2017

A Posteriori Estimates Using Auxiliary Subspace Techniques

Jeffrey S. Ovall, *Portland State University*

A Posteriori Estimates Using Auxiliary Subspace Techniques

Harri Hakula, Michael Neilan & Jeffrey S. Ovall

Journal of Scientific Computing

ISSN 0885-7474
Volume 72
Number 1

J Sci Comput (2017) 72:97-127
DOI 10.1007/s10915-016-0352-0

Springer

J Sci Comput (2017) 72:97–127
DOI 10.1007/s10915-016-0352-0

A Posteriori Estimates Using Auxiliary Subspace Techniques

Harri Hakula[1] · **Michael Neilan**[2] · **Jeffrey S. Ovall**[3]

Received: 12 November 2015 / Revised: 20 November 2016 / Accepted: 27 December 2016 /
Published online: 4 January 2017

Abstract A posteriori error estimators based on auxiliary subspace techniques for second order elliptic problems in $\mathbb{R}^d$ ($d \geq 2$) are considered. In this approach, the solution of a global problem is utilized as the error estimator. As the continuity and coercivity of the problem trivially leads to an efficiency bound, the main focus of this paper is to derive an analogous effectivity bound and to determine the computational complexity of the auxiliary approximation problem. With a carefully chosen auxiliary subspace, we prove that the error is bounded above by the error estimate up to oscillation terms. In addition, we show that the stiffness matrix of the auxiliary problem is spectrally equivalent to its diagonal. Several numerical experiments are presented verifying the theoretical results.

Keywords Finite element methods · A posteriori error estimation · High-order methods

1 Introduction

A posteriori error estimation and adaptive mesh refinement have become essential components of high-performace scientific computing using finite elements. One class of methods for estimating error that has proven to be very robust in practice is the so-called hierarchical basis approach, whose origins can be traced back to [16,17,39]. In this approach, given a finite

✉ Jeffrey S. Ovall
jovall@pdx.edu

Harri Hakula
harri.hakula@aalto.fi

Michael Neilan
neilan@pitt.edu

1 Department of Mathematics and Systems Analysis, Aalto University, Aalto, Finland

2 Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, USA

3 Fariborz Maseeh Department of Mathematics and Statistics, Portland State University, Portland, OR, USA

element approximation $\hat{u} \in V$ of the solution u, an approximate error function $\varepsilon \approx u - \hat{u}$ is computed in a finite dimensional auxiliary space W satisfying $V \cap W = \{0\}$. A global error estimate is obtained by measuring ε in an appropriate norm, and local norms of ε are used as local error indicators to drive an adaptive algorithm. Hierarchical basis methods belong to the broader class of implicit methods, which require the solution of additional, simpler (local or global) systems to obtain an error estimate. We note that "hierarchical basis error estimation" has occassionally been used to describe a class of explicit methods that bear some superficial similarities with what is considered here (cf. [25,35]). Such methods can be shown to be equivalent to the standard residual method, whereas hierarchical methods as discussed here cannot, in general.

Traditionally (cf. [2,5,8,38]) hierarchical basis estimators have been analyzed for self-adjoint, coercive problems, with error measured in the induced "energy" norm $|||\cdot|||$. In such cases, ε is the orthogonal projection of the error $u - \hat{u}$ onto the space W, so one clearly has the efficiency bound $|||\varepsilon||| \leq |||u - \hat{u}|||$. A complementary reliability bound, $|||u - \hat{u}||| \leq C|||\varepsilon|||$, is then obtained by exploiting a Strong Cauchy Inequality (cf. [19]) between the spaces V and W, and making a saturation assumption—i.e. that the best approximation of u in $V \oplus W$ is strictly better than its best approximation in V. Although the saturation assumption is generally expected to hold on sufficiently fine meshes for problems likely to be encountered in practice, it is not difficult to construct counter-examples (cf. [12,18]) for particular problems on particular meshes. In [12,18], notions of *data oscillation* are used to replace the saturation assumption with a quantity that is directly measurable, even it if is not generally measured in practice. A related notion of oscillation appears explicitly in the reliability bound of the present work.

The contributions [3,12,22,28] provide a representative sample of more recent work in the area, considering such issues as non-self-adjoint problems and asymptotic exactness of error estimates, primarily for lower-order elements. Here, we focus on higher-order elements on meshes in $\mathbb{R}^d$, and our choice of auxiliary space W becomes more obviously different from what would be considered a standard choice when $d \geq 3$. To clarify, W is typically chosen so that $V \oplus W$ is a natural approximation space—for example, if V is the Lagrange space having local polynomial degree p, $V \oplus W$ might be the space having local polynomial degree $> p$ (a p-hierarchy), or the space having local polynomial degree p on a refinement of the mesh used for V (an h-hierarchy). Our choice of W differs, and is motivated by the fact that finite element errors may be decomposed in terms of residuals on mesh cells and their $(d-1)$-dimensional interfaces, so W should be designed such that it can adequately "capture" these components of the error.

1.1 Problem Statement and Background

Let $\Omega \subset \mathbb{R}^d$ $(d \geq 2)$ be a bounded polytope, having boundary $\partial\Omega = \Gamma_N \cup \Gamma_D$, a disjoint union with Γ_D closed in the relative topology on $\partial\Omega$. We define the space

$$H^1_{0,D}(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D \text{ in the sense of trace}\},$$

and adopt the following notation for norms and semi-norms on Hilbert spaces $H^k(\omega)$ $(k \geq 0)$ for $\omega \subset \Omega$,

$$\|v\|^2_{k,\omega} = \sum_{|\alpha| \leq k} \|D^\alpha v\|^2_{L^2(\omega)}, \quad |v|^2_{k,\omega} = \sum_{|\alpha| = k} \|D^\alpha v\|^2_{L^2(\omega)}.$$

When $\omega = \Omega$, we omit it from the subscript. We also employ these Sobolev spaces and norms on subsets of Ω having lower dimension.

