Limerick, 2005.

[Kop07a]   N. Kopteva. Maximum norm error analysis for a 2d singularly perturbed semilinear reaction-diffusion problem. *Math. Comp.*, 76:631–646, 2007.

[Kop07b]   N. Kopteva. Pointwise error estimates for 2d singularly perturbed semilinear reaction-diffusion problems. In I. Farago, P. Vabishchevich, and L. Vulkov, editors, *Finite Difference Methods: Theory and Applications*, pages 105–114, 2007.

[Kop08]   N. Kopteva. Maximum norm a posteriori error estimate for a 2d singularly perturbed reaction-diffusion problem. *SIAM J. Numer. Anal.*, 2008. Electronic publication April 11.

[KP07]   O. A. Karakashian and F. Pascal. Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems. *SIAM J. Numer. Anal.*, 45(2):641–665 (electronic), 2007.

[KP08]   D. Kim and E.-J. Park. A posteriori error estimators for the upstream weighted mixed methods for convection-diffusion problems. *Comput. Meth. Appl. Mech. Engrg.*, 197:806–820, 2008.

[KR92]   F. Kratsch and H.-G. Roos. Monotonieerhaltende upwind-Schemata im zweidimensionalen Fall. *ZAMM*, 72:201–208, 1992.

[KR02]   G. Kanschat and R. Rannacher. Local error analysis of the interior penalty discontinuous Galerkin method for second order elliptic problems. *J. Numer. Math.*, 10:249–274, 2002.

[KR05]   S. Kaya and B. Rivière. A two-grid stabilization method for solving the steady-state Navier-Stokes equations. *Numer. Meth. Part. Diff. Equat.*, 22:728–743, 2005.

[Kra87]   F. Kratsch. *Ein invers monotones Verfahren zur numerischen Lösung singulär gestörter Probleme in der Ebene*. PhD thesis, TU Dresden, 1987.

[Kro97]   D. Kroener. *Numerical Schemes for Conservation Laws*. Wiley-Teubner, Stuttgart, 1997.

[KS97]   R. B. Kellogg and M. Stynes. Optimal approximability of solutions of singularly perturbed two-point boundary value problems. *SIAM J. Numer. Anal.*, 34(5):1808–1816, 1997.

[KS99]   R. B. Kellogg and M. Stynes. n-widths and singularly perturbed boundary value problems. *SIAM J. Numer. Anal.*, 36:1604–1620, 1999.

[KS01a]   R. B. Kellogg and M. Stynes. n-widths and singularly perturbed boundary value problems. II. *SIAM J. Numer. Anal.*, 39:690–707, 2001.

[KS01b]   N. Kopteva and M. Stynes. Approximation of derivatives in a convection-diffusion two-point boundary value problem. *Appl. Numer. Math.*, 39:47–60, 2001.

[KS01c]   N. Kopteva and M. Stynes. A robust adaptive method for a quasilinear one-dimensional convection-diffusion problem. *SIAM J. Numer. Anal.*, 39:1446–1467, 2001.

[KS04]   N. Kopteva and M. Stynes. Numerical analysis of a singularly perturbed nonlinear reaction-diffusion problem with multiple solutions. *Appl. Numer. Math.*, 51:273–288, 2004.

[KS05]   R.B. Kellogg and M. Stynes. Corner singularities and boundary layers in a simple convection-diffusion problem. *J. Differential Equations*, 213:81–120, 2005.

[KS06]   R.B. Kellogg and M. Stynes. A singularly perturbed convection-diffusion problem in a half-plane. *Appl. Anal.*, 85:1471–1485, 2006.

[KS07a]   R.B. Kellogg and M. Stynes. Sharpened bounds for corner singularities and boundary layers in a simple convection-diffusion problem. *Appl. Math. Lett.*, 20:539–544, 2007.

[KS07b]   N. Kopteva and M. Stynes. Approximation of interior-layer solutions in a singularly perturbed semilinear reaction-diffusion problem. Technical report, University of Limerick, 2007. (submitted for publication).

[KT78]   R. B. Kellogg and A. Tsan. Analysis of some difference approximations for a singularly perturbed problem without turning points. *Math. Comp.*, 32:1025–1039, 1978.

[KT08]   P. Knobloch and L. Tobiska. On the stability of finite element discretizations of convection–diffusion–reaction equations. Preprint Otto von Guericke University Magdeburg, 2008.

[Kuh99]   G. Kuhnert. A posteriori error estimation for anisotropic tetrahedral and triangular finite element meshes. PhD thesis, Univ. Chemnitz, 1999.

[Kuh05]   G. Kuhnert. A posteriori H1 error estimation for a singularly perturbed reaction-diffusion problem on anisotropic meshes. *IMA J. Numer. Anal.*, 25:408–428, 2005.

[Kun86]   M. Kunze. *Eine Finite-Element-Methode vom Galerkin-Petrov-Typ zur numerischen Lösung einer Klasse stationärer Diffusions-Konvektions-Gleichungen unter besonderer Berücksichtigung lokaler Abschätzungen.* PhD thesis, TH Magdeburg, 1986.

[Lax61]   P. D. Lax. On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients. *Comm. Pure Appl. Math.*, 14:497–520, 1961.

[Lay93]   W. Layton. Optimal difference schemes for 2-D transport problems. *J. Comp. Appl. Math.*, 45:337–341, 1993.

[LBD$^+$02] Y. Liu, R. E. Bank, T. F. Dupont, S. Garcia, and R. F. Santos. Symmetric error estimates for moving mesh mixed methods for advection-diffusion equations. *SIAM J. Numer. Anal.*, 40(6):2270–2291 (electronic) (2003), 2002.

[LCHR03]   J. Lang, W. Cao, W. Huang, and R. D. Russell. A two-dimensional moving finite element method with local refinement based on a posteriori error estimates. *Appl. Numer. Math.*, 46:75–94, 2003.

[Lel76]   E. F. Lelikova. On the asymptotic solution of an elliptic equation of the second order with a small parameter effecting the highest derivative (in Russian). *Diff. Equations*, 12:1852–1865, 1976.

[Len00a]   W. Lenferink. Pointwise convergence of approximations to a convection-diffusion equation on a Shishkin mesh. *Appl. Numer. Math.*, 32(1):69–86, 2000.

[Len00b]   W. Lenferink. Some superconvergence results for finite element discretizations on a Shishkin mesh of a convection-diffusion problem. In *Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems*, pages 193–198. Nova Science, Huntington, NY, 2000.

[Leo79a]   B. P. Leonard. A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Comput. Methods Appl. Mech. Eng.*, 19:59–98, 1979.

[Leo79b]   B. P. Leonard. A survey of finite differences of opinion on numerical modelling of the incompressible diffusion convection equation. In T.J.R. Hughes, editor, *Finite element methods for convection dominated flows*, pages 1–30. ASME, New York, 1979.

[LeV90]   R. J. LeVeque. *Numerical methods for conservation laws.* Birkhäuser, Basel, 1990.

[Ley08]   D. Leykekhman. Uniform error estimates in the finite element method for a singularly perturbed reaction-diffusion problem. *Math. Comp.*, 77:21–39 (electronic), 2008.

[LG97]   J.-Y. Lee and L. Greengard. A fast adaptive numerical method for stiff two-point boundary value problems. *SIAM J. Sci. Comput.*, 18(2):403–429, 1997.

[Li01]   J. Li. Convergence and superconvergence analysis of finite element methods on highly nonuniform anisotropic meshes for singularly perturbed reaction-diffusion problems. *Appl. Numer. Math.*, 36:129–154, 2001.

[Lin91]   Q. Lin. A rectangle test for finite element analysis. In *Proc. Syst. Sci. Eng.*, pages 213–216. Great Wall (H.K.) Culture Publish Co., 1991.

[Lin99]   T. Linß. An upwind difference scheme on a novel Shishkin-type mesh for a linear convection-diffusion problem. *J. Comput. Appl. Math.*, 110:93–104, 1999.

[Lin00a]   T. Linß. Analysis of a Galerkin finite element method on a Bakhvalov-Shishkin mesh for a linear convection-diffusion problem. *IMA J. Numer. Anal.*, 20:621–632, 2000.

[Lin00b]   T. Linß. Uniform superconvergence of a Galerkin finite element method on Shishkin-type meshes. *Numer. Methods Partial Differential Equations*, 16:426–440, 2000.

[Lin01a]   T. Linß. Sufficient conditions for uniform convergence on layer-adapted grids. *Appl. Numer. Math.*, 37:241–255, 2001.

[Lin01b]   T. Linß. Uniform pointwise convergence of finite difference schemes using grid equidistribution. *Computing*, 66:27–39, 2001.

