An hp finite element method for convection-diffusion problems in one dimension

Originally published in:
IMA Journal of Numerical Analysis 19(3), https://doi.org/10.1093/imanum/19.3.425

IMA Journal of Numerical Analysis (1999) **19**, 425–453

An hp finite element method for convection–diffusion problems in one dimension

JENS MARKUS MELENK AND CHRISTOPH SCHWAB
Seminar für Angewandte Mathematik, ETH Zürich, CH-8092 Zürich, Switzerland

[Received 1 December 1997 and in revised form 30 October 1998]

We analyse an hp FEM for convection–diffusion problems. Stability is achieved by suitably upwinded test functions, generalizing the classical α-quadratically upwinded and the Hemker test-functions for piecewise linear trial spaces (see, e.g., Morton 1995 *Numerical Solutions of Convection–Diffusion Problems*, Oxford: Oxford University Press, and the references therein). The method is proved to be stable independently of the viscosity. Further, the stability is shown to depend only weakly on the spectral order. We show how sufficiently accurate, approximate upwinded test functions can be computed on each element by a local least-squares FEM. Under the assumption of analyticity of the input data, we prove robust exponential convergence of the method. Numerical experiments confirm our convergence estimates and show robust exponential convergence of the hp FEM even for viscosities of the order of the machine precision, i.e., for the limiting transport problem.

1. The convection–diffusion problem

1.1 *Problem formulation*

In $\Omega = (-1, 1)$ we consider the model convection–diffusion problem

$$L_\varepsilon u_\varepsilon := -\varepsilon u_\varepsilon'' + a(x)u_\varepsilon' + b(x)u_\varepsilon = f(x) \tag{1.1}$$

with the boundary conditions

$$u(\pm 1) = \alpha^\pm \in \mathbb{R}. \tag{1.2}$$

Here $\varepsilon > 0$ is the diffusivity, $u(x)$ is, for example, the concentration of a transported substance, $a(x)$ is the velocity of the transporting medium, $b(x)$ specifies losses/sources of the substance and $f(x)$ is an external source term. Throughout this work, we make the following assumptions on the coefficients $a \in C^1[-1, 1]$, $b \in C^0[-1, 1]$: There are constants $\underline{b} \in \mathbb{R}$, $\underline{a}$, γ_1, $\gamma_2 > 0$ independent of $\varepsilon \in (0, 1]$ such that

$$\begin{gathered} a(x) \geqslant \underline{a}, \quad b(x) \geqslant \underline{b}, \quad \min\{\underline{a}^2, \underline{a}^2 + 4\underline{b}\varepsilon\} \geqslant \gamma_1^2, \\ \max\left\{\frac{\underline{a} - \sqrt{\underline{a}^2 + 4\underline{b}\varepsilon}}{2\varepsilon}, 0\right\} \leqslant \gamma_2. \end{gathered} \tag{1.3}$$

As we will also consider the adjoint problem of (1.1) (cf. Section 2 ahead) we stipulate the existence of $\underline{b}^* \in \mathbb{R}$ and $\gamma_1^*, \gamma_2^* > 0$ such that

$$a(x) \geqslant \underline{a} > 0, \quad b(x) - a'(x) \geqslant \underline{b}^*, \quad \min\{\underline{a}^2, \underline{a}^2 + 4\underline{b}^*\varepsilon\} \geqslant (\gamma_1^*)^2,$$
$$\max\left\{\frac{\underline{a} - \sqrt{\underline{a}^2 + 4\underline{b}^*\varepsilon}}{2\varepsilon}, 0\right\} \leqslant \gamma_2^*. \tag{1.4}$$

Expression (1.3) ensures the unique solvability of (1.1), (1.2) while (1.4) guarantees the unique solvability of the adjoint problem. Note that for given a, b, the constants γ_1, γ_2, γ_1^*, γ_2^* exist under the assumption that the diffusivity ε is sufficiently small.

The finite element approximation of (1.1), (1.2) for small ε is nontrivial due to the singular perturbation character of the problem which manifests itself in two distinct phenomena.

Firstly, the solution u_ε exhibits a boundary layer near the outflow boundary $x = 1$; the presence of sharp boundary layers brings about the problem of *robust consistency*, that is, the question of designing trial spaces that can approximate the solutions of (1.1), (1.2) uniformly in the parameter $\varepsilon \in (0, 1]$. Secondly, as is well known, variational formulations of (1.1), (1.2) based on the same piecewise polynomial subspace of H_0^1 as trial and test space are not uniformly stable with respect to the parameter ε. *Robust stability*, however, can be achieved by Petrov–Galerkin (PG) methods, where the test and trial spaces are distinct.

Numerous schemes have been proposed in the literature to address these two issues (see the monographs by Morton (1995), Roos *et al* (1996), Miller *et al* (1996) and the references therein). In most PG methods advocated in practice, the trial spaces consist of piecewise polynomials of a fixed, low degree p and accuracy is obtained by decreasing the mesh width h. Such h versions, given stability of the method, can only lead to *robust algebraic* rates of convergence of the form $O(N^{-q})$ (uniformly in ε) for some $q > 0$ where N stands for the number of degrees of freedom. Recently Schwab & Suri (1996), Schwab *et al* (1996) and Melenk (1997) demonstrated for the related reaction–diffusion equation, which also exhibits sharp boundary layers, that *robust exponential* convergence of the hp FEM can be obtained, i.e., the error is $O(\exp(-bN))$ for some $b > 0$ independent of ε. In the present paper it will also be shown that robust exponential approximability can be achieved by piecewise polynomials of increasingly higher degree on appropriate meshes for the solutions of convection–diffusion equations. Thus, provided that a stable method is available, this approximation result allows us to obtain robust exponential convergence for problem (1.1), (1.2). To a large extent, therefore, the present paper is devoted to the construction of a stable method yielding, for the first time, a class of exponentially converging PG methods for convection-dominated problems.

Two popular approaches to overcoming stability problems are *stabilized* methods, such as the streamline diffusion method (see, e.g., Johnson & Nävert (1981)) and the related SUPG (see, e.g., Hughes & Brooks (1979, 1982)) on the one hand, and PG schemes that are based on upwinded test functions on the other hand (see, e.g., Christie *et al* (1978), Hemker (1977)). The PG idea, originally formulated as an h version, is generalized here to a p and hp setting by introducing appropriate upwinded test functions; in particular, the

upwinded test functions of the low-order h version schemes of Christie *et al* (1978) and Hemker (1977) are special cases of our more general approach. The idea of stabilization can also be pursued in an hp context, and we refer to Melenk & Schwab (1998) for the analysis of the hp streamline diffusion method as a typical representative of stabilized methods.

We base our hp FEM on the variational framework of Szymczak (1982) and Szymczak & Babuška (1984), where the h-version FEM was analysed and optimal convergence rates, uniform in ε, were shown. The crucial ingredient in the work of Szymczak (1982) and Szymczak & Babuška (1984) (similarly to Christie *et al* (1978), Hemker (1977)) was the construction of suitable, upwinded test functions by asymptotic analysis of the elemental adjoint problem. The generalization of this asymptotic analysis (and the techniques used by Christie *et al* (1978), Hemker (1977)) to high-order elements and higher dimensions is not straightforward.

Here we propose therefore a fully numerical method. More precisely, we show how, for hp-trial spaces with any mesh-degree combination, sufficiently accurate *approximate upwinded test functions* can be stably computed. The calculation of the test functions is completely localized to either a single element or a patch of elements and done by a least-squares-like method (which is uniformly stable in ε). This can be simply performed as part of the usual element stiffness matrix generation in the hp FEM. Our analysis shows that (a) the approximate test functions thus obtained do ensure stability, and (b) fairly crude approximations of the test functions suffice, so that the work spent in computing these test functions can be expected to be moderate. Most importantly, no analytical input in the form of asymptotic expansions or boundary layer functions is necessary—the method is fully computational and at least conceptually generalizes to two- and three-dimensional problems.

The outline of the paper is as follows. For the robust exponential approximability result, one needs sharp regularity results for the solution u_ε of (1.1), (1.2). These results are collected in Section 1.2. In Section 2, the variational framework is provided that will be the basis of our analysis. Section 3 contains the main results of the paper. We show that for piecewise polynomial trial functions, one can construct upwinded test functions that lead to a uniformly stable method. In Section 3.3, we employ the regularity results of Section 1.2 to show that robust exponential approximability of the solutions of (1.1), (1.2) is possible on appropriate meshes. Together with the uniform stability of the method, we therefore obtain robust exponential convergence. By means of a perturbation argument we show in Section 4 that already fairly crude approximations to the upwinded test functions of Section 3 suffice to ensure the stability of the method. An application of the perturbation argument can be found in Section 4.2, where we propose a least-squares approach to approximate the upwinded test functions. Finally, in Section 5 our results are illustrated with some numerical examples.

1.2 *Regularity*

Let us consider (1.1) on $\Omega = (-1, 1)$ with *analytic* input data $a(x)$, $b(x)$, $f(x)$ satisfying

$$\|a^{(n)}\|_{L^\infty(\Omega)} \leqslant C_a \gamma_a^n \qquad \forall n \in \mathbb{N}_0, \tag{1.5}$$

$$\|b^{(n)}\|_{L^\infty(\Omega)} \leqslant C_b \gamma_b^n \qquad \forall n \in \mathbb{N}_0, \tag{1.6}$$

$$\|f^{(n)}\|_{L^\infty(\Omega)} \leqslant C_f \gamma_f^n \qquad \forall n \in \mathbb{N}_0, \tag{1.7}$$

for some constants C_a, C_b, C_f, γ_a, γ_b, $\gamma_f > 0$. Assumptions (1.3) and (1.5)–(1.7) ensure the existence of a unique, analytic solution u_ε of (1.1), (1.2). The purpose of this subsection is to provide the regularity properties of u_ε in terms of the parameter ε and the constants of (1.3), (1.5)–(1.7). These regularity results are necessary for the proof of *robust exponential convergence* of the *hp* FEM obtained in the present paper. Although regularity results related to the ones presented here can be found in the literature (e.g., in the books by Roos *et al* (1996), Morton (1995)), the specific derivative bounds seem to be new (see also Melenk (1997) for the related case of a reaction–diffusion equation).

The solution u_ε of (1.1), (1.2) is analytic on $\overline{\Omega}$; however, for small values of ε, it exhibits a boundary layer at the outflow boundary. This boundary layer behaviour can be characterized with the aid of asymptotic expansions: For any expansion order $M \in \mathbb{N}_0$, we have the standard decomposition (as done by, e.g., Gartland (1987))

$$u_\varepsilon = w_M + C_M u_\varepsilon^+ + r_M. \tag{1.8}$$

Here, u_M is the *asymptotic part* given by

$$\begin{aligned} w_M &:= \sum_{j=0}^{M} \varepsilon^j u_j + \alpha^- \mathrm{e}^{-\Lambda(x)}, \\ u_{j+1}(x) &:= \mathrm{e}^{-\Lambda(x)} \int_{-1}^{x} \frac{\mathrm{e}^{\Lambda(t)}}{a(t)} u_j''(t)\,\mathrm{d}t, \qquad j = 0,\ldots, M-1, \\ u_0(x) &:= \mathrm{e}^{-\Lambda(x)} \int_{-1}^{x} \frac{\mathrm{e}^{\Lambda(t)}}{a(t)} f(t)\,\mathrm{d}t, \\ \Lambda(x) &:= \int_{-1}^{x} \frac{b(x)}{a(x)}\,\mathrm{d}t. \end{aligned} \tag{1.9}$$

The *outflow boundary layer* u_ε^+ solves the problem

$$L_\varepsilon u_\varepsilon^+ = 0 \quad \text{on } \Omega, \qquad u_\varepsilon^+(-1) = 0, \quad u_\varepsilon^+(1) = 1, \tag{1.10}$$

and C_M is given by

$$C_M := \alpha^+ - w_M(1). \tag{1.11}$$

Finally, the remainder r_M is given as the solution of

$$L_\varepsilon r_M = \varepsilon^{M+1} u_M'' \quad \text{on } \Omega, \qquad r_M(\pm 1) = 0. \tag{1.12}$$

Note that for $M = 0$ the function w_0 solves the *limiting transport problem* given by (1.1) with $\varepsilon = 0$ and the boundary condition $w_0(-1) = \alpha^-$.