We consider variational problems of the form

$$\text{Find } u \in H^1_{0,D}(\Omega) \ \ni: \ \underbrace{\int_\Omega A\nabla u \cdot \nabla v + (\boldsymbol{b} \cdot \nabla u + cu)v}_{B(u,v)} = \underbrace{\int_\Omega fv + \int_{\Gamma_N} gv}_{F(v)} \ \forall v \in H^1_{0,D}(\Omega), \tag{1.1}$$

where the data $A : \Omega \to \mathbb{R}^{d\times d}$, $\boldsymbol{b} : \Omega \to \mathbb{R}^d$, $c, f : \Omega \to \mathbb{R}$ and $g : \Gamma_N \to \mathbb{R}$ are piecewise smooth with respect to some polyhedral partition of Ω. The matrix A is symmetric and uniformly positive definite a.e. in Ω, $A(x)z \cdot z \geq \alpha|z|^2$ for all $z \in \mathbb{R}^d$ and a.e. $x \in \Omega$. We further assume conditions on the coefficients so that B is continuous and coercive,

$$|B(v, w)| \leq \mathfrak{C}\|v\|_1\|w\|_1, \ B(v, v) \geq \mathfrak{c}\|v\|_1^2 \quad \text{for all } v, w \in H^1_{0,D}(\Omega). \tag{1.2}$$

Under these assumptions, the problem (1.1) is well-posed. We refer to $\mathfrak{C}$ and $\mathfrak{c}$, respectively, as the continuity constant and the coercivity constant.

Remark 1.1 The coercivity assumption is convenient for our analysis, but it is not essential for our key efficiency and reliability result, Theorem 1.4. That theorem holds, as stated, if we instead have standard inf-sup conditions, (cf. [20, Theorem 2.6]) on the space $H^1_{0,D}(\Omega)$ and the discrete spaces V and W described below, with common inf-sup constant $\mathfrak{c}$. However, coercivity is used in an essential way in the proof of our spectral equivalence result, Theorem 1.5.

Given a family $\{\mathcal{T}\}$ of conforming, shape-regular simplicial partitions of Ω, we define the standard piecewise polynomial finite element spaces on $\mathcal{T}$:

$$V = V_p = \{v \in H^1_{0,D}(\Omega) : \ v|_T \in \mathcal{P}_p(T) \text{ for each } T \in \mathcal{T}\}, \tag{1.3}$$

where $\mathcal{P}_p(T)$ is the space of polynomials of total degree $\leq p$ on T. More generally $\mathcal{P}_p(S)$ is taken to be the space of polynomials of total degree $\leq p$ having domain $S \subset \mathbb{R}^j$ for some $0 \leq j \leq d$. Given a finite dimensional space $W \subset H^1_{0,D}(\Omega)$ such that $V \cap W = \{0\}$, we consider the approximation problem

$$\text{Find } \hat{u} \in V \ \ni: \ B(\hat{u}, v) = F(v) \ \ (= B(u, v)) \ \forall v \in V, \tag{1.4}$$

and the error problem

$$\text{Find } \varepsilon \in W \ \ni: \ B(\varepsilon, v) = F(v) - B(\hat{u}, v) \ \big(= B(u - \hat{u}, v)\big) \ \forall v \in W. \tag{1.5}$$

Since V and W inherit the continuity and coercivity of B, these problems are also well-posed. In the present work, W is also a piecewise polynomial space with respect to $\mathcal{T}$, but we postpone its definition to later sections. Throughout, we implicitly assume that $\mathcal{T}$ is subordinate to the polyhedral partition of Ω on which the data is piecewise smooth—i.e., the data is smooth on each simplex $T \in \mathcal{T}$.

1.2 Main Results

In order to describe our basic approach to constructing and analyzing ε, we introduce some basic notation. Let $\mathcal{F}$ denote the set of $(d-1)$-dimensional subsimplices, the "faces" of $\mathcal{T}$, and further decompose this as $\mathcal{F} = \mathcal{F}_I \cup \mathcal{F}_D \cup \mathcal{F}_N$, where $\mathcal{F}_I$ comprises those faces in the interior of Ω, and $\mathcal{F}_D$ and $\mathcal{F}_N$ comprise those faces in Γ_D and Γ_N, respectively. The starting point of our analysis is the error identity in Proposition 1.2, which follows directly from (1.4)

and (1.5) and elementwise integration-by-parts as used in the derivation of residual methods. More specifically, for $\phi \in H^1_{0,D}(\Omega)$, $B(u-\hat{u},\phi) = F(\phi) - B(\hat{u},\phi)$, so

$$\begin{aligned} B(u-\hat{u},\phi) &= \sum_{T\in\mathcal{T}} \int_T (f - \boldsymbol{b}\cdot\nabla\hat{u} + c\hat{u})\phi - A\nabla\hat{u}\cdot\nabla\phi + \sum_{F\in\mathcal{F}_N}\int_F g\phi \\ &= \sum_{T\in\mathcal{T}} \left(\int_T (f - (-\nabla\cdot A\nabla\hat{u} + \boldsymbol{b}\cdot\nabla\hat{u} + c\hat{u}))\phi - \int_{\partial T} A\nabla\hat{u}\cdot\boldsymbol{n}\phi \right) \\ &\quad + \sum_{F\in\mathcal{F}_N}\int_F g\phi. \end{aligned}$$

Distributing the integrals on ∂T to the faces $F \in \mathcal{F}_I \cup \mathcal{F}_N$, we obtain