[Lin01c]   T. Linß. Uniform second-order pointwise convergence of a finite difference discretisation for a quasilinear problem. *Comput. Math. Math. Phys.*, 41:898–909, 2001.

[Lin02a]   T. Linß. Analysis of an upwind difference scheme on arbitrary meshes for convection-diffusion problems. *GAMM Mitt. Ges. Angew. Math. Mech.*, 25(1-2):47–86, 2002.

[Lin02b]   T. Linß. Solution decomposition for linear convection-diffusion problems. *Z. Anal. Anwendungen*, 21:209–214, 2002.

[Lin03a]   T. Linß. Layer-adapted meshes for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 192:1061–1105, 2003.

[Lin03b]   T. Linß. Robustness of an upwind finite difference scheme for semilinear convection-diffusion problems with boundary turning points. *J. Comput. Math.*, 21:401–410, 2003.

[Lin04]   T. Linß. Error expansion for a first-order upwind difference scheme applied to a model convection-diffusion problem. *IMA J. Numer. Anal.*, 24(2):239–253, 2004.

[Lin07a]   T. Linß. Analysis of a system of singularly perturbed convection-diffusion equations with strong coupling. Technical Report MATH-NM-02-2007, Technical University of Dresden, 2007.

[Lin07b]   T. Linß. Analysis of an upwind finite-difference scheme for a system of coupled singularly perturbed convection-diffusion equations. *Computing*, 79:23–32, 2007.

[Lin08a]   T. Linß. Analysis of a fem for a coupled system of singularly perturbed reaction-diffusion equations. Technical Report MATH-NM-04-2008, Technical University of Dresden, 2008.

[Lin08b]   T. Linß. A robust method for a time-dependent system of coupled singularly perturbed reaction-diffusion problems. Technical Report MATH-NM-03-2008, Technical University of Dresden, 2008.

[Lio73]   J. L. Lions. *Perturbations singuliéres dans les problémes aux limite et en contrôle optimal.* Springer-Verlag, Berlin, 1973.

[Lis83]   V. D. Liseĭkin. On the numerical solution of a two-dimensional elliptic equation with a small parameter in the highest derivatives (in Russian). *Num. Meth. Mech. Cont. Media*, 14:110–115, 1983.

[Lis90]   V. D. Liseĭkin. Application of special transformations for numerical solution of problems with boundary layers. *Comput. Math. and Math. Phys.*, 30:43–53, 1990.

[Lis99]   V. D. Liseĭkin. *Grid generation methods.* Scientific Computation. Springer-Verlag, Berlin, 1999.

[LL06]   Q. Lin and J. Lin. *Finite element methods: accuracy and improvement.* Science Press, Beijing, 2006.

[LM]   T. Linß and N. Madden. Layer-adapted meshes for a system of coupled singularly perturbed reaction-diffusion problems. *IMA J. Numer. Anal.* (to appear).

[LM03]   T. Linß and N. Madden. An improved error estimate for a numerical method for a system of coupled singularly perturbed reaction-diffusion equations. *Comput. Methods Appl. Math.*, 3(3):417–423, 2003.

[LM07]   T. Linß and N. Madden. Parameter uniform approximations for time-dependent reaction-diffusion problems. *Numer. Methods Partial Differential Equations*, 23:1290–1300, 2007.

[LMSZ08]   F. Liu, N. Madden, M. Stynes, and A. Zhou. A two-scale sparse grid method for a singularly perturbed reaction-diffusion problem in two dimensions. Technical report, National University of Ireland, Cork, 2008. (submitted for publication).

[LMV96]   R. D. Lazarov, I. D. Mishev, and P. S. Vassilevski. Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.*, 33(1):31–55, 1996.

[LOM98]   G. Lube, F. C. Otto, and H. Müller. A non-overlapping domain decom-
          position method for parabolic initial-boundary value problems. *Appl.
          Numer. Math.*, 28:359–369, 1998.

[Lor75]   J. Lorenz. *Die inverse Monotonie von Matrizen und ihre Anwendung
          beim Stabilitätsnachweis von Differenzenverfahren.* PhD thesis, Univ.
          Münster, 1975.

[Lor81]   J. Lorenz. Stability and consistency analysis of difference methods for
          singular perturbation problems. In O. Axelsson, editor, *Analytical and
          numerical approaches to asymptotic problems*, pages 141–156. North Hol-
          land, Amsterdam, 1981.

[Lor82]   J. Lorenz. Nonlinear boundary value problems with turning points and
          properties of difference schemes. In *Theory and applications of singu-
          lar perturbations (Oberwolfach, 1981)*, volume 942 of *Lecture Notes in
          Math.*, pages 150–169. Springer, Berlin, 1982.

[Lor84]   J. Lorenz. Analysis of difference schemes for a stationary shock problem.
          SIAM J. Numer. Anal., 21:1038–1053, 1984.

[LP89]   W. D. Liseĭkin and W. E. Petrenko. An adaptive-invariant method for
          the numerical solution of problems with boundary and interior layers (in
          Russian). Computer Center, Academy of Sci., Novosibirsk, 1989.

[LR74]   P. Lesaint and P. A. Raviart. On a finite element method for solving
          the neutron transport equation. In C. deBoor, editor, *Mathematical
          aspects of finite elements in partial differential equations*, pages 89–123.
          Academic Press, New York, 1974.

[LR80]   R. E. Lynch and J. R. Rice. A high order difference method for differ-
          ential equations. *Math. Comp.*, 34:333–372, 1980.

[LR95]   W. J. Layton and P. J. Rabier. Peaceman-Rachford procedure and do-
          main decomposition for finite element problems. *Numer. Linear Algebra
          Appl.*, 2:363–393, 1995.

[LR04]   T. Linß and H.-G. Roos. Analysis of a finite difference scheme for a
          singularly perturbed problem with two small parameters. *J. Math. Anal.
          Appl.*, 289:355–366, 2004.

[LR06]   G. Lube and G. Rapin. Residual-based stabilized higher-order FEM for a
          generalized Oseen problem. *Math. Models Methods Appl. Sc.*, 16(7):949–
          966, 2006.

[LRV00]   T. Linß, H.-G. Roos, and R. Vulanović. Uniform pointwise convergence
          on Shishkin-type meshes for quasi-linear convection-diffusion problems.
          SIAM J. Numer. Anal., 38:897–912, 2000.

[LS99]   T. Linß and M. Stynes. A hybrid difference scheme on a Shishkin mesh
          for linear convection-diffusion problems. *Appl. Num. Math.*, 31:255–270,
          1999.

[LS01a]   T. Linß and M. Stynes. Asymptotic analysis and Shishkin-type decom-
          position for an elliptic convection-diffusion problem. *J. Math. Anal.
          Applic.*, 261:604–632, 2001.

[LS01b]   T. Linß and M. Stynes. Numerical methods on Shishkin meshes for linear
          convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*,
          190:3527–3542, 2001.

[LS01c]   T. Linß and M. Stynes. The SDFEM on Shishkin meshes for linear
          convection-diffusion problems. *Numer. Math.*, 87:457–484, 2001.

[LSU67]   O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural′ceva. *Linear and quasilinear equations of parabolic type*. Transl. of Mathem. Monographs, Vol. 23. A.M.S., Providence, R.I., 1967.

[LT90]   G. Lube and L. Tobiska. A nonconforming finite element method of streamline-diffusion type for the incompressible Navier-Stokes equations. *J. Comp. Math.*, 8(2):147–158, 1990.

[LU68]   O. A. Ladyzhenskaya and N. N. Ural′tseva. *Linear and quasilinear elliptic equations*. Translated from the Russian by Scripta Technica, Inc. Academic Press, New York, 1968.

[Lu95]   H. Lu. A uniform consistency barrier on finite-difference schemes of positive type for convection-diffusion equations. *SIAM J. Sci. Comput.*, 16:169–172, 1995.

[Lub94]   G. Lube. Stabilized Galerkin finite element methods for convection dominated and incompressible flow problems. In *Numerical Analysis and Mathematical Modelling*, Banach Center Publications 29. Inst. of Math., Polish Academy of Sci., Warszawa, 1994.

[Lub06]   G. Lube. Stabilized FEM for incompressible flow. Critical review and new trends. In P. Wesseling, E. Onate, and J. Périaux, editors, *ECCO-MAS CFD 2006*, pages 1–20. TU Delft, 2006.

[LV01]   T. Linß and R. Vulanović. Uniform methods for semilinear problems with an attractive boundary turning point. *Novi Sad J. Math.*, 31:99–114, 2001.