THEOREM 1.1 Let u_ε be the solution of (1.1), (1.2). Then there are constants C, K depending only on the constants in (1.5)–(1.7) and on the constants $\underline{a}$, γ_1, γ_2 in (1.3) such that

$$\|u_\varepsilon^{(n)}\|_{L^\infty(\Omega)} \leqslant C K^n \max(n, \varepsilon^{-1})^n \qquad \forall n \in \mathbb{N}_0, \tag{1.13}$$

$$|u_\varepsilon^{+(n)}(x)| \leqslant C K^n \max(n, \varepsilon^{-1})^n \mathrm{e}^{-\underline{a}(1-x)/(2\varepsilon)} \quad \forall n \in \mathbb{N}_0, \quad x \in \Omega. \tag{1.14}$$

Furthermore, under the assumption $0 < \varepsilon MK \leqslant 1$, the terms in the decomposition (1.8) satisfy

$$\|w_M^{(n)}\|_{L^\infty(\Omega)} \leqslant C K^n n! \qquad \forall n \in \mathbb{N}_0, \tag{1.15}$$

$$\|r_M^{(n)}\|_{L^\infty(\Omega)} \leqslant C \varepsilon^{1-n} (\varepsilon MK)^M, \qquad n = 0, 1, 2, \tag{1.16}$$

$$|C_M| \leqslant C. \tag{1.17}$$

Theorem 1.1 can be proved using the same techniques that are employed by Melenk (1997). The lemmata that allow for the use of these techniques are collected in the Appendix; the reader is referred to Melenk & Schwab (1997) where the arguments are worked out in detail.

2. Variational formulation

Without loss of generality, we may analyse (1.1) with homogeneous Dirichlet data

$$\alpha^\pm = 0 \tag{2.1}$$

by the standard argument of seeking u_ε in the form $u_\varepsilon = \tilde{u}_\varepsilon + u_0$ (where u_0 is linear and satisfies the boundary conditions (1.2)) and then noting that this leads to (1.1), (2.1) for $\tilde{u}_\varepsilon$ with the same operator L_ε and suitably adjusted right-hand side f which is analytic and independent of ε.

To motivate our variational formulation, we observe that multiplication of (1.1) by a test function v and twofold integration by parts gives a so-called *very weak variational formulation*: Find $u \in L^2(\Omega)$ such that

$$B(u, v) := \int_\Omega u L_\varepsilon^* v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x =: F(v) \qquad \forall v \in (H^2 \cap H_0^1)(\Omega).$$

Here, L_ε^* denotes the adjoint of L_ε, i.e.

$$L_\varepsilon^* u = -\varepsilon u'' - a(x)u' + (b - a')(x)u \tag{2.2}$$

which is defined when $a \in C^1([-1, 1])$. There are several drawbacks with FEMs based on very weak variational formulations: first, a' is in general not globally continuous, but only elementwise smooth (if it stems, for example, from linearization of a nonlinear problem around a FE approximation of u); second, to obtain a good test space for a given trial space of possibly discontinuous functions, a global adjoint problem must be solved for each basis function; and third, the essential boundary conditions (1.2) are generally not satisfied by FE solutions. This leads us to a formulation which is situated 'between' the weak one based on $H_0^1 \times H_0^1$ and the very weak one based on $L^2 \times H^2 \cap H_0^1$.

We present Sobolev spaces with mesh-dependent norms introduced by Szymczak & Babuška (1984). For a collection of nodes $\{-1 = x_0 < x_1 < \cdots < x_N = 1\}$, we introduce the notation $I_j := (x_{j-1}, x_j)$, $h_j := |I_j| = x_j - x_{j-1}$, $m_j = (x_{j-1} + x_j)/2$ for $j = 1,..., N$. The elements I_j form a mesh $\mathcal{T} = \{I_j : j = 1,..., N\}$ on Ω. Let $\{\rho_j : j = 1,..., N - 1\}$ be given by

$$\rho_j := \tfrac{1}{2}(h_j + h_{j+1}) \tag{2.3}$$

and set $h := \max\{h_j : j = 1,..., N\}$. Then we define the trial space $H^0_{\mathcal{T}}$ as the completion of $H^1_0(\Omega)$ with respect to the mesh-dependent norm

$$\|u\|_{H^0_{\mathcal{T}}} := \left(\int_{-1}^{1} |u|^2 \, \mathrm{d}x + \sum_{j=1}^{N-1} \rho_j \left| u(x_j) \right|^2 \right)^{1/2}. \tag{2.4}$$

The space $H^0_{\mathcal{T}}$ thus obtained is a Hilbert space and is isomorphic to $L^2(\Omega) \oplus \mathbb{R}^{N-1}$ so that every $u \in H^0_{\mathcal{T}}$ is of the form $u = (\tilde{u}, d_1, d_2,..., d_{N-1})$ and

$$\|u\|_{H^0_{\mathcal{T}}} = \left(\|\tilde{u}\|^2_{L^2} + \sum_{j=1}^{N-1} \rho_j \left| d_j \right|^2 \right)^{1/2}. \tag{2.5}$$

Consistent with our definition, we say that $u \in H^0_{\mathcal{T}} \cap H^1_0$ if $\tilde{u} \in H^1_0$ and $d_j = u(x_j)$.

Next, we introduce the test space

$$H^2_{\mathcal{T}} := \left\{ v \in H^1_0(\Omega) : v|_{I_j} \in H^2(I_j),\ j = 1,..., N \right\}. \tag{2.6}$$

On the pair $H^0_{\mathcal{T}} \times H^2_{\mathcal{T}}$ we define the bilinear form $B_{\mathcal{T}}(\cdot, \cdot)$ by

$$B_{\mathcal{T}}(u, v) := \sum_{j=1}^{N} \int_{I_j} \tilde{u} L^*_\varepsilon v \, \mathrm{d}x - \sum_{j=1}^{N-1} d_j \left[\varepsilon v'(x_j) \right] \tag{2.7}$$

where $\left[v'(x_j) \right]$ denotes the jump of v' at $x_j \in \mathcal{T}$. We equip the space $H^2_{\mathcal{T}}$ with the norm

$$\|v\|_{H^2_{\mathcal{T}}} := \left(\sum_{j=1}^{N} \left\| L^*_\varepsilon v \right\|^2_{L^2(I_j)} + \sum_{j=1}^{N-1} \frac{\left| \left[\varepsilon v'(x_j) \right] \right|^2}{\rho_j} \right)^{1/2}. \tag{2.8}$$

We remark in passing that so far we have used $a \in C^0([-1, 1]) \cap C^1(\bar{I}_j)$, $j = 1,..., N$, rather than $a \in C^1[-1, 1]$. With these definitions we have

PROPOSITION 2.1 For any mesh $\mathcal{T}$ the bilinear form $B_{\mathcal{T}}(\cdot, \cdot)$ satisfies

$$|B_{\mathcal{T}}(u, v)| \leqslant \|u\|_{H^0_{\mathcal{T}}} \|v\|_{H^2_{\mathcal{T}}} \qquad \forall u \in H^0_{\mathcal{T}}, v \in H^2_{\mathcal{T}}, \tag{2.9}$$

$$\inf_{0 \neq v \in H^2_{\mathcal{T}}} \sup_{0 \neq u \in H^0_{\mathcal{T}}} \frac{B_{\mathcal{T}}(u, v)}{\|u\|_{H^0_{\mathcal{T}}} \|v\|_{H^2_{\mathcal{T}}}} \geqslant 1, \tag{2.10}$$

$$\forall 0 \neq u \in H^0_{\mathcal{T}} : \sup_{v \in H^2_{\mathcal{T}}} B_{\mathcal{T}}(u, v) > 0. \tag{2.11}$$

Proof. The bound (2.9) follows directly from the definition of the norms and the Cauchy–Schwarz inequality. (2.10), (2.11) follow immediately from Lemma 2.6 of Szymczak & Babuška (1984). □

REMARK 2.2 It is easy to see that for a continuous bilinear form $B : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ where $\mathcal{X}$, $\mathcal{Y}$ are two reflexive Banach spaces over $\mathbb{R}$, the conditions

$$\inf_{0\neq v\in\mathcal{Y}} \sup_{0\neq u\in\mathcal{X}} \frac{B(u,v)}{\|u\|_{\mathcal{X}}\,\|v\|_{\mathcal{Y}}} \geqslant \gamma > 0, \qquad \forall 0 \neq u \in \mathcal{X} : \ \sup_{v\in\mathcal{Y}} B(u,v) > 0 \tag{2.12}$$

and

$$\inf_{0\neq u\in\mathcal{X}} \sup_{0\neq v\in\mathcal{Y}} \frac{B(u,v)}{\|v\|_{\mathcal{Y}}\,\|u\|_{\mathcal{X}}} \geqslant \gamma > 0, \qquad \forall 0 \neq v \in \mathcal{Y} : \ \sup_{u\in\mathcal{X}} B(u,v) > 0 \tag{2.13}$$

are equivalent; in particular, the constant $\gamma > 0$ is the same. Furthermore, in the special case $\dim \mathcal{X} = \dim \mathcal{Y} < \infty$, the first condition alone of either (2.12) or (2.13) implies both conditions of (2.12) or (2.13).

Based on Remark 2.2, Proposition 2.1 allows us to infer the existence of a solution $u \in H^0_{\mathcal{T}}$ of the problem:

$$\text{find } u \in H^0_{\mathcal{T}} \text{ such that } B_{\mathcal{T}}(u,v) = F(v) \qquad \forall v \in H^2_{\mathcal{T}} \tag{2.14}$$

for continuous linear functionals F on the space $H^2_{\mathcal{T}}$:

PROPOSITION 2.3 Under assumption (1.3), for every $f \in L^1(\Omega)$, every $0 < \varepsilon \leqslant 1$, and every mesh $\mathcal{T}$, problem (2.14) admits a unique solution $u \in H^0_{\mathcal{T}}$ which satisfies

$$\|u\|_{H^0_{\mathcal{T}}} \leqslant C\|f\|_{L^1(\Omega)}$$

where $C > 0$ is independent of ε and $\mathcal{T}$.

Proof. This follows easily from the fact that $f \in L^1(\Omega)$ represents a bounded linear functional on $H^2_{\mathcal{T}}$ by the estimate

$$\|v\|_{L^\infty(\Omega)} \leqslant C_1\|v\|_{H^2_{\mathcal{T}}} \qquad \forall v \in H^2_{\mathcal{T}}, \tag{2.15}$$

which is equation (2.17) of Szymczak & Babuška (1984). □

REMARK 2.4 The proof of Proposition 2.3 rests on the continuous embedding $H^2_{\mathcal{T}} \subset H^1_0(\Omega) \subset C(\overline{\Omega})$. Hence, the right-hand side functional F in (2.14) may actually be represented by an L^1 function f plus a (finite) number of Dirac distributions.

The variational formulation (2.14) is the basis of the FE discretization.

3. *hp* finite element discretization

3.1 *The finite element spaces*

We associate with each element I_j a polynomial degree $p_j \geqslant 1$ and combine the p_j in the degree-vector $\vec{p}$. We also set $p := \max\{p_j : 1 \leqslant j \leqslant N\}$. The trial spaces $S_0^{\vec{p},\ell}(\mathcal{T})$ of our

finite element method are the usual spaces of continuous, piecewise polynomials of degree p_j satisfying the homogeneous boundary conditions (2.1) at the endpoints ± 1:

$$\begin{aligned} S^{\vec{p},\ell}(\mathcal{T}) &:= \left\{ u \in H^\ell(\Omega) : u|_{I_j} \in \Pi_{p_j}(I_j),\ j = 1,\ldots, N \right\}, \quad \ell = 1, 2,\ldots, \\ S_0^{\vec{p},\ell}(\mathcal{T}) &:= S^{\vec{p},\ell}(\mathcal{T}) \cap H_0^1(\Omega), \end{aligned} \tag{3.1}$$

where $\Pi_p(I)$ denotes the set of all polynomials of degree p on the interval I. If $\ell = 1$, we simply write $S_0^{\vec{p}}(\mathcal{T})$.

As test space we choose, following Szymczak & Babuška (1984), the space of L^*-splines of degree $\vec{p}$ defined by

$$S_L^{\vec{p}}(\mathcal{T}) := \left\{ v \in H_0^1(\Omega) :\ \left(L_\varepsilon^* v\right)\big|_{I_j} = 0 \text{ if } p_j = 1,\ \left(L_\varepsilon^* v\right)\big|_{I_j} \in \Pi_{p_j-2}(I_j) \text{ if } p_j \geqslant 2 \right\}. \tag{3.2}$$

Note that (1.3), (1.4) imply that L_ε and L_ε^* are injective. Hence (3.2) makes sense and the test functions belong to $H^2(I_j)$, $j = 1,\ldots, N$. We omit the argument $\mathcal{T}$ when it is clear from the context which mesh is meant. Note that, due to (2.2), the space $S_L^{\vec{p}}$ is well defined even if the coefficient $a(x)$ is only piecewise C^1. We also observe that

$$M = \dim(S_0^{\vec{p}}) = -1 + \sum_{j=1}^{N} p_j = \dim(S_L^{\vec{p}}). \tag{3.3}$$

The finite element approximation u_M is then obtained in the usual way:

$$\text{find } u_M \in S_0^{\vec{p}} \text{ such that } B_{\mathcal{T}}(u_M, v) = F(v) \qquad \forall v \in S_L^{\vec{p}}. \tag{3.4}$$

Due to (3.3), problem (3.4) amounts to solving a linear system of M equations for the M unknown coefficients of u_M.