Proposition 1.2 *For any $v \in H^1_{0,D}(\Omega)$, $w \in W$ and $\hat{v} \in V$, it holds that*

$$B(u-\hat{u},v) = B(\varepsilon,w) + \mathcal{R}(v-\hat{v}-w),$$

where

$$\mathcal{R}(\phi) = F(\phi) - B(\hat{u},\phi) = \sum_{T\in\mathcal{T}}\int_T R_T\phi + \sum_{F\in\mathcal{F}_I\cup\mathcal{F}_N}\int_F r_F\phi,$$

and

$$\begin{aligned} R_T &= f - (-\nabla\cdot A\nabla\hat{u} + \boldsymbol{b}\cdot\nabla\hat{u} + c\hat{u})|_T, \\ r_F &= \begin{cases} g - A\nabla\hat{u}\cdot\boldsymbol{n}, & F\in\mathcal{F}_N \\ (-A\nabla\hat{u}\cdot\boldsymbol{n}_T)|_T - (A\nabla\hat{u}\cdot\boldsymbol{n}_{T'})|_{T'}, & F\in\mathcal{F}_I \end{cases}. \end{aligned}$$

Here, T and T' are the simplices sharing the face F, and $\boldsymbol{n}_T$ and $\boldsymbol{n}_{T'}$ are their outward unit normals; for $F \in \mathcal{F}_N$, the outward normal to $\boldsymbol{n}$ for $\partial\Omega$ is used.

Remark 1.3 The identity $B(u-\hat{u},v) = \mathcal{R}(v-\hat{v})$ for $v \in H^1_{0,D}(\Omega)$ and $\hat{v} \in V$ is the starting point for residual error estimates, which are obtained by choosing $\hat{v}$ to be a suitable interpolant of v, and deriving corresponding bounds on the weak residual, $|\mathcal{R}(v-\hat{v})| \le C\eta\|v\|_1$. Here η is comprised of appropriate weights, involving the local mesh size, on the volumetric and face residuals, $\|R_T\|_{0,T}$ and $\|r_F\|_{0,F}$. We note that reliability bounds for residual estimators of this sort are very naturally obtained, and it is efficiency bounds, involving oscillation terms, which require more ingenuity to establish. This is the opposite of the situation for the error estimators discussed here.

With an appropriate choice of error space $W = W_{p+d}$, described in detail later, we obtain our key error theorem, the upper bound of which is proved in Sect. 2.3—the lower bound is a trivial consequence of the coercivity and continuity conditions (1.2).

Theorem 1.4 *There is a constant C depending only on the dimension d, polynomial degree p, continuity and coercivity constants $\mathfrak{C}$ and $\mathfrak{c}$, and the shape-regularity of $\mathcal{T}$ such that*

$$\frac{\mathfrak{c}}{\mathfrak{C}}\|\varepsilon\|_1 \le \|u-\hat{u}\|_1 \le C\left(\|\varepsilon\|_1 + \operatorname{osc}(R,r,\mathcal{T})\right),$$

where the residual oscillation is defined by

$$\operatorname{osc}(R,r,\mathcal{T})^2 = \sum_{T\in\mathcal{T}} h_T^2 \inf_{\kappa\in\mathcal{P}_{p-1}(T)} \|R_T - \kappa\|^2_{0,T} + \sum_{F\in\mathcal{F}_I\cup\mathcal{F}_N} h_F \inf_{\kappa\in\mathcal{P}_{p-1}(F)} \|r_F - \kappa\|^2_{0,F}.$$

Here and following, h_T is the diameter of T and h_F is the diameter of F.

The space W_{p+d} will be spanned by appropriate "face bubble functions" supported in the two (or one) simplices sharing a face, and "interior bubble functions" supported in a simplex. In [6,22] it was shown that interior bubbles are not needed for lowest order elements when $d = 2, 3$. A very different sort of analysis was used in [26,28] for lowest order elements and $d = 2$ to show that the (H^1 or energy) error estimates based on ε are often asymptotically identical to the actual error.

As stated, the computation of ε requires the formation and solution of a global system, so one might naturally be concerned that the approach is too expensive for practical consideration. However, even those implicit methods that are based on the solution local (elementwise or patchwise) problems require the computation of local stiffness matrices. In Sect. 3 we argue that the size and sparsity structure of the system for computing ε is comparable to that of setting up all of the local systems for other implicit methods. So when comparing the cost of this and other implicit methods, the real issue is whether or not it is more expensive to solve a single global system or a collection of local systems. Our key result in this regard is

Theorem 1.5 *The global stiffness matrix for W_{p+d} is spectrally-equivalent to its diagonal.*

Although this result implies that we could get away with solving a diagonal system, and that the modified $\tilde{\varepsilon}$ would still provide two-sided bounds as in Theorem 1.4, with suitably adjusted constants, we instead advocate approximately solving the full system using a few steps of a Krylov solver (CG, BiCG-Stab, GMRES) either with no preconditioning (e.g. when $d = 2$) or a simple Jacobi or Gauss-Seidel preconditioner.

We offer a few more remarks concerning the solution of local or global problems in the construction of error estimates. An approximate error function $\varepsilon \approx u - \hat{u}$ is very naturally obtained through the solution of the global problem (1.5), and can be used for driving anisotropic h-refinement or r-refinement (mesh smoothing) in practice. We note, however, that our analysis implicitly assumes isotropic meshes, though some of our experiments use elements with high aspect ratios. Although local problems might also be used do drive anisotropic refinement, empirical evidence [23,24] suggests that the solution of global problems are better suited for this purpose. A point in which some approaches based on local problems currently have a theoretical advantage over the approach presented here is that they are provably robust with respect to polynomial degree [13,21], a property which is known not to hold for standard residual-based error estimates [27]. The efficiency (lower) bound in Theorem 1.4 is clearly independent of the polynomial degree p, but the proof presented here for the reliability bound suggests that the constant C could depend on p, which is not ideal. Extensive numerical experiments, as reported in Sect. 4, provide empirical evidence that our estimate is robust with respect to p, and we hope to prove this in future work.