[LW95]   G. Lube and D. Weiss. Stabilized finite element methods for singularly perturbed parabolic problems. *Appl. Numer. Math.*, 17:431–459, 1995.

[LW00]   J. Li and M. F. Wheeler. Uniform convergence and superconvergence of mixed finite element methods on anisotropically refined grids. *SIAM J. Numer. Anal.*, 38:770–798, 2000.

[LX05]   S.-T. Liu and Y. Xu. Graded Galerkin methods for the high order convection-diffusion problem. Technical report, Syracuse University, 2005.

[LX06]   S.-T. Liu and Y. Xu. Galerkin methods based on Hermite splines for singular perturbation problems. *SIAM J. Numer. Anal.*, 43:2607–2623, 2006.

[LY96]   Q. Lin and N. Yan. *Construction and Analysis of High Efficient Finite Elements*. Hebei University Press, P.R.China, 1996. (In Chinese).

[Mac99]   J. Mackenzie. Uniform convergence analysis of an upwind finite-difference approximation of a convection-diffusion boundary value problem on an adaptive grid. *IMA J. Numer. Anal.*, 19(2):233–249, 1999.

[Mar77]   G. I. Marchuk. *Methods of Numerical Mathematics*. Springer-Verlag, Berlin, 1977.

[Mat]   G. Matthies. Local projection stabilisation for higher order discretisations of convection-diffusion problems on Shishkin meshes. *Adv. Comput. Math.* (to appear).

[Mat01]   G. Matthies. Mapped finite elements on hexahedra. Necessary and sufficient conditions for optimal interpolation errors. *Numer. Algorithms*, 27(4):317–327, 2001.

[Mat02]   P. Matus. The maximum principle and some of its applications. *Comput. Methods Appl. Math.*, 2:50–91, 2002.

[Mat07] G. Matthies. Inf-sup stable nonconforming finite elements of higher order on quadrilaterals and hexahedra. *M2AN Math. Model. Numer. Anal.*, 41:713–742, 2007.

[MB01]   K. Miller and M. J. Baines. Least squares moving finite elements. *IMA J. Numer. Anal.*, 21:621–642, 2001.

[Mel00] J. M. Melenk. On *n*-widths for elliptic problems. *J. Math. Anal. Appl.*, 247:272–289, 2000.

[Mel02] J. M. Melenk. *hp-Finite Element Methods for Singular Perturbations*. Springer, Heidelberg, 2002.

[Mey98] K.-H. Meyn. Monotonieaussagen für elliptische und parabolische Randwertaufgaben und Anwendungen auf Finite-Elemente-Funktionen. PhD thesis, Univ. Hamburg, 1998.

[MG79]   A. R. Mitchell and D. F. Griffiths. Semi-discrete generalised Galerkin methods for time-dependent conduction-convection problems. In J.R. Whiteman, editor, *MAFELAP III*, pages 19–34. Academic Press, New York, 1979.

[MG80]   A. R. Mitchell and D. F. Griffiths. Upwinding by Petrov-Galerkin methods in convection-diffusion problems. *J. Comp. Appl. Math.*, 6:219–228, 1980.

[MH85]   A. Mizukami and T. J. R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle. *Comput. Methods Appl. Mech. Eng.*, 50:181–193, 1985.

[Mic77] J. H. Michael. A general theory for linear elliptic partial differential equations. *J. diff. equations*, 23:1–29, 1977.

[Mil76] J. J. H. Miller. Construction of a fem for a singularly perturbed problem in 2 dimensions. In *Numerische Behandlung von Differentialgleichungen, Band 2 (Tagung, Math. Forschungsinst., Oberwolfach, 1975)*, pages 165–169. Internat. Ser. Numer. Math., Vol. 31. Birkhäuser, Basel, 1976.

[Mil83] K. Miller. Alternate modes to control the nodes in the moving finite element method. In I. Babuška, J. Chandra, and J. Flaherty, editors, *Adaptive computational methods for partial differential equations*, pages 165–182. SIAM, Philadephia, 1983.

[ML07]   G. Matthies and G. Lube. On streamline-diffusion methods of inf-sup stable discretisations of the generalised Oseen problem. Preprint 2007-02, Institut für Numerische und Angewandte Mathematik, Univ. Göttingen, 2007.

[MM07]   J. A. Mackenzie and W. R. Mekwi. An analysis of stability and convergence of a finite-difference discretization of a model parabolic PDE in 1D using a moving mesh. *IMA J. Numer. Anal.*, 27:507–528, 2007.

[MMS92] K. W. Morton, T. Murdoch, and E. Süli. Optimal error estimation for Petrov-Galerkin methods in two dimensions. *Numer. Math.*, 61:359–372, 1992.

[Mor80] K. W. Morton. Stability of finite difference approximations to a diffusion-convection equation. *Int. J. Numer. Methods Eng.*, 15:677–683, 1980.

[Mor96] K. W. Morton. *Numerical Solution of Convection-diffusion Problems*. Chapman & Hall, London, 1996.

[MOS96] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. *Fitted numerical methods for singular perturbation problems.* World Scientific Publishing Co. Inc., River Edge, NJ, 1996.

[MOSS98] J. J. H. Miller, E. O'Riordan, G. I. Shishkin, and L. P. Shishkina. Fitted mesh methods for problems with parabolic boundary layers. *Math. Proc. R. Ir. Acad.*, 98A:173–190, 1998.

[MQS97] L. S. Mulholland, Y. Qiu, and D. M. Sloan. Solution of evolutionary partial differential equations using adaptive finite differences with pseudospectral post-processing. *J. Comput. Phys.*, 131:280–298, 1997.

[MRS90] P. A. Markowich, C. Ringhofer, and S. Schmeiser. *Semiconductor Equations.* Springer, Vienna, 1990.

[MS85] K. W. Morton and B. W. Scotney. Petrov-Galerkin methods and diffusion-convection problems in 2D. In J.R. Whiteman, editor, *MAFELAP V*, pages 343–366. Academic Press, New York, 1985.

[MS93] K. W. Morton and I. J. Sobey. Discretization of a convection-diffusion equation. *IMA J. Numer. Anal.*, 13:141–160, 1993.

[MS96] N. Madden and M. Stynes. Linear enhancements of the streamline diffusion method for convection-diffusion problems. *Computers Math. Applic.*, 32:29–42, 1996.

[MS99a] J. M. Melenk and C. Schwab. An *hp* finite element method for convection-diffusion problems in one dimension. *IMA J. Numer. Anal.*, 19:425–453, 1999.

[MS99b] J. M. Melenk and C. Schwab. The *hp* streamline diffusion finite element method for convection dominated problems in one space dimension. *East-West J. Numer. Math.*, 7:31–60, 1999.

[MS03] N. Madden and M. Stynes. A uniformly convergent numerical method for a coupled system of two singularly perturbed linear reaction-diffusion problems. *IMA J. Numer. Anal.*, 23:627–644, 2003.

[MS07] G. Matthies and F. Schieweck. On the reference mapping for quadrilateral and hexahedral finite elements on multilevel adaptive grids. *Computing*, 80(2):95–119, 2007.

[MST07] G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.*, 41(4):713–742, 2007.

[MT02] G. Matthies and L. Tobiska. The inf-sup condition for the mapped $Q_k - P_{k-1}^{disc}$ element in arbitrary space dimension. *Computing*, 69(2):119–139, 2002.

[MT05] G. Matthies and L. Tobiska. Inf-sup stable non-conforming finite elements of arbitrary order on triangles. *Numer. Math.*, 102:293–309, 2005.

[MT07] G. Matthies and L. Tobiska. Local projection type stabilisation applied to inf-sup stable discretisations of the Oseen problem. Preprint, Fakultät für Mathematik, Univ. Magdeburg, 2007.

[MW94] J. J. H. Miller and S. Wang. A new non-conforming Petrov-Galerkin finite element method with triangular elements for an advection-diffusion problem. *IMA J. Numer. Anal.*, 14:257–276, 1994.

[Näv82] U. Nävert. *A finite element method for convection-diffusion problems.* PhD thesis, Chalmers University of Technology, Göteborg, 1982.

[NH00] J. Noordmans and P. W. Hemker. Application of an adaptive sparsegrid technique to a model singular perturbation problem. *Computing*, 65:357–378, 2000.

584   References

[Nii84]   K. Niijima. A uniformly convergent difference scheme for a semilinear singular perturbation problem. *Numer. Math.*, 43:175–198, 1984.

[Nii86]   K. Niijima. An error analysis for a difference scheme of exponential type applied to a nonlinear singular perturbation problem without turning points. *J. Comput. Appl. Math.*, 15:93–101, 1986.