3.2 *Stability*

Our main result in this section is

THEOREM 3.1 For all $0 < \varepsilon \leqslant 1$, $\mathcal{T}$ and $\vec{p}$ there holds

$$\inf_{0 \neq v \in S_L^{\vec{p}}} \sup_{0 \neq u \in S_0^{\vec{p}}} \frac{B_{\mathcal{T}}(u, v)}{\|u\|_{H_{\mathcal{T}}^0} \|v\|_{H_{\mathcal{T}}^2}} \geqslant \frac{1}{\gamma_M} \tag{3.5}$$

with $\gamma_M = \max\{\sqrt{5}, \sqrt{p}\}$.

REMARK 3.2 Theorem 3.1 extends the results of Szymczak (1982) where the case $p = 1$ is analysed.

Proof. We show that for every $v \in S_L^{\vec{p}}$ there exists $u_v \in S_0^{\vec{p}}$ such that

$$B_{\mathcal{T}}(u_v, v) \geqslant \|v\|_{H_{\mathcal{T}}^2}^2, \quad \|u_v\|_{H_{\mathcal{T}}^0} \leqslant \gamma_M \|v\|_{H_{\mathcal{T}}^2}.$$

To this end, we write

$$u_v|_{I_j} = \sum_{i=0}^{p_j} a_{ij} L_i \left(2\frac{x - m_j}{h_j}\right), \tag{3.6}$$

where L_i denotes the ith Legendre polynomial on $(-1, 1)$ normalized such that $L_i(1) = 1$.

A basis for $S_L^{\vec{p}}$ can be obtained as follows: First, we define *external, nodal upwinded shape functions* $\psi_{-1,j} \in H_0^1(\overline{I}_{j-1} \cup \overline{I}_j)$ by

$$\begin{aligned} L_\varepsilon^* \psi_{-1,j} &= 0 && \text{in } I_{j-1} \cup I_j, \quad j = 2,..., N, \\ \psi_{-1,j} &= 0 && \text{elsewhere}, \\ \psi_{-1,j}(x_k) &= \delta_{j,k+1}, && k = 1,..., N-1. \end{aligned} \tag{3.7}$$

Note that $\psi_{-1,j} \in H^2(I_j)$, $j = 1,..., N$. The nodal shape functions $\psi_{-1,j}$ are augmented for $p_j \geqslant 2$ by *internal, upwinded shape functions* $\psi_{i,j} \in \left(H^2 \cap H_0^1\right)(I_j)$. They are defined by

$$\begin{aligned} L_\varepsilon^* \psi_{i,j} &= L_i \left(2\frac{x - m_j}{h_j}\right) \text{ in } I_j, \quad i = 0,..., p_j - 2, \ j = 1,..., N, \\ \psi_{i,j} &= 0 \quad \text{elsewhere.} \end{aligned} \tag{3.8}$$

Any $v \in S_L^{\vec{p}}$ can be written as

$$v(x) = \sum_{j=2}^{N} v(x_{j-1}) \psi_{-1,j}(x) + \sum_{j=1}^{N} \sum_{i=0}^{p_j - 2} b_{ij} \psi_{i,j}(x), \tag{3.9}$$

where b_{ij} are the Legendre coefficients of $L_\varepsilon^* v\big|_{I_j}$. Further, from the definition (3.7) of the $\psi_{-1,j}$ we have

$$L_\varepsilon^* v\big|_{I_j} = \sum_{i=0}^{p_j - 2} b_{ij} L_i \left(2\frac{x - m_j}{h_j}\right), \quad j = 1,..., N,$$

which yields with the orthogonality properties of the Legendre polynomials and a scaling argument

$$\sum_{j=1}^{N} \left\| L_\varepsilon^* v \right\|_{L^2(I_j)}^2 = \sum_{j=1}^{N} h_j \sum_{i=0}^{p_j - 2} \frac{|b_{ij}|^2}{2i + 1} =: \sum_{j=1}^{N} h_j S_j. \tag{3.10}$$

Combining (3.10) with (2.8), we obtain for $v \in S_L^{\vec{p}}$ an expression for $\|v\|_{H_T^2}$ in terms of the b_{ij}:

$$\|v\|_{H_T^2} = \left(\sum_{j=1}^{N} h_j \sum_{i=0}^{p_j - 2} \frac{|b_{ij}|^2}{2i + 1} + \sum_{j=1}^{N-1} \frac{\left|\left[\varepsilon v'(x_j)\right]\right|^2}{\rho_j} \right)^{1/2} \quad \forall v \in S_L^{\vec{p}}. \tag{3.11}$$

Writing $v(x)$ in the form (3.9) and $u_v(x)$ as in (3.6) and inserting into (2.7), we find in the same way

$$B_{\mathcal{T}}(u_v, v) = \sum_{j=1}^{N} h_j \sum_{i=0}^{p_j-2} \frac{a_{ij} b_{ij}}{2i+1} - \sum_{j=1}^{N-1} u_v(x_j) \left[\varepsilon v'(x_j)\right]. \tag{3.12}$$

For given $v \in S_L^{\vec{p}}$, i.e. for given b_{ij}, we choose now a_{ij} as follows: First, we select

$$a_{ij} = b_{ij}, \quad i = 0,..., p_j - 2, \tag{3.13}$$

which leaves $a_{p_j-1,j}, a_{p_j,j}$ to be determined, for each I_j. Since u_v must be continuous, two conditions per interval must be enforced. We prescribe u_v at each endpoint of I_j as follows (by $u_v(x^{\pm})$ we denote the right/left limit of u_v at x):

$$u_v(x_{j-1}^+) = a_j^- := \begin{cases} -\left[\varepsilon v'(x_{j-1})\right]/\rho_{j-1} & \text{if } j > 1, \\ 0 & \text{if } j = 1 \end{cases} \tag{3.14}$$

and

$$u_v(x_j^-) = a_j^+ := \begin{cases} -\left[\varepsilon v'(x_j)\right]/\rho_j & \text{if } j < N, \\ 0 & \text{if } j = N. \end{cases} \tag{3.15}$$

Conditions (3.14), (3.15) ensure continuity of u_v. Since $L_i(\pm 1) = (\pm 1)^i$ implies

$$u_v(x_{j-1}^+) = \sum_{i=0}^{p_j} (-1)^i a_{ij}, \quad u_v(x_j^-) = \sum_{i=0}^{p_j} a_{ij}$$

we get with (3.13) the linear system

$$\begin{bmatrix} (-1)^{p_j-1} & (-1)^{p_j} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a_{p_j-1,j} \\ a_{p_j,j} \end{bmatrix} = \begin{bmatrix} a_j^- - \sum_{i=0}^{p_j-2} (-1)^i b_{ij} \\ a_j^+ - \sum_{i=0}^{p_j-2} b_{ij} \end{bmatrix}. \tag{3.16}$$

Its determinant is nonzero for any p_j, therefore u_v is uniquely determined by (3.13) and (3.16).

From (3.12), (3.13) and (3.14), (3.15) we get

$$B_{\mathcal{T}}(u_v, v) = \|v\|_{H^2_{\mathcal{T}}}^2.$$

It remains therefore to show

$$\|u_v\|_{H^0_{\mathcal{T}}} \leqslant \gamma_M \|v\|_{H^2_{\mathcal{T}}} \tag{3.17}$$

with γ_M as in (3.5).

Since u_v is continuous, we have

$$\|u_v\|_{H^0_{\mathcal{T}}}^2 = \sum_{j=1}^{N} \|u_v\|_{L^2(I_j)}^2 + \sum_{j=1}^{N-1} \rho_j \left|u_v(x_j)\right|^2 = \sum_{j=1}^{N} h_j \sum_{i=0}^{p_j} \frac{\left|a_{ij}\right|^2}{2i+1} + \sum_{j=1}^{N-1} \rho_j \left|a_j^+\right|^2$$

$$= \sum_{j=1}^{N} \left\{ h_j \sum_{i=0}^{p_j-2} \frac{|b_{ij}|^2}{2i+1} + h_j \sum_{i=p_j-1}^{p_j} \frac{|a_{ij}|^2}{2i+1} \right\} + \sum_{j=1}^{N-1} \rho_j^{-1} \left| \left[\varepsilon v'(x_j) \right] \right|^2$$

$$= \|v\|_{H_T^2}^2 + \sum_{j=1}^{N} h_j \sum_{i=p_j-1}^{p_j} \frac{|a_{ij}|^2}{2i+1}. \tag{3.18}$$

We estimate $|a_{ij}|^2$ for $i = p_j - 1, p_j$. From (3.16), we get

$$\begin{bmatrix} a_{p_j-1,j} \\ a_{p_j,j} \end{bmatrix} = \frac{-1}{2(-1)^{p_j}} \begin{pmatrix} 1 & (-1)^{p_j-1} \\ -1 & (-1)^{p_j-1} \end{pmatrix} \begin{pmatrix} a_j^- - \sum_{i=0}^{p_j-2} (-1)^i b_{ij} \\ a_j^+ - \sum_{i=0}^{p_j-2} b_{ij} \end{pmatrix}$$

$$= \frac{-1}{2(-1)^{p_j}} \begin{pmatrix} a_j^- + (-1)^{p_j-1} a_j^+ + \sum_{i=0}^{p_j-2} b_{ij}((-1)^{p_j} - (-1)^i) \\ -a_j^- + (-1)^{p_j-1} a_j^+ + \sum_{i=0}^{p_j-2} b_{ij}((-1)^i + (-1)^{p_j}) \end{pmatrix}.$$

We estimate

$$\max\{|a_{ij}| : i = p_j - 1, p_j\} \leqslant \frac{1}{2} \left(\left|a_j^-\right| + \left|a_j^+\right| \right) + \sum_{i=0}^{p_j-2} |b_{ij}|$$

and get with (3.14), (3.15) that

$$\max\{|a_{ij}|^2 : i = p_j - 1, p_j\} \leqslant \left(\frac{\left|\left[\varepsilon v'(x_{j-1})\right]\right|^2}{\rho_{j-1}^2} + \frac{\left|\left[\varepsilon v'(x_j)\right]\right|^2}{\rho_j^2} \right) + 2 \left(\sum_{i=0}^{p_j-2} |b_{ij}| \right)^2.$$

With the understanding that $\left[v'(x_0)\right] = \left[v'(x_N)\right] = 0$ and $\rho_0 = \rho_N = \infty$ we estimate further

$$\sum_{j=1}^{N} h_j \sum_{i=p_j-1}^{p_j} \frac{|a_{ij}|^2}{2i+1} \leqslant \sum_{j=1}^{N} \frac{h_j}{2p_j - 1} \left\{ \frac{\left|\left[\varepsilon v'(x_{j-1})\right]\right|^2}{\rho_{j-1}^2} + \frac{\left|\left[\varepsilon v'(x_j)\right]\right|^2}{\rho_j^2} + 2(p_j - 1)^2 \sum_{i=0}^{p_j-2} \frac{|b_{ij}|^2}{2i+1} \right\}.$$

Using now $h_j/\rho_j \leqslant 2$, $h_j/\rho_{j-1} \leqslant 2$ and

$$\frac{2(p_j-1)^2}{2p_j-1} = \frac{2(p_j-1)^2}{2(p_j-1)+1} \leqslant \frac{2(p_j-1)^2}{2(p_j-1)} = p_j - 1 \leqslant p - 1$$

we arrive at

$$\sum_{j=1}^{N} h_j \sum_{i=p_j-1}^{p_j} \frac{|a_{ij}|^2}{2i+1} \leqslant 4 \sum_{j=1}^{N-1} \frac{\left|\left[\varepsilon v'(x_j)\right]\right|^2}{\rho_j} + (p-1) \sum_{j=1}^{N} h_j \sum_{i=0}^{p_j-2} \frac{|b_{ij}|^2}{2i+1} \tag{3.19}$$

$$\leqslant \max\{4, p-1\} \|v\|_{H_T^2}^2,$$

where we used (2.8) and (3.10). Referring to (3.18) completes the proof. □

3.3 *Consistency and convergence*

As $\dim S_0^{\vec{p}} = \dim S_L^{\vec{p}} < \infty$, Theorem 3.1 and Remark 2.2 imply that the following two conditions hold:

$$\inf_{0\neq u\in S_0^{\vec{p}}} \sup_{0\neq v\in S_L^{\vec{p}}} \frac{B_{\mathcal{T}}(u,v)}{\|u\|_{H^0_{\mathcal{T}}}\,\|v\|_{H^2_{\mathcal{T}}}} \geqslant \frac{1}{\gamma_M}, \tag{3.20}$$

$$\forall 0\neq v\in S_L^{\vec{p}}:\ \sup_{u\in S_0^{\vec{p}}} B_{\mathcal{T}}(u,v) > 0. \tag{3.21}$$

We deduce therefore from (3.20), (3.21) that for every mesh-degree combination $(\vec{p},\mathcal{T})$ there exists a unique FE solution u_M of (3.4). In particular, the $M\times M$ (generally nonsymmetric) stiffness matrix corresponding to (3.4) is nonsingular. Moreover, the FE solution u_M is quasi-optimal, i.e.

$$\|u-u_M\|_{H^0_{\mathcal{T}}} \leqslant (1+\gamma_M)\,\|u-w\|_{H^0_{\mathcal{T}}} \qquad \forall w\in S_0^{\vec{p}}. \tag{3.22}$$

The rate of convergence of the FEM (3.4) is therefore determined by the approximability of the exact solution u from the trial space $S_0^{\vec{p}}$. We show that proper selection of the mesh $\mathcal{T}$ and of the polynomial degree distribution $\vec{p}$ yields an *exponential rate of convergence*, uniform in ε. We will consider the approximation of two types of solutions u_ε. In Theorem 3.3, we consider the case of analytic right-hand sides f. In that case, the solution u_ε exhibits only a boundary layer at the outflow boundary, and we will show that the 'two-element' mesh of Schwab & Suri (1996) with one small element in the outflow boundary layer leads to robust exponential convergence of our hp FEM. We also analyse the approximation of solutions stemming from right-hand sides that contain Dirac distributions (note that such right-hand sides are admissible by Remark 2.4). Such solutions exhibit additionally internal layers akin to viscous shock profiles. We show in Theorem 3.5 that inserting a small element of appropriate size in each internal layer allows for the resolution of these internal layers.