Before outlining the rest of the paper, we finally provide some motivation for the development in Sect. 2 by considering the residual oscillation term. We define the local residual oscillation for each $T \in \mathcal{T}$ by

$$\begin{aligned}\operatorname{osc}(R, r, T)^2 = h_T^2 \inf_{\kappa \in \mathcal{P}_{p-1}(T)} \|R_T - \kappa\|_{0,T}^2 + \frac{1}{2} \sum_{F \in \mathcal{F}_{I,T}} h_F \inf_{\kappa \in \mathcal{P}_{p-1}(F)} \|r_F - \kappa\|_{0,F}^2 \\ + \sum_{F \in \mathcal{F}_{N,T}} h_F \inf_{\kappa \in \mathcal{P}_{p-1}(F)} \|r_F - \kappa\|_{0,F}^2,\end{aligned}$$

where $\mathcal{F}_{I,T}$ are the faces of T in $\mathcal{F}_I$ and $\mathcal{F}_{N,T}$ are the faces of T in $\mathcal{F}_N$. By definition,

$$\operatorname{osc}(R, r, \mathcal{T})^2 = \sum_{T \in \mathcal{T}} \operatorname{osc}(R, r, T)^2.$$

The choice of W_{p+d} is such that, if the data is piecewise smooth (for example), then the local oscillation is of higher order than the local best-approximation error:

$$\frac{\operatorname{osc}(R, r, T)}{\inf_{v \in \mathcal{P}_p(T)} \|u - v\|_{1,T}} \to 0 \text{ as } |T| \to 0. \tag{1.6}$$

This is illustrated more explicitly in the following example.

Example 1.6 Suppose that A and $\boldsymbol{b}$ are piecewise constant and $c = 0$. In this case it holds that $(-\nabla \cdot A\nabla\hat{u} + \boldsymbol{b} \cdot \nabla\hat{u})|_T \in \mathcal{P}_{p-1}(T)$ and $A\nabla\hat{u}|_T \in [\mathcal{P}_{p-1}(T)]^d$, so we have

$$\operatorname{osc}(R, r, T)^2 = h_T^2 \inf_{\kappa \in \mathcal{P}_{p-1}(T)} \|f - \kappa\|_{0,T}^2 + \sum_{F \in \mathcal{F}_{N,T}} h_F \inf_{\kappa \in \mathcal{P}_{p-1}(F)} \|g - \kappa\|_{0,F}^2.$$

If $f \in H^p(T)$ and $g \in H^p(F)$, then $\operatorname{osc}(R, r, T) = \mathcal{O}(h_T^{p+1/2})$. If $f \in H^p(T)$ and $g \in \mathcal{P}_{p-1}(F)$, then $\operatorname{osc}(R, r, T) = \mathcal{O}(h_T^{p+1})$. Finally, if $f \in \mathcal{P}_{p-1}(T)$ and $g \in \mathcal{P}_{p-1}(F)$, then $\operatorname{osc}(R, r, T) = 0$.

1.3 Outline of Paper

The rest of the paper is organized as follows. In Sect. 2 we provide additional notation, define the auxiliary finite element space W, and state and prove some crucial properties of this space. With these results established, we prove Theorem 1.4 in Sect. 2.3. In Sect. 3 we discuss the computational complexity of the auxiliary problem, including the size and structure of the resulting system as well as its spectral properties. The proof of Theorem 1.5 is presented here. Finally, in Sect. 4, we give several numerical experiments verifying the theoretical results.

2 Reliability Analysis

2.1 Local and Global Polynomial Spaces for Estimating Error

Given a (non-degenerate) simplex $T \subset \mathbb{R}^d$ of diameter h_T, we define $\mathcal{S}_j(T)$, $0 \le j \le d$ to be the set of sub-simplices of T of dimension j; its cardinality is $|\mathcal{S}_j(T)| = \binom{d+1}{j+1}$. We denote by $\mathcal{S}_j$ the set of sub-simplices of the triangulation of dimension j, and point out the overlap of notation $\mathcal{S}_{d-1} = \mathcal{F}_I \cup \mathcal{F}_D \cup \mathcal{F}_N$ and $\mathcal{S}_d = \mathcal{T}$. Recall that $\mathcal{P}_m(S)$ is the set of polynomials of total degree $\le m$ with domain S, and note that $\dim \mathcal{P}_m(S) = \binom{m+j}{j}$ for $S \in \mathcal{S}_j(T)$. Taking the vertices of T to be $\{z_0, \ldots, z_d\}$; we let $\lambda_i \in \mathcal{P}_1(\mathbb{R}^d)$, $0 \le i \le d$, be the corresponding barycentric coordinates, uniquely defined by the relations $\lambda_i(z_j) = \delta_{ij}$. We let the *face* $F_j \in \mathcal{S}_{d-1}(T)$ be the sub-simplex not containing z_j, and n_j be the outward-pointing unit normal to F_j.

Definition 2.1 (*Element and face bubbles*) The fundamental element and face bubbles for T are given by ($j = 0, 1, \ldots, d$)

$$b_T = \prod_{k=0}^{d} \lambda_k \in \mathcal{P}_{d+1}(T), \quad b_{F_j} = \prod_{\substack{k=0 \\ k \ne j}}^{d} \lambda_k \in \mathcal{P}_d(T). \tag{2.1}$$

We also define general volume and face bubbles of degree m,

$$Q_{m,T} = \{v = b_T w \in \mathcal{P}_m(T) : \ w \in \mathcal{P}_{m-d-1}(T)\} \tag{2.2}$$

$$Q_{m,F_j} = \{v = b_{F_j} w \in \mathcal{P}_m(T) : \ w \in \mathcal{P}_{m-d}(T)\} \ominus Q_{m,T}. \tag{2.3}$$

Here and below, we use the shorthand $W_1 \ominus W_2 = \text{span}\{W_1 \setminus W_2\}$ for vector spaces W_1 and W_2. So $W_1 \ominus W_2$ is the largest subspace of W_1 that has only 0 in common with W_2.