[Nii90]   K. Niijima. Pointwise error estimates for a streamline diffusion finite element scheme. *Numer. Math.*, 56:707–719, 1990.

[NKS08]   A. Naughton, R. B. Kellogg, and M. Stynes. Regularity and derivative bounds for a convection-diffusion problem with a Neumann outflow condition. Preprint, National University of Ireland Cork, 2008.

[Noc95]   R. H. Nochetto. Pointwise a posteriori error estimates for elliptic problems on highly graded meshes. *Math. Comp.*, 64:1–22, 1995.

[NS03]   M. C. Natividad and M. Stynes. Richardson extrapolation for a convection-diffusion problem using a Shishkin mesh. *Appl. Numer. Math.*, 45(2-3):315–329, 2003.

[NSOS88]   M. J. Ng-Stynes, E. O'Riordan, and M. Stynes. Numerical methods for time-dependent convection-diffusion problems. *J. Comput. Appl. Math.*, 21:289–310, 1988.

[NY83]   K. Niederdrenk and H. Yserentant. The uniform stability of singularly perturbed discrete and continuous boundary value problems. *Numer. Math.*, 41:223–253, 1983.

[O'M91]   R. E. O'Malley. *Singular perturbation methods for ordinary differential equations*. Springer-Verlag, Berlin, 1991.

[OR70]   J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, 1970.

[OR73]   O. A. Oleĭnik and E. V. Radkevič. *Second order equations with nonnegative characteristic form*. Plenum Press, New York, 1973. Translated from the Russian by Paul C. Fife.

[OR79]   L. A. Oganesjan and L. A. Ruhovec. *Variatsionno-raznostnye metody resheniya ellipticheskikh uravnenii*. Akad. Nauk Armyan. SSR, Erevan, 1979.

[O'R84]   E. O'Riordan. Singular perturbation finite element methods. *Numer. Math.*, 44:425–434, 1984.

[O'R86]   M. O'Reilly. Exponential fitting applied to quasi-linear two-point boundary value problems. In *Proc. of BAIL IV*, pages 387–391, Dublin, 1986. Boole Press.

[OS]   E. O'Riordan and M. Stynes. Numerical analysis of a strongly coupled system of two singularly perturbed convection-diffusion problems. *Adv. Comput. Math.* (to appear).

[OS86]   E. O'Riordan and M. Stynes. An analysis of a superconvergence result for a singularly perturbed boundary value problem. *Math. Comp.*, 46:81–92, 1986.

[OS89]   E. O'Riordan and M. Stynes. A uniformly convergent difference scheme for an elliptic singular perturbation problem. In L. Tobiska, editor, *Discretization Methods of Singular Perturbations and Flow Problems*, pages 157–168. Univ. Magdeburg, 1989.

[OS91a]   E. O'Riordan and M. Stynes. An analysis of some exponentially fitted finite element methods for singularly perturbed elliptic problems. In J. J. H. Miller, editor, *Computational methods for boundary and interior layers in several dimensions*, pages 138–153. Boole Press, Dublin, 1991.

[OS91b]   E. O'Riordan and M. Stynes. A globally uniformly convergent finite element method for a singularly perturbed elliptic problem in two dimensions. *Math. Comp.*, 57:47–62, 1991.

[OS07a]   E. O'Riordan and G. I. Shishkin. Parameter uniform numerical methods for singularly perturbed elliptic problems with parabolic boundary layers. *Appl. Numer. Math.*, 2007. Published online November 22.

[OS07b]   E. O'Riordan and G. I. Shishkin. A technique to prove paramater-uniform convergence for a singularly perturbed convection-diffusion equation. *J. Comput. Appl. Math.*, 206:136–145, 2007.

[Osh81]   S. Osher. Nonlinear singular perturbation problems and one sided difference schemes. *SIAM J. Numer. Anal.*, 18:129–144, 1981.

[OSS]   E. O'Riordan, J. Stynes, and M. Stynes. A parameter-uniform finite difference method for a coupled system of convection-diffusion two-point boundary value problems. *Numer. Math. Theor. Meth. Appl.* (to appear).

[Osw91]   P. Oswald. On a BPX-preconditioner for P_1 elements. Technical report, FSU Jena, 1991.

[OU84]   K. Ohmori and T. Ushijima. A technique of upstream type applied to a linear nonconforming finite element approximation of convective diffusion equations. *RAIRO Numer. Anal.*, 18:309–332, 1984.

[Pac93]   J. Pach. *New Trends in Discrete and Computational Geometry.* Springer-Verlag, 1993.

[Pet91]   T. E. Peterson. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation. *SIAM J. Numer. Anal.*, 28:133–140, 1991.

[PHSM87]   S. Polak, C. Den Heijer, W. H. Schilders, and P. Markowich. Semiconductor device modelling from the numerical point of view. *Int. J. Numer. Methods Eng.*, 24:763–838, 1987.

[Pic03]   M. Picasso. An anisotropic error indicator based on the Z-Z error estimator: application to elliptic and parabolic problems. *SIAM J. Sci. Comp.*, 24:1328–1355, 2003.

[Pir89]   O. Pironneau. *Finite element methods for fluids.* John Wiley & Sons, Chichester, 1989.

[PR55]   D. W. Peaceman and H. H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3:28–41, 1955.

[Pri94]   A. Priestley. The positive and nearly conservative Lagrange-Galerkin method. *IMA J. Numer. Anal.*, 14:277–294, 1994.

[PW67]   M. H. Protter and H. F. Weinberger. *Maximum principles in differential equations.* Prentice-Hall, Englewood Cliffs, 1967.

[Qin94]   J. Qin. *On the convergence of some low order mixed finite elements for incompressible fluids.* PhD thesis, Pennsylvania State University, 1994.

[QS99]   Y. Qiu and D. M. Sloan. Analysis of difference approximations to a singularly perturbed two-point boundary value problem on an adaptively generated grid. *J. Comput. Appl. Math.*, 101(1-2):1–25, 1999.

[QST00]   Y. Qiu, D. M. Sloan, and T. Tang. Numerical solution of a singularly perturbed two-point boundary value problem using equidistribution: analysis of convergence. *J. Comput. Appl. Math.*, 116(1):121–143, 2000.

[RAF96]   H.-G. Roos, D. Adam, and A. Felgenhauer. A novel nonconforming
           uniformly convergent finite element methods in two dimensions. *J. Math.
           Anal. Appl.*, 201(3):715–755, 1996.

[Ran94]   R. Rannacher. Domain decomposition in the nonstationary streamline
           diffusion finite element method. In *Finite element methods (Jyväskylä,
           1993)*, volume 164 of *Lecture Notes in Pure and Appl. Math.*, pages
           367–380. Dekker, New York, 1994.

[Ran98]   R. Rannacher. A posteriori error estimation in least-squares stabilized
           finite element schemes. *Comput. Methods Appl. Mech. Engrg.*, 166:99–
           114, 1998.

[Ran00]   R. Rannacher. Finite element methods for the incompressible Navier-
           Stokes equations. In G.P. Galdi, J. Heywood, and R. Ranancher, editors,
           Fundamental Directions in Mathematical Fluid Mechanics, pages 191–
           293. Birkhäuser, Basel, Boston, Berlin, 2000.

[Ran04]   R. Rannacher. Incompressible viscous flows. In E. Stein, R. de Borst,
           and T.J.R. Hughes, editors, *Encyclopedia of Computational Mechanics*,
           volume 3, pages 155–181. John Wiley & Sons, 2004.

[Rau72]   J. Rauch. L_2 is a continuable initial condition for Kreiss' mixed prob-
           lems. *Comm. Pure Appl. Math.*, 25:265–285, 1972.

[REI$^+$07] A. Rap, L. Elliott, D. B. Ingham, D. Lesnic, and X. Wen. The inverse
           source problem for the variable coefficients convection-diffusion equa-
           tion. *Inverse Probl. Sci. Eng.*, 15:413–440, 2007.

[RF08]   V. Ramakkagari and J. E. Flaherty. A new stable method for singularly
           perturbed convection-diffusion equations. *Comput. Methods Appl. Mech.
           Engrg.*, 197:1507–1524, 2008.

[RH73]   W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron
           transport equation. Technical Report LA-UR-73-479. Los Alamos, 1973.

[Ric88]   G. R. Richter. An optimal order error estimate for the discontinuous
           Galerkin method. *Math. Comp.*, 50:75–88, 1988.

[Ric90]   G. R. Richter. A finite element method for time-dependent convection-
           diffusion equations. *Math. Comp.*, 54:81–106, 1990.

[Ric92]   G. R. Richter. The discontinuous Galerkin method with diffusion. *Math.
           Comp.*, 58(198):631–643, 1992.