THEOREM 3.3 Let u_ε be the solution of (1.1), (2.1), and assume that the coefficients a, b and the right-hand side f satisfy (1.3), (1.5)–(1.7). For every $\varepsilon,\kappa>0$ let the degree vector $\vec{p}$ and the mesh $\mathcal{T}=\mathcal{T}_{\kappa,\varepsilon}$ be given by

$$\begin{aligned} &\vec{p}=\{p,p\}, \quad \mathcal{T}_{\kappa,\varepsilon}=\{I_1,I_2\} \quad \text{if } \kappa p\varepsilon<1,\\ &\vec{p}=\{p\}, \qquad \mathcal{T}_{\kappa,\varepsilon}=\{\Omega\} \qquad \text{if } \kappa p\varepsilon\geqslant 1, \end{aligned} \tag{3.23}$$

where

$$I_1=(-1,1-\kappa p\varepsilon), \quad I_2=(1-\kappa p\varepsilon,1).$$

Then there is a constant κ_0 depending only on the constants of (1.3), (1.5)–(1.7) such that for every $0<\kappa<\kappa_0$ there are $C,\sigma>0$ independent of p and ε such that

$$\inf_{v_p\in S_0^{\vec{p},1}(\mathcal{T}_{\kappa,\varepsilon})} \|u_\varepsilon - v_p\|_{L^\infty(\Omega)} \leqslant C\mathrm{e}^{-\sigma p} \qquad \forall p\in\mathbb{N}. \tag{3.24}$$

Proof. The proof is very similar to that of Theorem 16 of Melenk (1997). □

As $\|u_\varepsilon - v_p\|^2_{H^0_{\mathcal{T}}} \leqslant 3\|u_\varepsilon - v_p\|^2_{L^\infty(\Omega)}$, Theorem 3.3 together with (3.22) shows that for analytic input data robust exponential convergence can be achieved by the FE scheme (3.4) provided the space $S^{\vec{p}}_0$ is designed properly (i.e. with one element of size $O(p\varepsilon)$ in the outflow boundary layer) and provided that the corresponding stable test space $S^{\vec{p}}_L(\mathcal{T})$ is available. Results analogous to Theorem 3.3 also hold true when f is piecewise analytic on $[-1, 1]$; then, however, additional internal layers arise at points of nonanalyticity of f which must be accounted for by adding further $O(\varepsilon p)$ elements. Theorem 3.5 below illustrates this point.

REMARK 3.4 An estimate on the value of the constant κ_0 is in principle available from the proof of Theorem 3.3. For *constant* coefficients a, b, the value of κ_0 can be determined explicitly as was done by Schwab & Suri (1996): $\kappa_0 = 4/(e\lambda)$ where $\lambda = (a+\sqrt{a^2 + 4b\varepsilon})/2 \geqslant a/2$ by assumption (1.3). The numerical experiments of Schwab & Suri (1996) show moreover that the approximation properties of piecewise polynomials on the meshes $\mathcal{T}_{\kappa,\varepsilon}$ are fairly insensitive to the precise choice of κ.

We conclude this section by demonstrating that the ideas behind the 'two-element' mesh can also be used for the scale resolution of viscous shock profiles. To that end, let us consider

$$L_\varepsilon u_\varepsilon = f + \delta_0 \quad \text{on } \Omega, \qquad u_\varepsilon(\pm 1) = \alpha^\pm \tag{3.25}$$

where δ_0 denotes the Dirac distribution concentrated at the point $x = 0$. The following analogue of Theorem 3.3 holds.

THEOREM 3.5 Let u_ε be the solution of (3.25) and assume that the coefficients a, b and the right-hand side f satisfy (1.3), (1.5)–(1.7). For every $\varepsilon, \kappa > 0$ let the degree vector $\vec{p}$ and the 'four-element' mesh $\mathcal{T} = \mathcal{T}^4_{\kappa,\varepsilon}$ be given by

$$\begin{aligned} &\vec{p} = \{p, p, p, p\}, \quad \mathcal{T}^4_{\kappa,\varepsilon} = \{I_1, I_2, I_3, I_4\} \quad \text{if } \kappa p\varepsilon < 1/2, \\ &\vec{p} = \{p\}, \qquad\qquad \mathcal{T}^4_{\kappa,\varepsilon} = \{\Omega\} \qquad\qquad \text{if } \kappa p\varepsilon \geqslant 1/2, \end{aligned} \tag{3.26}$$

where

$$I_1 = (-1, -\kappa p\varepsilon), \quad I_2 = (-\kappa p\varepsilon, 0), \quad I_3 = (0, 1 - \kappa p\varepsilon), \quad I_4 = (1 - \kappa p\varepsilon, 1).$$

Then there are $\varepsilon_0, \kappa_0 > 0$ depending only on the constants of (1.3), (1.5)–(1.7) such that for all $0 < \varepsilon \leqslant \varepsilon_0$ the following holds: For each $0 < \kappa < \kappa_0$ there are $C, \sigma > 0$ (independent of p and ε) such that

$$\inf_{v_p \in S^{\vec{p},1}_0(\mathcal{T}^4_{\kappa,\varepsilon})} \|u_\varepsilon - v_p\|_{L^\infty(\Omega)} \leqslant C\mathrm{e}^{-\sigma p} \qquad \forall p \in \mathbb{N}.$$

Proof. It is straightforward to construct a function $u_\delta \in C[-1, 1] \cap H^1_0(-1, 1)$ with the properties that $L_\varepsilon u_\delta = \delta_0$ in the distributional sense and $\|u_\delta\|_{L^\infty(-1,1)} \leqslant C$, where $C > 0$ is independent of ε.

By superposition, the solution u_ε of (3.25) can be written as

$$u_\varepsilon = u_\delta + \tilde{u}_\varepsilon$$

where $\tilde{u}_\varepsilon$ solves the auxiliary problem

$$L_\varepsilon \tilde{u}_\varepsilon = f \quad \text{on } \Omega, \qquad \tilde{u}_\varepsilon(-1) = \alpha^-, \quad \tilde{u}_\varepsilon(1) = \alpha^+. \tag{3.27}$$

It is now easy to see that for the solution u_ε restricted to the subintervals $(-1, 0)$, $(0, 1)$, the analytic regularity assertions of Theorem 1.1 hold true. Hence, by Theorem 3.3, the functions $u_\varepsilon|_{(-1,0)}$, $u_\varepsilon|_{(0,1)}$ can be approximated with the desired exponential accuracy on a two-element mesh on $(-1, 0)$ and $(0, 1)$. Such two-element meshes are contained in the 'four-element' mesh considered here. □

4. Approximate test functions

Theorem 3.1 shows that the use of the upwinded test space $S_L^{\vec{p}}$ of (3.2) gives rise to a stable numerical scheme. Unfortunately, however, the shape functions $\psi_{k,j}$ in (3.7), (3.8) are themselves solutions of (local) convection–diffusion problems. For the case $p = 1$ and constant coefficients a, b, these upwinded test functions can be computed explicitly and lead to the so-called 'Hemker test functions' of Hemker (1977). For non-constant coefficients, however, they are not explicitly available. We show therefore now that stability can be retained even if the $\psi_{k,j}$ are known only approximately. The perturbation analysis of Section 4.1 shows that fairly weak accuracy requirements on the test functions ψ_{ij} suffice to ensure stability of the FEM. Especially for low p, rather 'crude' approximations to the L-splines ψ_{ij} are sufficient; this is the reason why techniques such as the α-quadratic upwinding of Christie *et al* (1978) (see also Morton (1995) for an up-to-date account on these methods) and the use of Hemker test functions in De Groen (1978), De Groen & Hemker (1979) obtained by freezing coefficients lead to stable FEM. All these methods are in fact covered by our perturbation analysis.

4.1 *Stability with approximate test functions*

The functions $\psi_{k,j}$ are solutions of local Dirichlet problems. Let us assume that approximate test functions $\tilde{\psi}_{k,j} \in H_0^1(\Omega) \cap \prod_{i=1}^N H^2(I_i)$ are given (one possible scheme for obtaining such approximate test functions will be described in Section 4.2 below). For such approximate test functions $\tilde{\psi}_{k,j}$ we then introduce the approximate test space

$$\tilde{S}_L^{\vec{p}} = \operatorname{span}\left\{\tilde{\psi}_{k,j} : k = -1,\ j = 2,\ldots, N \text{ and } k = 0,\ldots, p_j - 2,\ j = 1,\ldots, N\right\}. \tag{4.1}$$

Let us assume that these approximate test functions $\tilde{\psi}_{k,j}$ satisfy for some *residuals* $\eta_{k,j} \in L^2(\Omega)$ the following equations:

$$\begin{aligned} L_\varepsilon^* \tilde{\psi}_{-1,j} &= \eta_{-1,j} && \text{in } I_{j-1} \cup I_j, \quad j = 2,\ldots, N, \\ \tilde{\psi}_{-1,j} &= 0 && \text{elsewhere}, \\ \tilde{\psi}_{-1,j}(x_k) &= \delta_{j,k+1}, && k = 1,\ldots, N-1, \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} L_\varepsilon^* \tilde{\psi}_{i,j} &= L_i\left(2\frac{x-m_j}{h_j}\right) + \eta_{i,j} \quad \text{in } I_j,\ i=0,\ldots,p_j-2,\ j=1,\ldots,N,\\ \tilde{\psi}_{i,j} &= 0 \quad \text{elsewhere.} \end{aligned} \tag{4.3}$$

For such approximate test spaces $\tilde{S}_L^{\vec{p}}$, the FEM reads: Find $u_M \in S_0^{\vec{p}}$ such that

$$B_{\mathcal{T}}(u_M, v) = F(v) \qquad \forall v \in \tilde{S}_L^{\vec{p}}. \tag{4.4}$$

We show first that the bilinear form $B_{\mathcal{T}}(\cdot,\cdot)$ is stable on $S_0^{\vec{p}} \times \tilde{S}_L^{\vec{p}}$ provided the residuals $\eta_{i,j}$ in (4.2), (4.3) are sufficiently small.

THEOREM 4.1 Assume that the approximate test functions $\tilde{\psi}_{k,j}$ satisfy (4.2), (4.3) with $\eta_{i,j}$ such that

$$\Lambda_1 \leqslant \frac{1}{20C_1^2}\gamma_M^{-2}, \quad \Lambda_2 \leqslant \frac{1}{20}\gamma_M^{-2} \tag{4.5}$$

where

$$\Lambda_1 := \sum_{j=1}^{N} \left\|\eta_{-1,j}\right\|_{L^2(I_j)}^2 + \left\|\eta_{-1,j+1}\right\|_{L^2(I_j)}^2, \tag{4.6}$$

$$\Lambda_2 := \max_{1\leqslant j\leqslant N}\left\{h_j^{-1}\sum_{i=0}^{p_j-2}(2i+1)\left\|\eta_{ij}\right\|_{L^2(I_j)}^2\right\} \tag{4.7}$$

(we set $\eta_{-1,1} = \eta_{-1,N+1} = 0$ and $\Lambda_2 := 0$ if $p = 1$; C_1 is the constant in (2.15)).