The functions in $Q_{m,T}$ are precisely those in $\mathcal{P}_m(T)$ that vanish on ∂T; and the functions in Q_{m,F_j} are precisely those in $\mathcal{P}_m(T)$ that vanish on $\partial T \setminus F_j$, with the additional constraint that, if $v \in Q_{m,F_j}$ and v vanishes on F_j, then v vanishes on T. It is evident from their definitions that

$$Q_{m,T} \cap Q_{m,F_j} = \{0\}, \quad Q_{m,F_i} \cap Q_{m,F_j} = \{0\} \text{ for } i \neq j, \tag{2.4}$$

$$\dim Q_{m,T} = \dim \mathcal{P}_{m-d-1}(T) = \binom{m-1}{d}, \tag{2.5}$$

$$\dim Q_{m,F_j} = \dim \left(\mathcal{P}_{m-d}(T) \ominus \mathcal{P}_{m-d-1}(T)\right) = \binom{m}{d} - \binom{m-1}{d} = \binom{m-1}{d-1}. \tag{2.6}$$

Here and elsewhere, we use the conventions that $\binom{n}{k} = 0$ when $k > n$, and $\mathcal{P}_n = \{0\}$ when $n < 0$. It will be useful to characterize the volume and face bubbles in terms of moments, as we do in the following lemma.

Lemma 2.2 *A function $v \in Q_{m,T}$ is uniquely determined by the moments*

$$\int_T v\kappa, \quad \forall \kappa \in \mathcal{P}_{m-d-1}(T), \tag{2.7}$$

and a function $v \in Q_{m,F_j}$ is uniquely determined by the moments

$$\int_{F_j} v\kappa, \quad \forall \kappa \in \mathcal{P}_{m-d}(F_j). \tag{2.8}$$

Proof As is shown, for example, in [4], a function $v \in \mathcal{P}_m(T)$ is uniquely determined by the moments

$$\int_S v\kappa, \quad \forall \kappa \in \mathcal{P}_{m-\ell-1}(S) \ , \ \forall S \in \mathcal{S}_\ell(T) \ , \ 0 \leq \ell \leq d, \tag{2.9}$$

where $\int_S v\kappa$ with $S \in \mathcal{S}_0(T)$ is understood to be the evaluation of v at the vertex S. Since $v \in Q_{m,T}$ vanishes on S for $S \in \mathcal{S}_j(T)$ and $j < d$, v is determined by the moments on T alone. Similarly, any $v \in \{v = b_{F_j} w \in \mathcal{P}_m(T) : \ w \in \mathcal{P}_{m-d}(T)\}$ is uniquely determined by its moments on T and F_j, so any $v \in Q_{m,F_j}$ is uniquely determined by its moments on F_j alone. □

Definition 2.3 (*Local error space*) Given $m \in \mathbb{N}$, we define the local space

$$R_m(T) = Q_{m,T} \oplus \left(\bigoplus_{j=0}^{d} Q_{m-1,F_j} \right). \tag{2.10}$$

Given $p \in \mathbb{N}$, we define the local error space

$$W_{p+d}(T) = (Q_{p+d,T} \ominus Q_{p,T}) \oplus \left(\bigoplus_{j=0}^{d} (Q_{p+d-1,F_j} \ominus Q_{p,F_j}) \right) = R_{p+d}(T) \ominus R_p(T), \tag{2.11}$$

so that $\mathcal{P}_p(T) + R_{p+d}(T) = \mathcal{P}_p(T) \oplus W_{p+d}(T)$.

The dimension of $W_{p+d}(T)$ is readily deduced from (2.4)–(2.6),

$$\dim W_{p+d}(T) = \binom{p+d-1}{d} - \binom{p-1}{d} + (d+1)\left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1}\right). \tag{2.12}$$

We further note that, by Lemma 2.2 and its proof, a function $v \in R_{p+d}(T)$ is uniquely determined by the values

$$\int_S v\kappa \qquad \forall \kappa \in \mathcal{P}_{p-1}(S),\ \forall S \in \mathcal{S}_\ell(T),\ d-1 \le \ell \le d. \tag{2.13}$$

Remark 2.4 Starting with the standard basis for $\mathcal{P}_1(T)$, $\{\lambda_j : 0 \le j \le d\}$, a p-hierarchical basis for $\mathcal{P}_m(T)$, $m > 1$, is built from a p-hierarchical basis for $\mathcal{P}_{m-1}(T)$ by adding basis functions of degree m. Three approaches to such constructions, at least in $d = 2, 3$, are described in [11,15,34], with a useful summary of the constructions from [15,34] provided in [1]. In these constructions, hierarchical basis functions are associated with each sub-simplex $S \in \mathcal{S}_\ell$, $0 \le \ell \le d$, so it is simple in this setting to construct a basis for $W_{p+d}(T)$.

The corresponding global finite element spaces, defined by the degrees of freedom and local spaces, are given by

$$R_{p+d} := \{v \in H^1_{0,D}(\Omega) :\ v|_T \in R_{p+d}(T) \text{ for each } T \in \mathcal{T}\},$$
$$W_{p+d} := \{w \in H^1_{0,D}(\Omega) :\ v|_T \in W_{p+d}(T) \text{ for each } T \in \mathcal{T}\},$$

and we recall that the Lagrange finite element space V_p is defined by (1.3). Similar to the local setting, the global spaces satisfy the relation $R_{p+d} = R_p \oplus W_{p+d}$ so that $V_p + R_{p+d} = V_p \oplus W_{p+d}$.