[Ris86]   U. Risch. *Ein hybrides upwind-FEM-Verfahren und dessen Anwen-
           dung auf schwach gekoppelte elliptische Differentialgleichungssysteme
           mit dominanter Konvektion.* PhD thesis, TH Magdeburg, 1986.

[Ris90]   U. Risch. An upwind finite element method for singularly perturbed
           elliptic problems and local estimates in the L$^\infty$-norm. *Math. Modél.
           Anal. Numér.*, 24(2):235–264, 1990.

[Ris01]   U. Risch. Convergence analysis of the residual free bubble method for
           bilinear elements. *SIAM J. Numer. Anal.*, 39(4):1366–1379, 2001.

[RL99]   H.-G. Roos and T. Linß. Sufficient conditions for uniform convergence
           on layer-adapted grids. *Computing*, 63:27–45, 1999.

[RL01a]   H.-G. Roos and T. Linß. Gradient recovery for singularly perturbed
           boundary value problems. I. One-dimensional convection-diffusion.
           Computing, 66:163–178, 2001.

[RL01b]   H.-G. Roos and T. Linß. Gradient recovery for singularly perturbed
           boundary value problems. II. Two-dimensional convection-diffusion.
           Math. Models Methods Appl. Sci., 11:1169–1179, 2001.

[RM94]   R. D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems*. Robert E. Krieger Publishing Co. Inc., Malabar, FL, 2nd edition, 1994.

[Roo85] H.-G. Roos. Necessary convergence conditions for upwind schemes in the two-dimensional case. *Int. J. Numer. Methods Eng.*, 21:1459–1469, 1985.

[Roo86a] H.-G. Roos. Beziehungen zwischen Diskretisierungsverfahren für Konvektions-Diffusions-Gleichungen und für die Grundgleichungen der inneren Elektronik. Report Math. 07-10-86, TU Dresden, 1986.

[Roo86b] H.-G. Roos. Second order monotone upwind schemes. *Computing*, 36:57–67, 1986.

[Roo96] H.-G. Roos. A note on the conditioning of upwind schemes on Shishkin meshes. *IMA J. Numer. Anal.*, 16:529–538, 1996.

[Roo02] H.-G. Roos. Optimal convergence of basic schemes for elliptic boundray value problems with strong parabolic layers. *J. Math. Anal. Appl.*, 267:194–208, 2002.

[RS89]   U. Risch and F. Schieweck. A multigrid method for solving the stationary Navier-Stokes equations by fem. In *Third MG Seminar*, volume 89 of *Rep. MATH.*, pages 74–87. Akad. Wiss. DDR, Berlin, 1989.

[RS92]   P. L. Roe and D. Sidilkover. Optimum positive linear schemes for advection in two and three dimensions. *SIAM J. Numer. Anal.*, 29:1542–1568, 1992.

[RS96]   H.-G. Roos and M. Stynes. Necessary conditions for uniform convergence of finite difference schemes for convection-diffusion problems with exponential and parabolic layers. *Applications of Mathematics*, 41:269–280, 1996.

[RST96] H.-G. Roos, M. Stynes, and L. Tobiska. *Numerical methods for singularly perturbed differential equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1996. Convection-diffusion and flow problems.

[RSW08] T. F. Russell, M. Stynes, and H. Wang. Error estimates for the FVEL-LAM for time-dependent advection-diffusion problems. Technical report, National University of Ireland, Cork, 2008. (In preparation).

[RU03]   H.-G. Roos and Z. Uzelac. The SDFEM for a convection-diffusion problem with two small parameters. *Comp. Meth. Appl. Math.*, 3:443–458, 2003.

[Rus06] A. Russo. Streamline-upwind Petrov/Galerkin method (SUPG) vs. residual free bubbles. *Comput. Methods Appl. Mech. Engrg.*, 195:1608–1620, 2006.

[RV93]   H.-G. Roos and R. Vulanović. A higher order uniform convergence result for a turning point problem. *Zeitschrift f. Analysis u. Anwendungen*, 12:723–728, 1993.

[RV07]   H.-G. Roos and R. Vanselow. A comparison of four- and five-point difference approximations for stabilizing the one-dimensional stationary convection-diffusion equation. Preprint MATH NM-06-07, 2007.

[RW03]   B. Riviere and M. Wheeler. A posteriori error estimates for a discontinuous Galerkin method applied to elliptic problems. *Comput. Math. Appl.*, 46:141–163, 2003.

[RZ03]   H.-G. Roos and H. Zarin. The discontinuous Galerkin finite element method for singularly perturbed problems. In E. Bänsch, editor, *CISC 2002*, volume 35 of *Lecture Notes in Comput. Sci. and Engng.*, pages 246–267. Springer, Berlin, 2003.

[RZ07]   H.-G. Roos and H. Zarin. A supercloseness result for the discontinuous Galerkin stabilization of convection-diffusion problems on Shishkin meshes. *Numer. Methods Partial Differential Equations*, 23:1560–1576, 2007.

[Sal98] R. Salmon. *Lectures on Geophysical Fluid Dynamics.* Oxford University Press, New York, 1998.

[San00] G. Sangalli. Global and local analysis for the residual-free bubbles method aplied to advection-dominated problems. *SIAM J. Numer. Anal.*, 38(5):1496–1522, 2000.

[San01] G. Sangalli. A robust a posteriori estimate for the residual free bubbles method aplied to advection-dominated problems. *Numer. Math.*, 89:379–399, 2001.

[San05] G. Sangalli. A uniform analysis of nonsymmetric and coercive linear operators. *SIAM J. Math. Anal.*, 36(6):2033–2048, 2005.

[San08] G. Sangalli. Robust a-posteriori estimator for advection-diffusion-reaction problems. *Math. Comp.*, 77(261):41–70 (electronic), 2008.

[Sav95] I. A. Savin. On the uniform convergence with respect to a small parameter of difference schemes for an ordinary differential equation. *Zh. Vychisl. Mat. i Mat. Fis.*, 35:1758–1765, 1995.

[SB84]   W. G. Szymczak and I. Babuška. Adaptivity and error estimation for the finite element method applied to convection diffusion problems. *SIAM J. Numer. Anal.*, 21(5):910–954, 1984.

[Sch84] F. Schieweck. Finite element methods for singularly perturbed partial differential equations. In J. Förste, editor, *Singularly perturbed differential equations and applications*, number 03/84 in R-MECH, pages 37–45. Akademie der Wissenschaften der DDR, 1984.

[Sch86] F. Schieweck. *Eine asymptotisch angepaßte Finite-Element-Methode für singulär gestörte elliptische Randwertaufgaben.* PhD thesis, TH Magdeburg, 1986.

[Sch87] F. Schieweck. Numerische Integration bei der Finite-Element-Diskretisierung singulär gestörter elliptischer Randwertaufgaben. *Wiss. Z. Techn. Universität Magdeburg*, 31:95–101, 1987.

[Sch94] F. Schieweck. On the order of two nonconforming finite element approximations of upwind type for the Navier-Stokes equations. In F.-K. Hebeker, R. Rannacher, and G. Wittum, editors, *Numerical methods for the Navier-Stokes equations. Proc.,*, pages 249–258, Heidelberg, 1994. Vieweg-Verlag.

[Sch00] F. Schieweck. A general transfer operator for arbitrary finite element spaces. Technical report, Institute for Analysis and Computational Mathematics, Univ. Magdeburg, 2000.

[Sch07] F. Schieweck. On the role of boundary conditions for the CIP stabilization of higher order finite elements. Technical Report 41, Institute for Analysis and Computational Mathematics, Univ. Magdeburg, 2007.

[Sch08] F. Schieweck. The stability of the CIP method for higher order finite elements applied to convection-diffusion equations. Technical report,

Institute for Analysis and Computational Mathematics, Univ. Magdeburg, 2008.

[SE99]   Y.-T. Shih and H. C. Elman. Modified streamline diffusion schemes for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 174(1-2):137–151, 1999.

[SE00]   Y.-T. Shih and H. C. Elman. Iterative methods for stabilized discrete convection-diffusion problems. *IMA J. Numer. Anal.*, 20:333–358, 2000.

[Seg82]   A. Segal. Aspects of numerical methods for elliptic singular perturbation problems. *SIAM J. Sci. Statist. Comput.*, 3:327–349, 1982.

[Sel84]   S. Selberherr. *Analysis and simulation of semiconductor devices*. Springer-Verlag, Wien–New-York, 1984.

[SGG99]   R. Sacco, E. Gatti, and L. Gotusso. A nonconforming exponentially fitted finite element method for two-dimensional drift-diffusion models in semiconductors. *Numer. Methods Partial Differential Equations*, 15:133–150, 1999.