Then there exists $C > 0$ independent of ε, $\vec{p}$ and $\mathcal{T}$ such that

$$\inf_{0\neq v\in\tilde{S}_L^{\vec{p}}}\sup_{0\neq u\in S_0^{\vec{p}}}\frac{B_{\mathcal{T}}(u,v)}{\|u\|_{H_{\mathcal{T}}^0}\|v\|_{H_{\mathcal{T}}^2}} \geqslant \frac{C}{\gamma_M} > 0. \tag{4.8}$$

Proof. Let $\tilde{v} \in \tilde{S}_L^{\vec{p}}$ be given. It can be written as

$$\tilde{v}(x) = \sum_{j=2}^{N}\tilde{v}(x_{j-1})\tilde{\psi}_{-1,j}(x) + \sum_{j=1}^{N}\sum_{i=0}^{p_j-2} b_{ij}\tilde{\psi}_{ij}(x).$$

We select $u_{\tilde{v}}$ as in the proof of Theorem 3.1, i.e.

$$u_{\tilde{v}}|_{I_j} = \sum_{i=0}^{p_j} a_{ij} L_i\left(2\frac{x-m_j}{h_j}\right)$$

where

$$a_{ij} = b_{ij}, \qquad i = 0,\ldots,p_j-2,$$

and $a_{p_j-1,j}, a_{p_j,j}$ are selected as in (3.14)–(3.16) with $\tilde{v}$ in place of v. Then $u_{\tilde{v}}$ is continuous on $[-1, 1]$. With the test functions ψ_{ij} in (3.7), (3.8) we define also

$$v(x) := \sum_{j=2}^{N} \tilde{v}(x_{j-1})\psi_{-1,j}(x) + \sum_{j=1}^{N} \sum_{i=0}^{p_j-2} b_{ij}\psi_{ij}(x)$$

and we set $\delta v := \tilde{v} - v$. Then

$$B_{\mathcal{T}}(u_{\tilde{v}}, \tilde{v}) = \sum_{j=1}^{N} \int_{I_j} u_{\tilde{v}} L_\varepsilon^* v \,\mathrm{d}x - \sum_{j=1}^{N-1} u_{\tilde{v}}(x_j)\left[\varepsilon \tilde{v}'(x_j)\right] + \sum_{j=1}^{N} \int_{I_j} u_{\tilde{v}} L_\varepsilon^* \delta v \,\mathrm{d}x.$$

We calculate

$$\int_{I_j} u_{\tilde{v}} L_\varepsilon^* v \,\mathrm{d}x = h_j \sum_{i=0}^{p_j-2} \frac{|b_{ij}|^2}{2i+1} =: h_j S_j \tag{4.9}$$

and by (3.14), (3.15),

$$-\sum_{j=1}^{N-1} u_{\tilde{v}}(x_j)\left[\varepsilon \tilde{v}'(x_j)\right] = \sum_{j=1}^{N-1} \rho_j^{-1} \left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2.$$

Hence we find upon setting

$$S := \left\{ \sum_{j=1}^{N} h_j S_j + \sum_{j=1}^{N-1} \rho_j^{-1} \left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2 \right\} \tag{4.10}$$

that

$$B_{\mathcal{T}}(u_{\tilde{v}}, \tilde{v}) \geqslant S - \sum_{j=1}^{N} \|u_{\tilde{v}}\|_{L^2(I_j)} \left\| L_\varepsilon^* \delta v \right\|_{L^2(I_j)}. \tag{4.11}$$

Now

$$\|u_{\tilde{v}}\|^2_{L^2(I_j)} = h_j \sum_{i=0}^{p_j-2} \frac{|b_{ij}|^2}{2i+1} + h_j \sum_{i=p_j-1}^{p_j} \frac{|a_{ij}|^2}{2i+1}.$$

Using (3.19), we then find

$$\sum_{j=1}^{N} \|u_{\tilde{v}}\|^2_{L^2(I_j)} \leqslant 4 \sum_{j=1}^{N-1} \frac{\left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2}{\rho_j} + p \sum_{j=1}^{N} h_j S_j, \tag{4.12}$$

and therefore in particular

$$\|u_{\tilde{v}}\|^2_{H^0_{\mathcal{T}}} \leqslant \sum_{j=1}^{N} \|u_{\tilde{v}}\|^2_{L^2(I_j)} + \sum_{j=1}^{N-1} \rho_j^{-1} \left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2 \leqslant \gamma_M^2 S. \tag{4.13}$$

Consider now $\|L_\varepsilon^* \delta v\|_{L^2(I_j)}$. We have by (4.2), (4.3)

$$\left(L_\varepsilon^* \delta v\right)\big|_{I_j} = \tilde{v}(x_{j-1})\eta_{-1,j} + \tilde{v}(x_j)\eta_{-1,j+1} + \sum_{i=0}^{p_j-2} b_{ij}\eta_{ij}.$$

This leads to

$$\begin{aligned}\|L_\varepsilon^* \delta v\|_{L^2(I_j)} &\leqslant \|\tilde{v}\|_{L^\infty}\left(\|\eta_{-1,j}\|_{L^2(I_j)} + \|\eta_{-1,j+1}\|_{L^2(I_j)}\right) + \sum_{i=0}^{p_j-2} |b_{ij}|\,\|\eta_{ij}\|_{L^2(I_j)} \\ &\leqslant \|\tilde{v}\|_{L^\infty}\left(\|\eta_{-1,j}\|_{L^2(I_j)} + \|\eta_{-1,j+1}\|_{L^2(I_j)}\right) \\ &\quad + \left(h_j S_j\right)^{1/2}\left(h_j^{-1}\sum_{i=0}^{p_j-2}(2i+1)\,\|\eta_{ij}\|_{L^2(I_j)}^2\right)^{1/2}\end{aligned}$$

i.e.

$$\|L_\varepsilon^* \delta v\|_{L^2(I_j)}^2 \leqslant 4\,\|\tilde{v}\|_{L^\infty}^2\,\Lambda_{1j} + 2h_j S_j \Lambda_{2j}, \qquad j = 1,\ldots, N, \tag{4.14}$$

where we defined

$$\begin{aligned}\Lambda_{1j} &:= \|\eta_{-1,j}\|_{L^2(I_j)}^2 + \|\eta_{-1,j+1}\|_{L^2(I_j)}^2, \\ \Lambda_{2j} &:= h_j^{-1}\sum_{i=0}^{p_j-2}(2i+1)\,\|\eta_{ij}\|_{L^2(I_j)}^2.\end{aligned}$$

Hence we may estimate with (4.12) and the definition of S in (4.10)

$$\begin{aligned}&\sum_{j=1}^{N}\|u_{\tilde{v}}\|_{L^2(I_j)}\,\|L_\varepsilon^*\delta v\|_{L^2(I_j)} \\ &\quad\leqslant \left(4\sum_{j=1}^{N-1}\frac{\left|\left[\varepsilon\tilde{v}'(x_j)\right]\right|^2}{\rho_j} + p\sum_{j=1}^{N}h_j S_j\right)^{1/2}\left(4\Lambda_1\,\|\tilde{v}\|_{L^\infty}^2 + 2\Lambda_2\sum_{j=1}^{N}h_j S_j\right)^{1/2} \\ &\quad\leqslant \max\{4,\,p\}^{1/2}S^{1/2}\left(4\Lambda_1\,\|\tilde{v}\|_{L^\infty}^2 + 2\Lambda_2\sum_{j=1}^{N}h_j S_j\right)^{1/2}.\end{aligned} \tag{4.15}$$

With (4.14), the embedding (2.15) and the definition of the $H_{\mathcal{T}}^2$ norm we get further

$$\begin{aligned}\|\tilde{v}\|_{H_{\mathcal{T}}^2}^2 &\leqslant 2\sum_{j=1}^{N}\left(\|L_\varepsilon^* v\|_{L^2(I_j)}^2 + \|L_\varepsilon^*\delta v\|_{L^2(I_j)}^2\right) + \sum_{j=1}^{N-1}\frac{\left|\left[\varepsilon\tilde{v}'(x_j)\right]\right|^2}{\rho_j} \\ &\leqslant 2(1+2\Lambda_2)\sum_{j=1}^{N}h_j S_j + 8C_1^2\Lambda_1\,\|\tilde{v}\|_{H_{\mathcal{T}}^2}^2 + \sum_{j=1}^{N-1}\frac{\left|\left[\varepsilon\tilde{v}'(x_j)\right]\right|^2}{\rho_j}.\end{aligned}$$

After regrouping terms, it follows that

$$S = \sum_{j=1}^{N}h_j S_j + \sum_{j=1}^{N-1}\frac{\left|\left[\varepsilon\tilde{v}'(x_j)\right]\right|^2}{\rho_j} \geqslant D(\eta)\,\|\tilde{v}\|_{H_{\mathcal{T}}^2}^2, \tag{4.16}$$

where, by (4.5) and the fact that $\gamma_M^2 \geqslant 5$

$$D(\eta) := \frac{1 - 8C_1^2\Lambda_1}{2(1+2\Lambda_2)} \geqslant \frac{1 - 8/100}{2(1+2/100)} = \frac{23}{51} > 0.$$

The inequality which is converse to (4.16) also holds. To obtain it, we proceed as follows: Using the estimate $\|a+b\|^2 \geqslant \frac{1}{2}\|a\|^2 - \|b\|^2$, which is valid for any norm $\|\cdot\|$, we get

$$\begin{aligned}\|\tilde{v}\|_{H^2_{\mathcal{T}}}^2 &= \sum_{j=1}^{N} \left\| L_\varepsilon^* v + L_\varepsilon^* \delta v \right\|_{L^2(I_j)}^2 + \sum_{j=1}^{N-1} \frac{\left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2}{\rho_j} \\ &\geqslant \left(\tfrac{1}{2} - 2\Lambda_2\right) \sum_{j=1}^{N} h_j S_j + \sum_{j=1}^{N-1} \frac{\left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2}{\rho_j} - 4C_1^2\Lambda_1 \|\tilde{v}\|_{H^2_{\mathcal{T}}}^2\end{aligned}$$

and obtain after rearranging terms

$$S = \sum_{j=1}^{N} h_j S_j + \sum_{j=1}^{N-1} \frac{\left|\left[\varepsilon \tilde{v}'(x_j)\right]\right|^2}{\rho_j} \leqslant C(\eta) \|\tilde{v}\|_{H^2_{\mathcal{T}}}^2 \tag{4.17}$$

where

$$C(\eta) := \frac{1 + 4C_1^2\Lambda_1}{1/2 - 2\Lambda_2} \leqslant \frac{13}{6}.$$

Finally, (4.11) and (4.15) imply with (4.13) and (4.16), (4.17) the bound

$$\begin{aligned}B_{\mathcal{T}}(u_{\tilde{v}}, \tilde{v}) &\geqslant S - \max\{4, p\}^{1/2} S^{1/2} \left(4\Lambda_1 \|\tilde{v}\|_{L^\infty}^2 + 2\Lambda_2 S\right)^{1/2} \\ &\geqslant S^{1/2} \left[S^{1/2} - \gamma_M \left(4\Lambda_1 C_1^2 + 2\Lambda_2 C(\eta)\right)^{1/2} \|\tilde{v}\|_{H^2_{\mathcal{T}}} \right] \\ &\geqslant \gamma_M^{-1} \|u_{\tilde{v}}\|_{H^0_{\mathcal{T}}} \left[\sqrt{D(\eta)} - \gamma_M \left(4\Lambda_1 C_1^2 + 2\Lambda_2 C(\eta)\right)^{1/2} \right] \|\tilde{v}\|_{H^2_{\mathcal{T}}}.\end{aligned}$$

It is easy to see now that the expression in brackets can be bounded from below by $\sqrt{23/51} - \sqrt{5/12} > 0$ which concludes the proof of the inf-sup condition (4.8). □

REMARK 4.2 The test functions $\tilde{\psi}_{ij}$ in (4.2), (4.3) are *conforming*, i.e., globally in $H_0^1(\Omega)$ and elementwise in $H^2(I_j)$. As we shall show shortly, it is possible to obtain numerical approximations $\tilde{\psi}_{ij}$ by solving the problems (3.7), (3.8) with a least-squares FEM on a subgrid $\tilde{\mathcal{T}}$ of $\mathcal{T}$. The assumption $\tilde{\psi}_{ij} \in H^2(I_j)$ for $I_j \in \mathcal{T}$ then implies that the least-squares FEM must be *locally* C^1 conforming. Although this can be achieved, we can also admit C^0-approximations $\tilde{\psi}_{ij}$ in Theorem 4.1, if we penalize the flux-jumps of $\tilde{\psi}_{ij}$ on the subgrid appropriately. This will complicate the analysis, but does not pose any essential difficulties.

The stability (4.8) together with the fact that the perturbed test functions are $H^2_{\mathcal{T}}$-conforming and the approximation property Theorem 3.3 imply the following convergence result.

THEOREM 4.3 For any mesh $\mathcal{T}$ the *hp* FE solution $\tilde{u}_M \in S_0^{\vec{p}}(\mathcal{T})$ in (4.4) corresponding to the approximate test space $\tilde{S}_L^{\vec{p}}$ defined in (4.1)–(4.3) and satisfying (4.5), exists and is quasi-optimal, i.e., with C, γ_M of (4.8)

$$\|u - \tilde{u}_M\|_{H^0_{\mathcal{T}}} \leqslant (1 + \gamma_M/C)\, \|u - v\|_{H^0_{\mathcal{T}}} \qquad \forall v \in S_0^{\vec{p}}. \tag{4.18}$$

In particular, if the coefficients a, b, and the right-hand side f are analytic and satisfy (1.3), (1.5)–(1.7) and the mesh $\mathcal{T} = \mathcal{T}_{\kappa,\varepsilon}$ is chosen as in (3.23) with κ sufficiently small, we have robust exponential convergence, i.e.,

$$\|u - \tilde{u}_M\|_{H^0_{\mathcal{T}}} \leqslant C \exp(-\theta M) \tag{4.19}$$

where $C, \theta > 0$ are independent of ε, p.