2.2 A Quasi-Interpolant Based on Moment Conditions

Lemma 2.5 *Given $v \in H^1(\Omega)$, there exits a $\hat{v} \in V_p$ and $\hat{w} \in W_{p+d}$ such that*

(i) $\int_T (v - \hat{v} - \hat{w})\kappa = 0$ *for all* $\kappa \in \mathcal{P}_{p-1}(T)$ *and* $T \in \mathcal{T}$.
(ii) $\int_F (v - \hat{v} - \hat{w})\kappa = 0$ *for all* $\kappa \in \mathcal{P}_{p-1}(F)$ *and* $F \in \mathcal{F}_I \cup \mathcal{F}_N$.
(iii) $|v - \hat{v} - \hat{w}|_{m,T} \le C h_T^{1-m} |v|_{1,\Omega_T}$ *for* $m = 0, 1$, *where* Ω_T *is a local patch of elements containing* T.
(iv) $|v - \hat{v} - \hat{w}|_{0,F} \le C h_F^{1/2} |v|_{1,\Omega_F}$, *where* h_F *is the diameter of* $F \in \mathcal{F}$, *and* $\Omega_F = \Omega_T$ *for some* $T \in \mathcal{T}$ *with* $F \subset \partial T$.
(v) $|\hat{w}|_{1,T} \le C|v|_{1,\Omega_T}$ *for each* $T \in \mathcal{T}$.

The constant C can be taken to be a global constant that depends only on d, p and the shape-regularity of the $\mathcal{T}$.

Proof Since functions in $R_{p+d}(T)$ are uniquely determined by the values (2.13), for $m = 0, 1$ the function $\langle\!\langle \cdot \rangle\!\rangle_{m,T} : R_{p+d}(T) \to \mathbb{R}^+$ defined by

$$\langle\!\langle \phi \rangle\!\rangle_{m,T} = \max_{\substack{S \in \mathcal{S}_\ell(T) \\ d-1 \le \ell \le d}} \sup_{\kappa \in \mathcal{P}_{p-1}(S)} \frac{h_T^{d/2-\ell/2-m}}{\|\kappa\|_{0,S}} \int_S \phi\kappa \tag{2.14}$$

is a norm on $R_{p+d}(T)$.

Let $\tilde{T} = \{y = h_T^{-1}x :\ x \in T\}$, and for each $\psi : T \to \mathbb{R}$, define $\tilde{\psi} : \tilde{T} \to \mathbb{R}$ by $\psi(y) = \psi(h_T x)$. Analogous definitions are given for the sub-simplices of T and $\tilde{T}$ and

functions defined on them. It is clear that $|\phi|_{m,T} = h_T^{d/2-m}|\tilde{\phi}|_{m,\tilde{T}}$, where $|\cdot|_{0,T} = \|\cdot\|_{0,T}$. We also have for any $S \in \mathcal{S}_\ell(T)$

$$\frac{h_T^{d/2-\ell/2-m}}{\|\kappa\|_{0,S}}\int_S \phi\kappa = \frac{h_T^{d/2-\ell/2-m}}{h_T^{\ell/2}\|\tilde{\kappa}\|_{0,\tilde{S}}}\int_{\tilde{S}} \tilde{\phi}\tilde{\kappa}h_T^\ell = \frac{h_T^{d/2-m}}{\|\tilde{\kappa}\|_{0,\tilde{S}}}\int_{\tilde{S}}\tilde{\phi}\tilde{\kappa}.$$

Since $h_{\tilde{T}} = 1$, we see that $\langle\!\langle\phi\rangle\!\rangle_{m,T} = h_T^{d/2-m}\langle\!\langle\tilde{\phi}\rangle\!\rangle_{m,\tilde{T}}$. Therefore there exists a scale-invariant constant $C > 0$ that depends solely on p, d and m such that

$$|\phi|_{m,T} = h_T^{d/2-m}|\tilde{\phi}|_{m,\tilde{T}} \le Ch_T^{d/2-m}\langle\!\langle\tilde{\phi}\rangle\!\rangle_{m,\tilde{T}} = C\langle\!\langle\phi\rangle\!\rangle_{m,T}. \tag{2.15}$$

At this stage, we see that the local constant $C = C(T)$ in (2.15) may depend on the shape of T, but not its diameter. For the rest of the argument, we make a shape-regularity assumption on $\mathcal{T}$.

Next, denote by $\hat{v}_1 \in V_p$ the Scott-Zhang interpolant of v satisfying [32]

$$\|v - \hat{v}_1\|_{m,T} \le Ch_T^{1-m}|v|_{1,\Omega_T} \quad (m = 0, 1), \tag{2.16a}$$

$$\|v - \hat{v}_1\|_{0,\partial T} \le Ch_T^{1/2}|v|_{1,\Omega_T}, \tag{2.16b}$$

on each $T \in \mathcal{T}$. Set $\hat{v}_2 \in R_{p+d}$ such that

$$\int_S \hat{v}_2\kappa = \int_S (v - \hat{v}_1)\kappa \qquad \forall\kappa \in \mathcal{P}_{p-1}(S),\ \forall S \in \mathcal{S}_\ell,\ d-1 \le \ell \le d.$$

By (2.15) and (2.16) we find

$$\begin{aligned}
|\hat{v}_2|_{m,T} &\le C \max_{\substack{S\in\mathcal{S}_\ell(T)\\ d-1\le\ell\le d}} \sup_{\kappa\in\mathcal{P}_{p-1}(S)} \frac{h_T^{d/2-\ell/2-m}}{\|\kappa\|_{0,S}}\int_S \hat{v}_2\kappa \\
&= C \max_{\substack{S\in\mathcal{S}_\ell(T)\\ d-1\le\ell\le d}} \sup_{\kappa\in\mathcal{P}_{p-1}(S)} \frac{h_T^{d/2-\ell/2-m}}{\|\kappa\|_{0,S}}\int_S (v - \hat{v}_1)\kappa \\
&\le C\big(h_T^{1/2-m}\|v - \hat{v}_1\|_{0,\partial T} + h_T^{-m}\|v - \hat{v}_1\|_{0,T}\big) \le Ch_T^{1-m}|v|_{1,\Omega_T}.
\end{aligned}$$