[Shi]   G. I. Shishkin. *Difference Methods for Singular Perturbation Problems*. CRC Press, Boca Raton. (to appear).

[Shi83]   G. I. Shishkin. A difference scheme on a non-uniform mesh for a differential equation with a small parameter in the highest derivative. *U.S.S.R. Comput. Maths. Math. Phys.*, 23:59–66, 1983.

[Shi84]   G. I. Shishkin. A method of improving the accuracy of the solution of difference schemes for parabolic equations with a small parameter in the highest derivative. *U.S.S.R. Comput. Maths. Math. Phys.*, 24:150–157, 1984.

[Shi86]   G. I. Shishkin. Solution of a boundary value problem for an elliptic equation with a small parameter for the leading derivatives. *U.S.S.R. Comput. Maths. Math. Phys.*, 26:38–46, 1986.

[Shi88]   G. I. Shishkin. Grid approximation of singularly perturbed parabolic equations with internal layers. *Soviet J. Numer. Anal. Math. Modelling*, 3:393–407, 1988.

[Shi89]   G. I. Shishkin. Approximation of the solutions of singularly perturbed boundary-value problems with a parabolic boundary layer. *U.S.S.R. Comput. Maths. Math. Physics*, 29:1–10, 1989.

[Shi90a]   G. I. Shishkin. Grid approximation of singularly perturbed boundary value problems with convective terms. *Sov. J. Numer. Anal. Math. Modelling*, 5:173–187, 1990.

[Shi90b]   G. I. Shishkin. Grid approximation of singularly perturbed elliptic and parabolic equations (in Russian). Second Doctoral thesis. Keldysh Institute Moscow, 1990.

[Shi92a]   G. I. Shishkin. Difference approximation of a singularly perturbed boundary value problem for quasilinear elliptic equations that degenerate into a first-order equation. *Comput. Math. Math. Phys.*, 32:467–480, 1992.

[Shi92b]   G. I. Shishkin. *Discrete approximation of singularly perturbed elliptic and parabolic equations* (in Russian). Russian Academy of Sciences, Ural Section, Ekaterinburg, 1992.

[Shi92c]   G. I. Shishkin. Methods of constructing grid approximations for singularly perturbed boundary-value problems. condensing grid methods. *Russ. J. Numer. Anal. Math. Modelling*, 7:537–562, 1992.

[Shi93]   G. I. Shishkin. Method of splitting for singularly perturbed parabolic equations. *East-West J. Numer. Math.*, 1:147–163, 1993.

[Shi96a]   S.-D. Shih. A novel uniform expansion for a singularly perturbed parabolic problem with corner singularity. *Methods Appl. Anal.*, 3:203–227, 1996.

[Shi96b]   G. I. Shishkin. Method of improving the accuracy of the approximate solutions to singularly perturbed equations by defect correction. *Russian J. Numer. Anal. Math. Modelling*, 11:539–557, 1996.

[Shi97a]   G. I. Shishkin. Acceleration of the process of the numerical solution to singularly perturbed boundary value problems for parabolic equations on the basis of parallel computations. *Russian J. Numer. Anal. Math. Modelling*, 12:271–291, 1997.

[Shi97b]   G. I. Shishkin. On finite difference fitted schemes for singularly perturbed boundary value problems with a parabolic boundary layer. *J. Math. Anal. Appl.*, 208:181–204, 1997.

[Shi98a]   G. I. Shishkin. Grid approximation of parabolic equations with small parameters multiplying the space and time derivatives. Reaction-diffusion equations. *Math. Balkanica (N.S.)*, 12:179–214, 1998.

[Shi98b]   G. I. Shishkin. Grid approximation of singularly perturbed systems of elliptic and parabolic equations with convective terms. *Differential Equations*, 34:1693–1704, 1998.

[Shi98c]   G. I. Shishkin. Grid approximations of singularly perturbed systems for parabolic convection-diffusion equations with counterflow. *Sib. Zh. Vychisl. Mat.*, 1:281–297, 1998.

[Shi00]   G. I. Shishkin. Approximation of singularly perturbed convection-diffusion problems with low smoothness of the derivatives involved in the equation. In L. G. Vulkov, J. J. H. Miller, and G. I. Shishkin, editors, *Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems*, pages 111–121, New York, 2000. Nova Science Publishers.

[Shi01]   S.-D. Shih. Angular layer of a singularly perturbed parabolic problem with corner singularity. *Canad. Appl. Math. Quart.*, 9:159–188, 2001.

[Shi03]   G. I. Shishkin. Grid approximation of a singularly perturbed one-dimensional heat equation on an unbounded domain excluding a rectangle with sides noncollinear to the axes. *Russian J. Numer. Anal. Math. Modelling*, 18:429–454, 2003.

[Shi04a]   G. I. Shishkin. Discrete approximations of solutions and derivatives for a singularly perturbed parabolic convection-diffusion equation. *J. Comput. Appl. Math.*, 166:247–266, 2004.

[Shi04b]   G. I. Shishkin. Limitations of adaptive mesh refinement techniques for singularly perturbed problems with a moving interior layer. *J. Comput. Appl. Math.*, 166:267–280, 2004.

[Shi07a]   S.-D. Shih. Internal layers of some parabolic singularly perturbed problems. *ZAMM Z. Angew. Math. Mech.*, 87:831–844, 2007.

[Shi07b]   G. I. Shishkin. Approximation of systems of singularly perturbed elliptic reaction-diffusion equations with two parameters. *Comput. Math. Math. Phys.*, 47:797–828, 2007.

[Shi08]   G. I. Shishkin. Optimal difference schemes on piecewise uniform meshes for a singularly perturbed parabolic convection-diffusion equation. *Math. Model. Anal.*, 13:99–112, 2008.

[SK87]   S. D. Shih and R. B. Kellogg. Asymptotic analysis of a singular perturbation problem. *SIAM J. Math. Anal.*, 18:1467–1511, 1987.

[Sma74] D. R. Smart. *Fixed point theorems.* Cambridge University Press, London, 1974. Cambridge Tracts in Mathematics, No. 66.

[SO86]   M. Stynes and E. O'Riordan. A finite element method for a singularly perturbed boundary value problem. *Numer. Math.*, 50:1–15, 1986.

[SO87]   M. Stynes and E. O'Riordan. L^1 and L^∞ uniform convergence of a difference scheme for a semilinear singular perturbation problem. *Numer. Math.*, 50:519–531, 1987.

[SO89]   M. Stynes and E. O'Riordan. Uniformly convergent difference schemes for singularly perturbed parabolic diffusion-convection problems without turning points. *Numer. Math.*, 55:521–544, 1989.

[SO91]   M. Stynes and E. O'Riordan. An analysis of a singularly perturbed two-point boundary value problem using only finite element techniques. *Math. Comp.*, 56:663–675, 1991.

[SO97]   M. Stynes and E. O'Riordan. A uniformly convergent Galerkin method on a Shishkin mesh for a convection-diffusion problem. *J. Math. Anal. Appl.*, 214:36–54, 1997.

[SR97]   M. Stynes and H.-G. Roos. The midpoint upwind scheme. *Appl. Numer. Math.*, 23(3):361–374, 1997.

[SS94]   G. Sun and M. Stynes. Finite element methods on piecewise equidistant meshes for interior turning point problems. *Numer. Algor.*, 8:111–129, 1994.

[SS95a] G. Sun and M. Stynes. Finite element methods for singularly perturbed higher order elliptic two-point boundary value problems I: reaction-diffusion type. *IMA J. Numer. Anal.*, 15:117–139, 1995.

[SS95b] G. Sun and M. Stynes. Finite element methods for singularly perturbed higher order elliptic two-point boundary value problems II: convection-diffusion type. *IMA J. Numer. Anal.*, 15:197–219, 1995.

[SS98]   R. Sacco and M. Stynes. Finite element methods for convection-diffusion problems using exponential splines on triangles. *Comput. Math. Appl.*, 35(3):35–45, 1998.

[SSH04] G. I. Shishkin, L. P. Shishkina, and P. W. Hemker. A class of singularly perturbed convection-diffusion problems with a moving interior layer. An *a posteriori* adaptive mesh technique. *Comput. Methods Appl. Math.*, 4:105–127, 2004.

[SSX98] C. Schwab, M. Suri, and C. Xenophontos. The *hp* finite element method for problems in mechanics with boundary layers. *Comput. Methods Appl. Mech. Engrg.*, 157(3-4):311–333, 1998.

[ST89]   F. Schieweck and L. Tobiska. A nonconforming finite element method of upstream type applied to the stationary Navier-Stokes equations. *RAIRO Numer. Anal.*, 23:627–647, 1989.