REMARK 4.4 The meshes $\mathcal{T}_{\kappa,\varepsilon}$ considered in Theorem 3.3 are essentially the 'minimal' meshes that can resolve the boundary layer behaviour of the solution u_ε in a p-version setting at a robust exponential rate. Clearly, the approximation results of the form (3.24) hold true for any mesh $\mathcal{T}$ that contains one small element of size $O(\varepsilon p)$ at the outflow boundary, i.e., if $\mathcal{T}_{\kappa,\varepsilon} \subset \mathcal{T}$. Due to the quasi-optimality (4.18), error estimates analogous to (4.19) hold for all meshes $\mathcal{T}$ with $\mathcal{T}_{\kappa,\varepsilon} \subset \mathcal{T}$. These 'minimal' meshes $\mathcal{T}_{\kappa,\varepsilon}$ depend on the polynomial degree p. In practice, it may be more convenient to fix a mesh $\mathcal{T}$ and then increase p until the desired accuracy is reached. For example, piecewise polynomials on a mesh which is graded geometrically towards the outflow boundary have approximation properties similar to the minimal meshes $\mathcal{T}_{\kappa,\varepsilon}$ provided that the smallest element of the geometric mesh is of size $O(\varepsilon)$ (see Melenk (1997)).

4.2 *Computation of the approximate test functions*

To obtain approximate test functions $\tilde{\psi}_{ij}$, many strategies are possible: classical approaches use approximate analytical expressions (e.g., the Hemker test functions or the α-quadratic upwinding mentioned above) or asymptotic expansions (as was done by Szymczak & Babuška (1984)). These semianalytical approaches work well for low-order methods; to accommodate high p together with arbitrary meshes, a fully numerical method for the computation of the test functions seems to be desirable.

Here we propose a *local least-squares FEM* to approximate the ψ_{ij} stably and completely computationally. The approach allows moreover for controlling the quantities Λ_1, Λ_2 in (4.6), (4.7) *a posteriori*.

The plan for the remainder of this section is as follows. In Section 4.2.1 we define the local least-squares problems, which yield the approximate test functions $\tilde{\psi}_{-1,j}$, $\tilde{\psi}_{ij}$, as minimization problems of appropriate quadratic functionals over finite-dimensional spaces $\mathcal{A}_j^{hq}$. In this framework, the choice of the spaces $\mathcal{A}_j^{hq}$ determines completely the test functions $\tilde{\psi}_{-1,j}$, $\tilde{\psi}_{ij}$ and thus the method (4.4). The exact test functions are also solutions of singularly perturbed convection–diffusion equations with analytic coefficients. We will therefore choose $\mathcal{A}_j^{hq}$ as spaces of piecewise polynomials of degree q on a two-element mesh (one small element at the outflow boundary of the local problem and one large element) in complete analogy to our approximation theory for the global solution u_ε. An approximation result for this choice is presented in Section 4.2.2.

Other choices of the spaces $\mathcal{A}_j^{hq}$ lead to different methods. For example, for $p = 1$ and $\mathcal{A}_j^{hq}$ consisting of quadratic polynomials, the least-squares method yields approximate test functions $\tilde{\psi}_{-1,j}$ very similar to those obtained by α-quadratic upwinding. The Hemker test functions for $p = 1$ and constant coefficients a, b may be obtained with our least-squares method if one includes in the spaces $\mathcal{A}_j^{hq}$ exponentials which solve the homogeneous adjoint problem.

4.2.1 *Approximate test functions via local least-squares FEM.* To motivate the method, we define $\mathcal{A}_j := \left(H^2 \cap H_0^1\right)(I_j)$, $j = 1,..., N$. We define further $\varphi_j(x)$ to be the piecewise linear 'hat' function with

$$\varphi_j(x_{j-1}) = \varphi_j(x_{j+1}) = 0, \qquad \varphi_j(x_j) = 1, \qquad \varphi_j(x) = 0 \text{ on } \Omega \backslash \overline{I_{j-1} \cup I_j}.$$

Then $\psi_{-1,j} - \varphi_{j-1} \in \mathcal{A}_{j-1} \cup \mathcal{A}_j$ and we have the variational characterization

$$\left(\psi_{-1,j} - \varphi_{j-1}\right)\big|_{I_k} = \arg \min_{\psi \in \mathcal{A}_k} \left\| L_\varepsilon^* \left(\psi - \varphi_{j-1}\right) \right\|_{L^2(I_k)}^2, \qquad k = j-1, j, \tag{4.20}$$

for $j = 2,..., N$ and

$$\psi_{ij}\big|_{I_j} = \arg \min_{\psi \in \mathcal{A}_j} \left\| L_\varepsilon^* \psi - L_i \left(2 \frac{\cdot - m_j}{h_j}\right) \right\|_{L^2(I_j)}^2 \qquad i = 0,..., p_j - 2, \qquad j = 1,..., N. \tag{4.21}$$

The assumptions (1.3), (1.4) imply that the operator L_ε and also its adjoint L_ε^* are injective from $\mathcal{A}_j \to L^2(I_j)$. Hence the expression

$$\|\psi\|_{*,j} := \left\| L_\varepsilon^* \psi \right\|_{L^2(I_j)} \tag{4.22}$$

is a norm on $\mathcal{A}_j$ (homogeneity and triangle inequality being obvious) and the quadratic functionals in (4.20), (4.21) are strictly convex and lower semicontinuous. Therefore (4.20), (4.21) admit unique solutions $\psi_{-1,j}$, ψ_{ij} which coincide with those in (3.7), (3.8).

For a numerical approximation of the functions $\psi_{-1,j}$, ψ_{ij}, let $\mathcal{A}_j^{hq} \subset \mathcal{A}_j$ be a finite-dimensional subspace. We obtain external approximate test functions $\tilde{\psi}_{-1,j}$ by

$$\left(\tilde{\psi}_{-1,j} - \varphi_{j-1}\right)\Big|_{I_k} = \arg \min_{\psi \in \mathcal{A}_k^{hq}} \left\| L_\varepsilon^* \left(\psi - \varphi_{j-1}\right) \right\|_{L^2(I_k)}^2, \qquad k = j-1, j \tag{4.23}$$

for $j = 2,..., N$ and internal approximate test functions $\tilde{\psi}_{ij}$, $i = 0,..., p_j - 2$ by

$$\tilde{\psi}_{ij}\Big|_{I_j} = \arg \min_{\psi \in \mathcal{A}_j^{hq}} \left\| L_\varepsilon^* \psi - L_i \left(2 \frac{\cdot - m_j}{h_j}\right) \right\|_{L^2(I_j)}^2 \qquad j = 1,..., N. \tag{4.24}$$

These approximations are also uniquely defined. Moreover, they are optimal in the norm $\|\circ\|_{*,j}$, for we have

$$\left\| \psi_{-1,j} - \tilde{\psi}_{-1,j} \right\|_{*,k} \leqslant \left\| \psi_{-1,j} - \varphi_{j-1} - \psi \right\|_{*,k} \quad \forall \psi \in \mathcal{A}_k^{hq}, \tag{4.25}$$

$$k = j-1, j, \quad j = 2,..., N,$$

$$\left\| \psi_{ij} - \tilde{\psi}_{ij} \right\|_{*,j} \leqslant \left\| \psi_{ij} - \psi \right\|_{*,j}, \quad \forall \psi \in \mathcal{A}_j^{hq}, \quad j = 1,..., N. \tag{4.26}$$

Thus, the design of the approximation spaces $\mathcal{A}_j^{hq}$ proceeds in the usual fashion: based on the regularity of the exact test functions ψ_{ij}, we show that we can select the least-squares approximation spaces $\mathcal{A}_j^{hq}$ so that exponential convergence rates for the approximate test functions can be achieved.

REMARK 4.5 The calculation of the approximate test functions $\tilde{\psi}_{-1,j}, \tilde{\psi}_{ij}$ can be done efficiently if one observes that (4.23), (4.24) represent completely decoupled local problems on the elements I_j, which can be solved in parallel. Furthermore, on each element I_j the local least-squares problems (4.23), (4.24) can be solved efficiently because the equivalent matrix formulations lead to problems with the same stiffness matrix and merely different right-hand sides. Thus, once a convenient decomposition of the elemental least-squares matrix is found (e.g., its LU decomposition), the approximate test functions $\tilde{\psi}_{-1,j}|_{I_j}, \tilde{\psi}_{-1,j+1}|_{I_j}, \tilde{\psi}_{ij}, i = 0,..., p_j - 2$ can be obtained by $p_j + 1$ backsolves.

4.2.2 *Analysis of the least-squares approach.* Here, we focus on the choice of piecewise polynomials of degree q for the local spaces $\mathcal{A}_j^{hq}$. In view of the fact that (4.23), (4.24) are minimization problems whose solutions solve (local) adjoint problems with polynomial right-hand sides, one can expect that the mesh design principles that were presented in Section 3.3 can be employed. For the sake of brevity, we merely state here the main results—the proof is similar to that of Theorem 3.3 and the detailed arguments can be found in Melenk & Schwab (1997).

Let $\mathcal{T} = \{I_j : j = 1,..., N\}$ be the mesh for the trial space $S_0^{\vec{p}}(\mathcal{T})$. Let q be the polynomial degree to be used on the subgrid. For each $I_j = (x_{j-1}, x_j)$, $h_j = x_j - x_{j-1}$, and $\kappa > 0$ define the 'two-element' *subgrid* $\mathcal{T}_{j,\kappa}^*$ by

$$\mathcal{T}_{j,\kappa}^* := \begin{cases} \{(x_{j-1}, x_{j-1} + \kappa q \varepsilon), (x_{j-1} + \kappa q \varepsilon, x_j)\} & \text{if } \kappa q \varepsilon \leqslant h_j/2 \\ \{(x_{j-1}, x_j)\} & \text{else} \end{cases} \tag{4.27}$$

with $\vec{q} = (q, q)$ in the first case and $\vec{q} = q$ in the second case. Based on these 'two-element' subgrid meshes, we then set (cf. (3.1))

$$\mathcal{A}_j^{hq} := S_0^{\vec{q},2}(\mathcal{T}_{j,\kappa}^*). \tag{4.28}$$

It can be shown (cf. Melenk & Schwab (1997) for the details) that, if the polynomial degree q is sufficiently large, then the approximate test functions $\tilde{\psi}_{k,j}$ are sufficiently accurate and the assumptions of Theorem 4.1 are met. This is formulated in the following theorem.

THEOREM 4.6 Let $\mathcal{T} = \{I_j : j = 1,..., N\}$ be a mesh with N elements and let $S_0^{\vec{p}}(\mathcal{T})$ be the trial space. For $\kappa > 0$ and $q \in \mathbb{N}$ define spaces $\mathcal{A}_j^{hq}$ by (4.28).

Then there is $\kappa_0 > 0$ depending only on the constants in (1.3)–(1.6) such that for every $0 < \kappa < \kappa_0$ there is $c > 0$ such that for $q \geqslant c \max(p, |\ln \varepsilon| + \ln N)$ the FEM (4.4) corresponding to $\tilde{S}_L^{\vec{p}}$ (computed by (4.23), (4.24) based on the spaces $\mathcal{A}_j^{hq}$) is stable and hence quasi-optimal.

5. Implementational aspects and numerical example

5.1 *Implementational issues*

It is essential to have a quadrature scheme that can calculate accurately integrals of functions with boundary layer character for the evaluation of the load vector and for the evaluation of the contribution $\int_{I_j} uL_\varepsilon^* v\,\mathrm{d}x$ in the stiffness matrix as in general both the approximate test functions v and $L_\varepsilon^* v$ have boundary layer character. Integrals of functions of boundary layer type, however, can be evaluated numerically in a very efficient way using standard composite Gaussian quadrature formulae as is shown in the following lemma.

LEMMA 5.1 Let w, f be analytic on $\Omega = (-1, 1)$ and $\varepsilon \in (0, 1]$. Assume that there are C_f, K_f, C_w, K_w such that for all $n \in \mathbb{N}_0$, $x \in \Omega$ there holds

$$|f^{(n)}(x)| \leqslant C_f (K_f)^n n!, \quad |w^{(n)}(x)| \leqslant C_w (K_w)^n \mathrm{e}^{-(1-x)/\epsilon} \max(n, \varepsilon^{-1})^n.$$

For $q \in \mathbb{N}$ let $\mathcal{T}_{\kappa,\varepsilon}$ be the 'two-element' meshes introduced in (3.23) (with q taking the role of p) and denote by $G_q(\mathcal{T}_{\kappa,\varepsilon}, wf)$ the composite Gaussian quadrature rule with q points in each element applied to the function wf. Then there is $\kappa_0 > 0$ such that for $0 < \kappa < \kappa_0$ there are $C, \sigma > 0$ (independent of ε, q) such that

$$\left| \int_\Omega w(x) f(x)\,\mathrm{d}x - G_q(\mathcal{T}_{\kappa,\varepsilon}, wf) \right| \leqslant C\mathrm{e}^{-\sigma q}, \quad q = 1, 2, 3,... \tag{5.1}$$

If f is a polynomial of degree p with $\|f\|_{L^\infty(\Omega)} \leqslant 1$ then under the assumption $q \geqslant p+1$, estimate (5.1) holds with C, σ independent of ε, p, q.