Uniquely decomposing $\hat{v}_2$ as $\hat{v}_2 = \hat{v}_3 + \hat{w}$ with $\hat{v}_3 \in R_p$ and $\hat{w} \in W_{p+d}$, and setting $\hat{v} := \hat{v}_1 + \hat{v}_3$ so that $\hat{v} + \hat{w} = \hat{v}_1 + \hat{v}_2$, we see that properties (i)–(ii) clearly hold, and

$$\|v - \hat{v} - \hat{w}\|_{m,T} \le \|v - \hat{v}_1\|_{m,T} + \|\hat{v}_2\|_{m,T} \le Ch_T^{1-m}|v|_{1,\Omega_T}.$$

Therefore by standard trace inequalities and the shape regularity of the mesh, we also have on $F \subset \partial T$

$$\|v - \hat{v} - \hat{w}\|_{0,F} \le C\big(h_F^{-1/2}\|v - \hat{v} - \hat{w}\|_{0,T} + h_F^{1/2}|v - \hat{v} - \hat{w}|_{1,T}\big) \le Ch_F^{1/2}|v|_{1,\Omega_F}.$$

Hence, properties (iii)–(iv) are satisfied.

Finally, since $R_p(T) \cap W_{p+d}(T) = \{0\}$, the Strong Cauchy–Schwarz inequality [2] gives the existence of a constant $\gamma \in [0, 1)$ such that $\int_T \nabla\hat{w}\cdot\nabla\hat{v}_3 \le \gamma|\hat{w}|_{1,T}|\hat{v}_3|_{1,T}$. Consequently, we have

$$\begin{aligned}
|\hat{v}_2|_{1,T}^2 &= |\hat{w}|_{1,T}^2 + |\hat{v}_3|_{1,T}^2 + 2\int_T \nabla\hat{w}\cdot\nabla\hat{v}_3 \\
&\ge |\hat{w}|_{1,T}^2 + |\hat{v}_3|_{1,T}^2 - 2\gamma|\hat{w}|_{1,T}|\hat{v}_3|_{1,T} \ge (1-\gamma^2)|\hat{w}|_{1,T}^2.
\end{aligned}$$

Therefore we find $|\hat{w}|_{H^1(T)} \le \sqrt{(1-\gamma^2)^{-1}}|\hat{v}_2|_{H^1(T)} \le C|v|_{H^1(\Omega_T)}$. □

Remark 2.6 The moment conditions (i)-(ii) of Lemma 2.5 imply the conditions

$$\int_T \nabla(v - \hat{v} - \hat{w}) \cdot \phi = 0 \text{ for all } \phi \in RT_{p-1}(T) \text{ and all } T \in \mathcal{T}, \tag{2.17}$$

where $RT_{p-1}(T) = x\mathcal{P}_{p-1}(T) + [\mathcal{P}_{p-1}(T)]^d = \{\phi = \sum_{j=0}^{d}(x - z_j)\kappa_j : \kappa_j \in \mathcal{P}_{p-1}(T)\}$ is the local Raviart-Thomas space. Recalling the vertex, face and normal vector notation above, this equivalence is most readily seen through the following simple consequence of integration-by-parts on a simplex:

$$\text{For } f \in H^1(T), \int_T (x - z_j) \cdot \nabla f = a_j \int_{F_j} f - d \int_T f, \tag{2.18}$$

where a_j is the distance (altitude) between z_j and F_j. Choosing $f = (v - \hat{v} - \hat{w})\kappa_j$ for $\kappa_j \in \mathcal{P}_{p-1}(T)$, and combining results for each j, makes the comparison between (2.17) and (i)-(ii) apparent. The conditions (2.17) are not independent, so they do not impose $\dim(RT_{p-1}(T)) = d\binom{p+d-1}{d} + \binom{p+d-2}{d-1}$ independent constraints on $R_{p+d}(T)$, whose dimension, $\binom{p+d-1}{d} + (d+1)\binom{p+d-2}{d-1}$, is generally smaller.

2.3 Proof of Theorem 1.4

Proof of Theorem 1.4 Combining Proposition 1.2 and Lemma 2.5, we determine that

$$\begin{aligned}
|B(u - \hat{u}, v)| &\le |B(\varepsilon, \hat{w})| + \sum_{T \in \mathcal{T}} \|v - \hat{v} - \hat{w}\|_{0,T} \inf_{\kappa \in \mathcal{P}_{p-1}(T)} \|R_T - \kappa\|_{0,T} \\
&\quad + \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_N} \|v - \hat{v} - \hat{w}\|_{0,F} \inf_{\kappa \in \mathcal{P}_{p-1}(F)} \|r_F - \kappa\|_{0,F} \\
&\le C\|\varepsilon\|_1 \|\hat{w}\|_1 + C \sum_{T \in \mathcal{T}} h_T \|v\|_{1,\Omega_T} \inf_{\kappa \in \mathcal{P}_{p-1}(T)} \|R_T - \kappa\|_{0,T} \\
&\quad + C \sum_{F \in \mathcal{F}_I \cup \mathcal{F}_N} h_F^{1/2} \|v\|_{1,\Omega_T} \inf_{\kappa \in \mathcal{P}_{p-1}(F)} \|r_F - \kappa\|_{0,F} \\
&\le C \left(\|\varepsilon\|_1 \|v\|_1 + \operatorname{osc}(R, r, \mathcal{T})\|v\|_1\right).
\end{aligned}$$

For the final inequality, we have used Lemma 2.5 (v), the (discrete) Cauchy–Schwarz inequality and the bounded overlap of the patches Ω_T and Ω_F (which is also a consequence of shape-regularity). Finally, we choose $v = u - \hat{u}$ and use the coercivity of B to complete the proof.