[ST95]   M. Stynes and L. Tobiska. Necessary L_2-uniform conditions for difference schemes for two-dimensional convection-diffusion problems. *Comput. Math. Applic.*, 29:45–53, 1995.

[ST96]   F. Schieweck and L. Tobiska. An optimal order error estimate for an upwind discretization of the Navier-Stokes equations. *Numer. Methods Partial Different. Equations*, 12:407–421, 1996.

[ST98]   M. Stynes and L. Tobiska. A finite difference analysis of a streamline diffusion method on a Shishkin mesh. *Numer. Algorithms*, 18:337–360, 1998.

[ST03]   M. Stynes and L. Tobiska. 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, 2003.

[Ste99]   R. Stenberg. Analysis of mixed finite element methods for the Stokes problem: A unified approach. *Math. Comput.*, 42:9–23, 1999.

[Ste05]   R. Stevenson. The uniform saturation property for a singularly perturbed reaction-diffusion equation. *Num. Math.*, 101: 355–379, 2005.

[Ste07]   R. Stevenson. Optimality of a standard adaptive finite element method. *Foundations of Comput. Math.*, 7:245–269, 2007.

[Sto79]   G. Stoyan. Monotone difference schemes for convection diffusion problems. *ZAMM*, 59:361–372, 1979.

[Sto82]   G. Stoyan. A monotone mesh approximation for a partial differential equation with one space variable. *Diff. Equations (USSR)*, 18:886–897, 1982.

[Str04]   J. C. Strikwerda. *Finite difference schemes and partial differential equations.* SIAM, Philadelphia, PA, 2004.

[Sty03]   M. Stynes. A jejune heuristic mesh theorem. *Comput. Meth. Appl. Math.*, 3:488–492, 2003.

[Su87]   Y.-C. Su. The boundary layer scheme for a singularly perturbed problem for the second order elliptic equation in the rectangle. *Appl. Math. and Mech.*, 8:203–210, 1987.

[SU91]   K. Surla and Z. Uzelac. The exponential spline collocation method for boundary value problems. In H.-G. Roos, A. Felgenhauer, and L. Angermann, editors, *Numerical methods in singularly perturbed problems. Proc.,*, pages 147–154. TU Dresden, 1991.

[SV82]   J. A. Spriet and G. C. Vansteenkiste. *Computer-aided modelling and simulation.* Academic Press, London, 1982.

[SV85]   L. R. Scott and M. Vogelius. Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials. *RAIRO Modél. Math. Anal. Numér.*, 19:111–143, 1985.

[SW83]   A. H. Schatz and L. B. Wahlbin. On the finite element method for singularly perturbed reaction-diffusion problems in two and one dimensions. *Math. Comp.*, 40:47–89, 1983.

[SW96]   P. M. Selwood and A. J. Wathen. Convergence rates and classification for one-dimensional finite-element meshes. *IMA J. Numer. Anal.*, 16:65–74, 1996.

[SW05]   S. Sun and M. F. Wheeler. Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media. *SIAM J. Numer. Anal.*, 43:195–219, 2005.

[SYZ07]   Q. S. Song, G. Yin, and Z. Zhang. An ϵ-uniform finite element method for singularly perturbed two-point boundary value problems. *Int. J. Numer. Anal. Model.*, 4:127–140, 2007.

[SZ90]   L. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.

[Tab77]   M. Tabata. A finite element approximation corresponding to the upwind differencing. *Memoirs of Numerical Mathematics*, 1:47–63, 1977.

[TBA+92] T. E. Tezduyar, M. Behr, S. K. Aliabadi, S. Mittal, and S. E. Ray. A new mixed preconditioning method for the finite element computations. *Comput. Methods Appl. Mech. Engrg.*, 99(1):27–42, 1992.

[Tem83]   R. Temam. *Navier-Stokes equations and nonlinear functional analysis*, volume 41 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1983.

[TMRS92] T. E. Tezduyar, S. Mittal, S. E. Ray, and R. Shih. Incompressible flow computations with stabilized bilinear and linear equal order interpolation velocity pressure elements. *Comput. Methods Appl. Mech. Eng.*, 95:221–242, 1992.

[TMS00]   L. Tobiska, G. Matthies, and M. Stynes. Convergence properties of the streamline-diffusion finite element method on a Shishkin mesh for singularly perturbed elliptic equations with exponential layers. In L. G. Vulkov, J. J. H. Miller, and G. I. Shishkin, editors, *Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems*, pages 123–132, New York, 2000. Nova Science Publishers.

[Tob81]   L. Tobiska. A priori Abschätzungen für singulär gestörte elliptische Probleme zweiter Ordnung. *Beiträge zur Analysis*, 17:41–47, 1981.

[Tob83]   L. Tobiska. Diskretisierungsverfahren zur Lösung singulär gestörter Randwertprobleme. *ZAMM*, 63:115–123, 1983.

[Tob89]   L. Tobiska. Full and weighted upwind finite element methods. In J.W.Schmidt and H. Spät, editors, *Splines in Numerical Analysis*, volume 52 of *Mathematical Research*, pages 181–188. Akademie-Verlag, Berlin, 1989.

[Tob95]   L. Tobiska. A note on the artificial viscosity of numerical schemes. *Comp. Fluid Dyn.*, 5:281–290, 1995.

[Tob06]   L. Tobiska. Analysis of a new stabilized higher order finite element method for advection–diffusion equations. *Comput. Methods Appl. Mech. Eng.*, 196(1–3):538–550, 2006.

[TP86]   T. E. Tezduyar and Y. Park. Discontinuity-capturing finite element formulations for nonlinear convection-diffusion-reaction equations. *Comput. Methods Appl. Mech. Engrg.*, 59:307–325, 1986.

[Tro87]   G. M. Troianiello. *Elliptic differential equations and obstacle problems*. Plenum Press, New York, 1987.

[TS62]   A. N. Tihonov and A. A. Samarskiĭ. Homogeneous difference schemes on irregular meshes. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 2:812–832, 1962.

[TS76]   V. A. Titov and G. I. Shishkin. A numerical solution of a parabolic equation with small parameters multiplying the derivatives with respect to the space variables (in Russian). *Trudy. Inst. Mat. i Meh. Ural. Nauču Centr. Akad. Nauk SSR*, 21:38–43, 1976.

[TS03]   A. Toselli and C. Schwab. Mixed *hp*-finite element approximations on geometric edge and boundary layer meshes in three dimensions. *Numer. Math.*, 94:771–801, 2003.

[TT89]   A. Thiele and L. Tobiska. A weighted upwind finite element method for solving the stationary Navier-Stokes equations. *Wiss. Z. Techn. Univ. Magdeburg*, 33:13–20, 1989.

[Tur91]   S. Turek. *Ein robustes und effizientes Mehrgitterverfahren zur Lösung der instationäeren, inkompressiblen, 2D Navier-Stokes-Gleichungen mit diskret divergenzfreien finiten Elementen*. PhD thesis, Universität Heidelberg, 1991.

[Tur99]   S. Turek. *Efficient Solvers for incompressible flow problems. An algorith-mic and computational approach*, volume 6 of *Lecture Notes in Compu-tational Science and Engineering*. Springer-Verlag, Berlin, Heidelberg, New York, 1999.

[TV96]   L. Tobiska and R. Verfürth. Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations. *SIAM J. Numer. Anal.*, 33:107–127, 1996.

[VB90]   A. B. Vassiljewa and W. F. Butusov. *Asymptotic expansions in the theory of singular perturbations (in Russian)*. Higher School, Moscow, 1990.

[VBK95]   A. B. Vasil′eva, V. F. Butuzov, and L. V. Kalachev. *The boundary function method for singular perturbation problems*, volume 14 of *SIAM Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1995.

[vD94]   M. v. Dyke. Nineteenth-century roots of the boundary layer idea. *SIAM Review*, 36:415–424, 1994.

[Ver89]   R. Verfürth. A posteriori error estimators for the Stokes equation. *Numer. Math.*, 55:309–325, 1989.

[Ver96]   R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner, 1996.

[Ver98a]   R. Verfürth. A posteriori error estimators for convection-diffusion prob-lems. *Numer. Math.*, 80:641–663, 1998.

[Ver98b]   R. Verfürth. Robust a posteriori error estimators for a singularly per-turbed reaction-diffusion equation. *Numer. Math.*, 78:479–493, 1998.

[Ver05]   R. Verfürth. Robust a posteriori error estimates for stationary convection-diffusion equations. *SIAM J. Numer. Anal.*, 43:1766–1782, 2005.