Proof. Observe that for the composite Gaussian quadrature formula of order q the quadrature error may be estimated by twice the size of the integration domain times a L^∞ best approximation of the integrand:

$$\left| \int_\Omega w(x) f(x)\,\mathrm{d}x - G_q(\mathcal{T}_{\kappa,\varepsilon}, wf) \right| \leqslant 2\,|\Omega| \inf_{\pi_{2q-1}} \left\| wf - \pi_{2q-1} \right\|_{L^\infty(\Omega)}$$

where the infimum is taken over all piecewise polynomials π_{2q-1} of degree $2q-1$ on the mesh $\mathcal{T}_{\kappa,\varepsilon}$. Let $\pi_{q-1}(f)$, $\pi_q(w)$ be the piecewise Gauss–Lobatto interpolants of f, w of orders $q-1$, q, respectively. Upon setting $\pi_{2q-1} := \pi_{q-1}(f)\pi_q(w)$ and using the stability result of the Gauss–Lobatto interpolation operator (cf., e.g., equation (18) of Melenk (1997)), we can bound

$$\begin{aligned} \|wf - \pi_{2q-1}\|_{L^\infty(\Omega)} &\leqslant \|f - \pi_{q-1}(f)\|_{L^\infty(\Omega)} \|w\|_{L^\infty(\Omega)} \\ &\quad + C_{GL}(1 + \ln q) \|f\|_{L^\infty(\Omega)} \|w - \pi_q(w)\|_{L^\infty(\Omega)}. \end{aligned} \tag{5.2}$$

The proof of Theorem 16 of Melenk (1997) shows that for the function w, which is of boundary layer type,

$$\|w - \pi_q(w)\|_{L^\infty(\Omega)} \leqslant C\mathrm{e}^{-\sigma q}$$

with $C, \sigma > 0$ independent of q, ε provided that κ is sufficiently small. A similar estimate holds for $\|f - \pi_{q-1}(f)\|_{L^\infty(\Omega)}$ by Lemma 11 of Melenk (1997), and thus the right-hand side of (5.2) is exponentially small (in q). Finally, if f is a polynomial of degree p and $q \geqslant p+1$, then the term involving $f - \pi_{q-1}(f)$ vanishes in (5.2), and hence the claim of the lemma follows. □

REMARK 5.2 Note that Lemma 5.1 shows that accurate numerical integration of boundary layer functions is possible without constructing special, 'exponentially' weighted quadrature rules. The present approach works even without explicit knowledge of the boundary layer function. As we pointed out in Remark 3.4, the value of κ_0 is in principle available from the proof. For the special weight function $w = \mathrm{e}^{-(1-x)/\varepsilon}$, the analysis of Schwab & Suri (1996) shows that $\kappa_0 = 4/e$. Furthermore, note that the use of geometrically refined meshes outlined in Remark 4.4 eliminates the need for bounds on κ_0. We chose π_{2q-1} as the product of (piecewise) polyomials of degree $q-1$ and q. However, other 'splittings' are possible and thus the condition $q \geqslant p+1$ for polynomial right-hand sides f may be relaxed to a condition of the form $q \geqslant \tau p$ with $\tau > 1/2$.

It is interesting to note that complete knowledge of the test functions is not required to formulate the method. The stiffness matrix corresponding to $B_{\mathcal{T}}(u, v)$ consists of two parts: a mass-matrix-like term stemming from the domain integrals and a finite-volume-like term stemming from the flux-jumps at interelement boundaries. For the mass matrix part, the test functions $\psi_{-1,j}$, $\psi_{i,j}$ need not be known completely. Rather, only $L_\varepsilon^*\psi_{-1,j}$, $L_\varepsilon^*\psi_{i,j}$ (which are chosen to be Legendre polynomials and hence known) are required. For the flux jumps, the only information employed from the Dirichlet problems (3.7), (3.8) are the normal derivatives in the endpoints, i.e., a Dirichlet-to-Neumann map is needed. In Section 4, we proposed a least-squares method to approximate the test functions and in our numerical example ahead we used exact solution formulae for the problem with frozen coefficients, but for the generation of the stiffness matrix, any sufficiently accurate Dirichlet-to-Neumann map may be taken.

Concerning the potential for parallelism and the band width of the resulting global system, the PG method presented in this paper is very similar to the usual *hp* FEM for standard second-order problems. The local adjoint problems determining the (approximate) test functions may be solved completely independently and in parallel and the local stiffness matrices can then be computed in parallel as well. The resulting global system is banded with band width p just as in the usual *hp* FEM. Moreover, the PG FEM allows readily for robust *a posteriori* error estimation, as has been shown by Szymczak & Babuška (1984).

5.2 *Numerical example*

The aim of the present section is to corroborate with a numerical example our results of Theorem 4.1 and Theorem 4.3, which state that fairly crude approximations of the test functions $\psi_{i,j}$ suffice to retain stability and hence quasi-optimality of our *hp* PG method.

We consider the model problem

$$-\varepsilon u_\varepsilon'' + a(x)u_\varepsilon' = f(x) \qquad \text{on } \Omega = (-1, 1), \qquad u_\varepsilon(\pm 1) = 0, \tag{5.3}$$
$$a(x) = 2 - x, \qquad f(x) = 2 - x.$$

The exact solution has a boundary layer at the outflow boundary and is given by

$$u_\varepsilon = x + \alpha + \beta \operatorname{erfc}\left(\frac{2-x}{\sqrt{2\varepsilon}}\right), \qquad \operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-t^2}\, \mathrm{d}t,$$

$$\alpha = \frac{\operatorname{erfc}(1/\sqrt{2\varepsilon}) + \operatorname{erfc}(3/\sqrt{2\varepsilon})}{\operatorname{erfc}(1/\sqrt{2\varepsilon}) - \operatorname{erfc}(3/\sqrt{2\varepsilon})}, \qquad \beta = -\frac{2}{\operatorname{erfc}(1/\sqrt{2\varepsilon}) - \operatorname{erfc}(3/\sqrt{2\varepsilon})}.$$

Guided by the approximation result Theorem 3.3, we choose for the trial space the spaces $S_0^{\vec{p},1}(\mathcal{T}_{\kappa,\varepsilon})$ with $\kappa = 1$ (cf. (3.1)) where the meshes $\mathcal{T}_{\kappa,\varepsilon}$ are given by (3.23). A specific basis of $S_0^{\vec{p},1}$ is given by the usual piecewise linear 'nodal' shape functions and the integrated Legendre polynomials (the 'internal' shape functions).

The test functions are taken as those that arise as the 'exact' test functions for the operator L_ε^* with *frozen* coefficients. More specifically, with $m_j = (x_{j-1} + x_j)/2$, we define the operators

$$\widetilde{L^*}_j u := -\varepsilon u'' - a(m_j)u' - a'(m_j)u$$

and take the test functions $\tilde{\psi}_{i,j}$ as the solutions of the following problems:

$$\begin{aligned}
\widetilde{L^*}_{j-1}\tilde{\psi}_{-1,j} &= 0 \quad \text{on } I_{j-1},\\
\widetilde{L^*}_j\tilde{\psi}_{-1,j} &= 0 \quad \text{on } I_j,\\
\tilde{\psi}_{-1,j} &= 0 \quad \text{on } I_k, \quad k \neq j, j-1,\\
\tilde{\psi}_{-1,j}(x_k) &= \delta_{j,k+1}, \quad k = 1,\ldots, N-1,\\
\widetilde{L^*}_j\tilde{\psi}_{i,j} &= L_i\left(2\frac{x-m_j}{h_j}\right) \quad \text{on } I_j, \quad i = 0, 1,\ldots,\\
\tilde{\psi}_{i,j} &= 0 \quad \text{on } I_k, \quad k \neq j, \quad i - 0, 1,\ldots.
\end{aligned}$$

Note that our particular choice reduces to the classical Hemker test functions for $p = 1$; for $p > 1$ therefore, our scheme could be viewed as an hp version of the 'Hemker test function' method. The approximate test functions $\tilde{\psi}_{i,j}$ are available in closed form since the Green's function for the constant coefficient operators $\widetilde{L^*}_j$ are known. The evaluation of the test functions $\tilde{\psi}_{i,j}$ (and their derivatives) at a point x requires the integration of Green's function, which has boundary layer character, against polynomials of degree p. We note that the 'two-element' composite Gaussian quadrature scheme of Lemma 5.1 allows us to evaluate the test functions very efficiently.

Our numerical experiments were performed using MATLAB, i.e., with double precision (16 decimal) accuracy. For the evaluation of the test functions $\tilde{\psi}_{i,j}$, the load vector, and the stiffness matrix, the 'two-element' composite Gaussian quadrature rule with $p+q$ points in each of the two subelements is used as suggested by Lemma 5.1. Figures 1 and 2 show the results when overintegration with $q = 30$ is used, i.e., all integrals can be assumed to be evaluated exactly (within machine precision). Figures 3 and 4 are obtained using the quadrature rule with $q = 0$.

In Fig. 1, the relative error in L^2 versus the polynomial degree is plotted. It illustrates that robust exponential convergence can be achieved in the L^2 norm as predicted by Theorem 4.3; the error curves are fairly straight lines in the semilogarithmic plot and they are practically on top of each other for $\varepsilon = 10^{-2}$ down to $\varepsilon = 10^{-14}$. Figure 2 shows the performance of the method when the relative error is measured in the H^1 semi-norm; again, robust exponential convergence is observed. Figures 3 and 4 show the effect of quadrature, i.e., when 'two-element' quadrature rules with p points in each subelement are used. We note that this implies in particular that the test functions are evaluated less

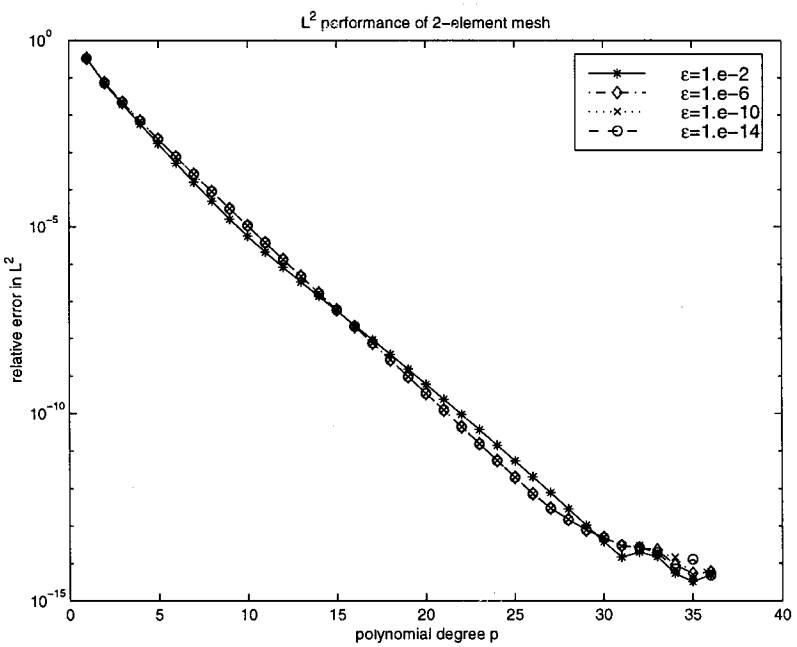

FIG. 1. L^2 performance of 'two-element mesh'; integration exact.

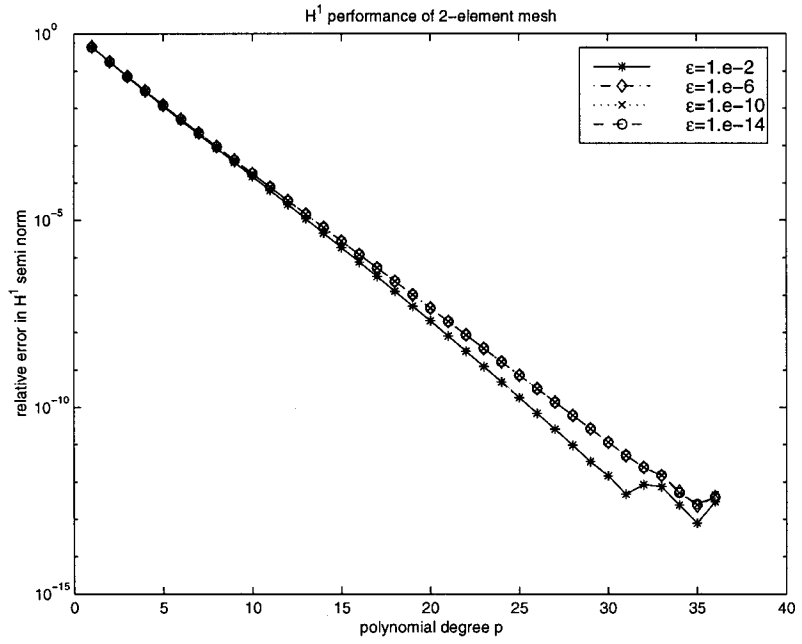

FIG. 2. H^1 semi-norm performance of 'two-element mesh'; integration exact.

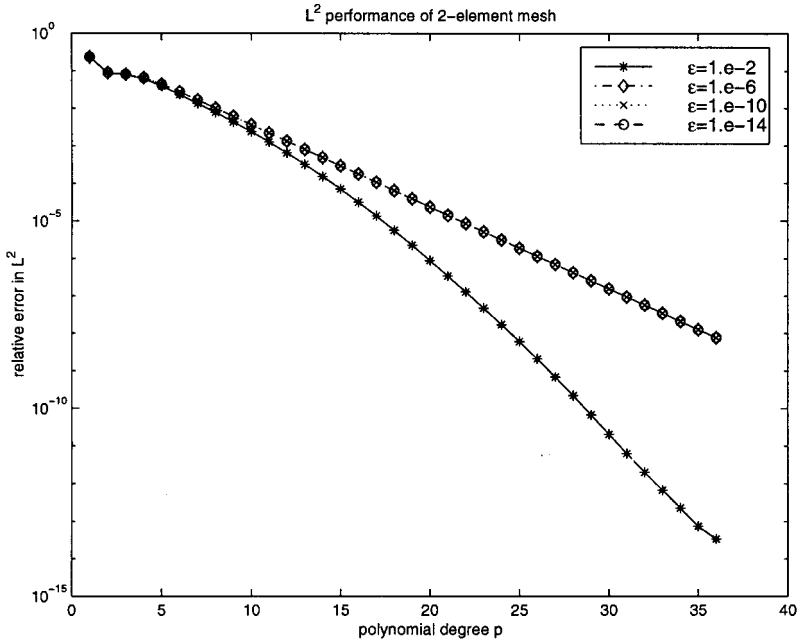

FIG. 3. L^2 performance of 'two-element mesh'; effect of quadrature.

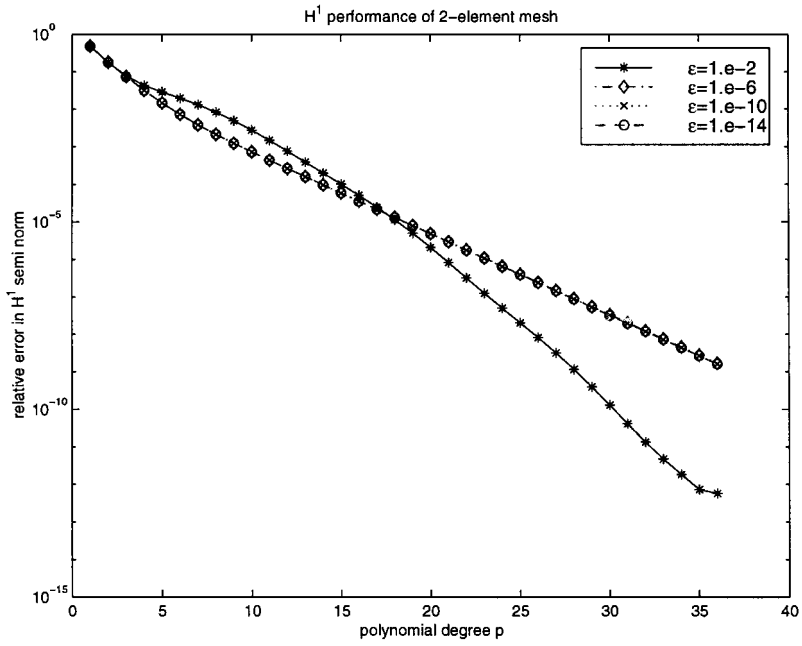

FIG. 4. H^1 semi-norm performance of 'two-element mesh'; effect of quadrature.

accurately. Nevertheless, even with quadrature, we observe robust exponential convergence in both the L^2 and the H^1 norm and the method remains stable.

Theorem 4.1 stipulates that the test functions have to be computed more accurately as the spectral order p is increased as $\gamma_M^{-2} = O(p^{-1})$. This was not observed in the present numerical example. The residuals η_{ij} as defined in (4.2), (4.3) (and hence also the quantities Λ_1, Λ_2 of (4.6), (4.7)) do *not* tend to zero as $p \to \infty$. Nevertheless, robust exponential convergence is visible in Figs 1–4.

In summary, our numerical experiments show that our error estimates are sharp and that they describe accurately the performance of the Petrov–Galerkin *hp* FEM: the impact of the quadrature order on the stability and consistency follows closely the predictions made in Theorem 4.3 and the method performs uniformly well for the viscosity parameter ε ranging from $\varepsilon = 10^{-2}$ to the order of machine precision, $\varepsilon = 10^{-14}$.

REFERENCES

CHRISTIE, J., GRIFFITHS, D. F., MITCHELL, A. R., & ZIENKIEWICZ, O. C. 1978 Finite element methods for second order differential equations with significant first derivatives. *Int. J. Numer. Methods Eng.* **10**, 1764–1771.

DE GROEN, P. P. N. 1978 A finite element method with large mesh width for stiff two-point boundary value problems. *Preprint*, Department of Mathematics, Eindhoven University, Eindhoven, The Netherlands.

DE GROEN, P. P. N. & HEMKER, P. W. 1979 Error bounds for exponentially fitted Galerkin methods applied to stiff two-point boundary value problems. *Numerical Analysis of Singular Perturbation Problems* (P. W. Hemker and J. J. H. Miller, eds). New York: Academic.

GARTLAND, E. C. 1987 Uniform high-order difference schemes for a singularly perturbed two-point boundary value problem. *Math. Comput.* **48**, 551–564.

HEMKER, P. W. 1977 A numerical study of stiff two-point boundary problems. *PhD thesis*, Mathematisch Centrum, Amsterdam.

HUGHES, T. J. & BROOKS, A. 1979 A multidimensional upwind scheme with no crosswind diffusion. *Finite Element Methods for Convection Dominated Flows (AMD 34)* (T. J. Hughes, ed). New York: American Society of Mechanical Engineers.

HUGHES, T. J. & BROOKS, A. 1982 A theoretical framework for Petrov–Galerkin methods with discontinuous weighting functions. Applications to the streamline diffusion procedure. *Finite Elements in Fluids*, vol. 4 (Gallagher, ed). New York: Wiley.

JOHNSON, C. & NÄVERT, U. 1981 An analysis of some finite element methods for advection-diffusion problems *Analytical and Numerical Approaches to Asymptotic Problems in Analysis* (O. Axelsson, L. S. Frank and A. van der Sluis, eds). Amsterdam: North-Holland, pp 99–116

KELLOGG, R. B. & TSAN, A. 1978 Analysis of some difference approximations for a singularly perturbed problem without turning points. *Math. Comput.* **32**, 1025–1039.

MELENK, J. M. 1997 On the robust exponential convergence of finite element methods for problems with boundary layers. *IMA J. Numer. Anal.* **17**, 577–601.

MELENK, J. M. & SCHWAB, C. 1997 An *hp* finite element method for convection–diffusion problems. *Research Report 97-05*, Seminar für Angewandte Mathematik, ETH Zürich. http://www.sam.math.ethz.ch/Reports/reports.html.

MELENK, J. M. & SCHWAB, C. 1998 The *hp* streamline–diffusion method for convection dominated problems in one space dimension. *Research Report 98-10*, Seminar für Angewandte Mathematik, ETH Zürich. http://www.sam.math.ethz.ch/Reports/reports.html.

MILLER, J. J. H, O'RIORDAN, E., & SHISHKIN, G. I. 1996 *Fitted Numerical Methods for Singular Perturbation Problems*. Singapore: World Scientific.

MORTON, K. W. 1995 *Numerical Solution of Convection–Diffusion Problems*. Oxford: Oxford University Press.

ROOS, H.-G., STYNES, M., & TOBISKA, L. 1996 *Numerical Solution of Singularly Perturbed Boundary Value Problems*. Heidelberg: Springer.

SCHWAB, C. & SURI, M. 1996 The p and hp versions of the finite element method for problems with boundary layers. *Math. Comput.* **65**, 1403–1429.

SCHWAB, C., SURI, M., & XENOPHONTOS, C. A. 1996 Boundary layer approximation by spectral/hp methods. *Houston J. Math. Spec. Issue of ICOSAHOM '95 Conf.* (A. V. Illin and L. R. Scott, eds), pp 501–508. (ISSN: 0362-1588).

SZYMCZAK, G. W. 1982 An adaptive finite element method for convection–diffusion problems. *PhD Dissertation*, University of Maryland, College Park.

SZYMCZAK, G. W. & BABUŠKA, I. 1984 Adaptivity and error estimation for the finite element method applied to convection diffusion problems. *SIAM J. Numer. Anal.* **21**, 910–954.

Appendix: Regularity

The proof of Theorem 1.1 is very similar to those of Theorems 3, 5, 6, and Corollary 4 of Melenk (1997) and rests on the following four lemmata.

The ensuing Lemmata A1, A2, A3 follow in the case $\underline{b} \geqslant 0$ by arguments similar to those of Kellogg & Tsan (1978); when $\underline{b} < 0$, this can be reduced to the case $\underline{b} \geqslant 0$ by a transformation like that on p 70 of Miller *et al* (1996).

LEMMA A.1 There is $C > 0$ depending only on the constants appearing in (1.3), (1.5), (1.6), and C_f such that the solution u_ε of (1.1) satisfies

$$\|u_\varepsilon\|_{L^\infty} \leqslant C, \qquad \|u_\varepsilon'\|_{L^\infty} \leqslant C\varepsilon^{-1}.$$

LEMMA A.2 Let u_ε^+ be the outflow boundary layer defined in (1.10). Then there is $C > 0$ depending only on the constants of (1.3), (1.5), (1.6) such that

$$|u_\varepsilon^+(x)| \leqslant \mathrm{e}^{-\underline{a}(1-x)/(2\varepsilon)}, \qquad |u_\varepsilon^{+\prime}(x)| \leqslant C\varepsilon^{-1}\mathrm{e}^{-\underline{a}(1-x)/(2\varepsilon)}.$$

LEMMA A.3 Let u_ε^+ be the outflow boundary layer defined in (1.10). Then there are constants C_1, $C_2 > 0$ depending only on the constants of (1.3), (1.5), (1.6) such that

$$C_1\varepsilon^{-1} \leqslant u_\varepsilon^{+\prime}(1) \leqslant C_2\varepsilon^{-1}.$$

LEMMA A.4 Let G be an open, complex neighbourhood of $I = [-1, 1]$. Assume that the functions $\Lambda, a, u_0 : G \to \mathbb{C}$ are holomorphic and bounded on G. Assume additionally that $|a| \geqslant \underline{a} > 0$ on G. Then there are constants $C, K_1, K_2 > 0$ depending only on $\underline{a}$, $\|a'\|_{L^\infty(G)}$, $\|\Lambda\|_{L^\infty(G)}$, $\|\Lambda'\|_{L^\infty(G)}$, and G such that the functions u_j defined recursively as in (1.9) satisfy

$$\|u_j^{(n)}\|_{L^\infty(I)} \leqslant C K_1^j K_2^n j! n! \|u_0\|_{L^\infty(G)} \qquad \forall j, n \in \mathbb{N}_0.$$

Proof. We will prove a stronger statement. Without loss of generality we may assume that G is star shaped with respect to $z = -1$. For $\delta > 0$ (sufficiently small) denote $G_\delta := \{z \in G \mid \mathrm{dist}(z, \partial G) > \delta\}$. Then we claim that there are $C, K > 0$ such that

$$\|u_j\|_{L^\infty(G_\delta)} \leqslant C K^j \delta^{-j} j! \|u_0\|_{L^\infty(G)} \qquad \forall j \in \mathbb{N}_0.$$

The proof of the lemma follows from this estimate by Cauchy's integral theorem for derivatives.

It remains therefore to establish the claim. One proceeds by induction on j. It is true for $j = 0$ and for any $C \geqslant 1$. We write

$$\begin{aligned} u_{j+1}(z) &= \mathrm{e}^{-\Lambda(z)} \int_{-1}^{z} \mathrm{e}^{\Lambda(t)} \frac{1}{a(t)} u_j''(t)\,\mathrm{d}t \\ &= \mathrm{e}^{-\Lambda(z)} \left[\mathrm{e}^{\Lambda(t)} \frac{1}{a(t)} u_j'(t) \right]_{-1}^{z} - \mathrm{e}^{-\Lambda(z)} \int_{-1}^{z} \mathrm{e}^{\Lambda(t)} \frac{\Lambda'(t)a(t) + a'(t)}{a(t)^2} u_j'(t)\,\mathrm{d}t. \end{aligned}$$

Hence there is $C_1 > 0$ such that

$$\|u_{j+1}\|_{L^\infty(G_\delta)} \leqslant C_1 \|u_j'\|_{L^\infty(G_\delta)}.$$

The remainder of the proof resembles that of Lemma 2 of Melenk (1997). □