[VF93]   R. Vulanović and P. A. Farrell. Continuous and numerical analysis of a multiple boundary turning point problem. *SIAM J. Numer. Anal.*, 30:1400–1418, 1993.

[VK93]   C. B. Vreugdenhil and B. Koren, editors. *Numerical methods for advection-diffusion problems*. Vieweg, Braunschweig, 1993.

[Voh07]   M. Vohralik. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.*, 45:1570–1599, 2007.

[Vol65]   E. A. Volkov. Differentiability properties of solutions of boundary value problems for the Laplace and Poisson equations. *Proc. Steklov Inst. Math.*, 77:101–126, 1965.

[VR90]   K. Vajravelu and D. Rollins. On solutions of some unsteady flows over a continuous, moving, porous flat surface. *J. Math. Anal. Appl.*, 153:52–63, 1990.

[Vul83]   R. Vulanović. On a numerical solution of a type of singularly perturbed boundary value problem by using a special discretization mesh. *Univ. u Novom Sadu Zb. Rad. Prirod.-Mat. Fak. Ser. Mat.*, 13:187–201, 1983.

[Vul86]   R. Vulanović. Mesh construction for discretization of singularly per-turbed boundary value problems. Doctoral dissertation. University of Novi Sad, 1986.

[Vul89]   R. Vulanović. A uniform numerical method for quasilinear singular per-turbation problems without turning points. *Computing*, 41:97–106, 1989.

[Vul91]   R. Vulanović. Non-equidistant finite difference methods for elliptic sin-
           gular perturbation methods. In J.J.H. Miller, editor, *Computational
           methods for boundary and interior layers in several dimensions*, pages
           203–223. Boole Press, Dublin, 1991.

[Vul01]   R. Vulanović. A priori meshes for singularly perturbed quasilinear two-
           point boundary value problems. *IMA J. Numer. Anal.*, 21:349–366,
           2001.

[Vul07]   R. Vulanović. The layer-resolving transformation and mesh generation
           for quasilinear singular perturbation problems. *J. Comput. Appl. Math.*,
           203:177–189, 2007.

[vV78]   M. van Veldhuizen. Higher order methods for a singularly perturbed
           problem. *Numer. Math.*, 30:267–279, 1978.

[Wah74]   L. B. Wahlbin. A dissipative Galerkin method for the numerical solution
           of first order hyperbolic equations. In *Mathematical aspects of finite
           elements in pde's (Proc. Sympos., Math. Res. Center,*, pages 147–169.
           Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison,
           Academic Press, New York, 1974.

[Wah95]   L. B. Wahlbin. *Superconvergence in Galerkin finite element methods.*
           Lecture Notes in Mathematics 1605. Springer-Verlag, Berlin, 1995.

[Wan99]   S. Wang. A new exponentially fitted triangular finite element method for
           the continuity equations in the drift-diffusion model of semiconductor
           devices. *M2AN Math. Model. Numer. Anal.*, 33:99–112, 1999.

[Wan00]   H. Wang. An optimal-order error estimate for an ELLAM scheme for
           two-dimensional linear advection-diffusion equations. *SIAM J. Numer.
           Anal.*, 37:1338–1368, 2000.

[Was65]   W. Wasow. *Asymptotic expansions for ordinary differential equations.*
           John Wiley, New York, 1965.

[WDE$^+$99] H. Wang, H. K. Dahle, R. E. Ewing, M. S. Espedal, R. C. Sharpley, and
           S. Man. An ELLAM scheme for advection-diffusion equations in two
           dimensions. *SIAM J. Sci. Comput.*, 20:2160–2194, 1999.

[Wen60]   B. Wendroff. On centered difference equations for hyperbolic systems.
           J. SIAM, 8:549–555, 1960.

[WER95]   H. Wang, R. E. Ewing, and T. F. Russell. Eulerian-Lagrangian localized
           adjoint methods for convection-diffusion equations and their convergence
           analysis. *IMA J. Numer. Anal.*, 15:405–459, 1995.

[Wes96]   P. Wesseling. Uniform convergence of discretization error for a singular
           perturbation problem. *Numer. Methods Partial Differential Equations*,
           12(6):657–671, 1996.

[Whe73]   M. F. Wheeler. An optimal L_∞ error estimate for Galerkin approxi-
           mations to solutions of two point boundary value problems. *SIAM J.
           Numer. Anal.*, 10:914–917, 1973.

[Wid71]   O. B. Widlund. On Lax's theorem on Friedrichs type finite difference
           schemes. *Comm. Pure Appl. Math.*, 24:117–123, 1971.

[Wig70]   N. M. Wigley. Mixed boundary value problems in domains with corners.
           Math. Z., 115:33–52, 1970.

[WS89]   J. J. Westerink and D. Shea. Consistent higher degree Petrov-Galerkin
           methods for the solution of the transient convection-diffusion equation.
           Int. J. Numer. Methods Eng., 28:1077–1101, 1989.

[WW]   K. Wang and H. Wang. A uniform estimate for the ELLAM scheme for transport equations. *Numer. Methods Partial Differential Equations.* (to appear).

[WW07]   H. Wang and K. Wang. Uniform estimates for Eulerian-Lagrangian methods for singularly perturbed time-dependent problems. *SIAM J. Numer. Anal.*, 45:1305–1329, 2007.

[WZ03]   H. Wang and W. Zhao. A modified alternating-direction finite volume method for modeling secondary hydrocarbon migration and accumulation processes. *Numer. Methods Partial Differential Equations*, 19:254–270, 2003.

[Xen03]   C. Xenophontos. A note on the convergence rate of the finite element method for singularly perturbed problems using the Shishkin mesh. *Appl. Math. Comput.*, 142(2-3):545–559, 2003.

[XF03]   C. Xenophontos and S. R. Fulton. Uniform approximation of singularly perturbed reaction-diffusion problems by the finite element method on a Shishkin mesh. *Numer. Methods Partial Differential Equations*, 19:89–111, 2003.

[XZ99]   J. Xu and L. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comp.*, 68(228):1429–1446, 1999.

[XZ03]   J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94:195–202, 2003.

[XZ07]   Z. Xie and Z. Zhang. Superconvergence of dG method for one-dimensional singularly perturbed problems. *J. Comput. Math.*, 25:185–200, 2007.

[YH86]   C.-C. Yu and J. C. Heinrich. Petrov-Galerkin methods for the time-dependent convective transport equation. *Int. J. Numer. Methods Eng.*, 23:883–901, 1986.

[YJS99]   X.-Y. Yue, L.-S. Jiang, and T.-M. Shih. Finite element analysis of a local exponentially fitted scheme for time-dependent convection-diffusion problems. *J. Comput. Math.*, 17:225–232, 1999.

[Yse83]   H. Yserentant. Die maximale Konsistenzordnung von Differenzenapproximationen nichtnegativer Art. *Numer. Math.*, 42:119–123, 1983.

[ZB92]   P. A. Zegeling and J. G. Blom. A note on the grid movement induced by mfe. *Int. J. Numer. Methods Eng.*, 35:623–636, 1992.

[Zha03]   Z. Zhang. Finite element superconvergence on Shishkin mesh for 2-d convection-diffusion problems. *Math. Comp.*, 72:1147–1177, 2003.

[Zha04]   Z. Zhang. Polynomial preserving gradient recovery and a posteriori error estimates for bilinear elements on irregular quadrilaterals. *Int. J. Numer. Anal. and Modelling*, 1:1–24, 2004.

[Zha05]   S. Zhang. A new family of stable mixed finite elements for the 3D Stokes equations. *Math. Comp.*, 74:543–554, 2005.

[Zho95]   G. Zhou. Local pointwise error estimates for the streamline diffusion method applied to nonstationary hyperbolic problems. *East-West J. Numer. Math.*, 3(3):217–235, 1995.

[Zho97]   G. Zhou. How accurate is the streamline diffusion finite element method? *Math. Comp.*, 66(217):31–44, 1997.

[ZL94]   A. Zhou and Q. Lin. Optimal and superconvergence estimates of the finite element method for a scalar hyperbolic equation. *Acta Mathematica Scientia*, 14:90–94, 1994.

[ZN05]   Z. Zhang and A. Nhaga. A new finite element gradient recovery method: superconvergence property. *SIAM J. Sci. Comp.*, 26:1192–1213, 2005.

[ZR96]   G. Zhou and R. Rannacher. Pointwise superconvergence of the streamline diffusion finite element method. *Numer. Methods Partial Differ. Equations*, 12(1):123–145, 1996.

[ZR05]   H. Zarin and H.-G. Roos. Interior penalty discontinuous approximations of convection-diffusion problems with parabolic layers. *Numer. Math.*, 100:735–759, 2005.

Index