Remark 2.7 We note that the continuity constant enters in the bound $|B(\varepsilon, \hat{w})| \le \mathfrak{C}\|\varepsilon\|_{1,\Omega}\|\hat{w}\|_1$, and only affects the term $\|\varepsilon\|_1$ in the reliability bound of Theorem 1.4. The coercivity constant $\mathfrak{c}$ affects both terms in the reliability bound.

Remark 2.8 Although our approach is analyzed as an h-method with global fixed p, the general approach is very naturally adjusted to both p and hp-methods. As indicated in the introduction, the driving motivation for the choice of $W_{p+d}(T)$ is to make sure that the local oscillation is of higher order than the local best approximation error. The development suggests that, if the local approximation space is $V(T) = \mathcal{P}_{p_T}(T)$, then the local error space $W(T)$ should be spanned by face bubbles of degree $p_T + d - 1$ and interior bubbles of degree $p_T + d$ that are not already represented in $\mathcal{P}_{p_T}(T)$. Again, although our approach is analyzed for simplicial elements, the shapes of the elements are irrelevant for much of our

development. In particular, it is straight-forward to apply the prescription above for choosing $W(T)$ on tensor-product elements such as quadrilaterals or bricks. In Sect. 4 we investigate our approach as a p-method on meshes that include tensorial elements.

2.4 Scalar Diffusion Problems with Jump-Discontinuities

An important subclass of problems of type (1.1) are those for which $\boldsymbol{b} = \mathbf{0}$ and $A = aI$ where a is piecewise-constant with respect to a polyhedral partition of Ω, so $B(v, w) = \int_\Omega a\nabla v\cdot\nabla w$. We define the corresponding energy norm $|\!|\!| \cdot |\!|\!|$ by $|\!|\!| v |\!|\!|^2 = B(v, v)$. For a d-dimensional subset S (think of a single simplex or a union of adjacent simplices), we define $|\!|\!| \cdot |\!|\!|_S$ by $|\!|\!| v |\!|\!|_S^2 = \int_S a\nabla v \cdot \nabla v$. Let a_T denote the value of a on T, and a_F denote the maximum a_T among the (one or two) T having F as a face.

It was shown in [10,30] that, if a satisfies a *quasi-monotonicity property*, there is a quasi-interpolation operator $\Pi : H^1_{0,D}(\Omega) \to V_1$, such that

$$a_T^{1/2}|v - \Pi v|_{m,T} \le Ch_T^{1-m}|\!|\!| v |\!|\!|_{\Omega_T}, \quad a_F^{1/2}|v - \Pi v|_{0,F} \le Ch_F^{1/2}|\!|\!| v |\!|\!|_{\Omega_T} \tag{2.19}$$

where C depends only on the shape-regularity of the mesh, and $m = 0, 1$. In particular, C is independent of the jumps in a. Although [10,30] focus on the case $c = 0$, it is obvious from their arguments that more general $c \ge 0$ If we follow through essentially the same arguments in the proof of Lemma 2.5, taking $\hat{v}_1 = \Pi v$, we deduce the following variants of (iii)–(iv),

$$\begin{aligned} &a_T^{1/2}|v - \hat{v} - \hat{w}|_{m,T} \le Ch_T^{1-m}|\!|\!| v |\!|\!|_{\Omega_T}\ ,\ a_F^{1/2}|v - \hat{v} - \hat{w}|_{0,F} \le Ch_F^{1/2}|\!|\!| v |\!|\!|_{\Omega_F}\ , \\ &|\!|\!| \hat{w} |\!|\!|_T \le C|\!|\!| v |\!|\!|_{\Omega_T}. \end{aligned} \tag{2.20}$$

Revisiting the proof of Theorem 1.4 in this setting, we obtain

$$|\!|\!| \varepsilon |\!|\!| \le |\!|\!| u - \hat{u} |\!|\!| \le C\left(|\!|\!| \varepsilon |\!|\!| + \mathrm{osc}_a(R, r, \mathcal{T})\right), \tag{2.21}$$

where C is independent of the data of the problem. As before, C may depend on p and d. The residual oscillation term in this case is given by

$$\begin{aligned} \mathrm{osc}_a(R, r, \mathcal{T})^2 = &\sum_{T\in\mathcal{T}} a_T^{-1}h_T^2 \inf_{\kappa\in\mathcal{P}_{p-1}(T)} \|R_T - \kappa\|_{0,T}^2 \\ &+ \sum_{F\in\mathcal{F}_I\cup\mathcal{F}_N} a_F^{-1}h_F \inf_{\kappa\in\mathcal{P}_{p-1}(F)} \|r_F - \kappa\|_{0,F}^2. \end{aligned} \tag{2.22}$$

We point the reader to [30] for a thorough discussion of the quasi-monotonicity property, and which cases it includes, but note that this property is violated by the Kellogg problem (see Sect. 4.2.3). Despite the fact that the argument provided here does not cover that case, we see empirically that $|\!|\!| \varepsilon |\!|\!|$ is a robust estimator of $|\!|\!| u - \hat{u} |\!|\!|$ for such problems as well. We finally remark that, while [10,30] focus on the case of pure diffusion problems, $c = 0$, their results extend to more general reaction-diffusion problems, $c \ge 0$, with no difficulty. As such, the error bounds (2.21) hold as stated in this case as well, with the obvious adjustment to the definition of the energy norm. We also point to [10] for analysis that allows piecewise-smooth diffusion coefficients, having jumps as before, and again (2.21) holds.

3 Computational Considerations

As presented above, the computation of ε requires the solution of a global system involving the stiffness matrix associated with W_{p+d}. At first glance this would seem to rule out the approach

as too expensive for practical computations, but we argue herein that this is not the case. Our argument is based on considerations of sparsity structure and size of the linear systems, and on their spectral properties. Using standard (p-hierarchical) bases for the spaces V_p and W_{p+d}, we compare and contrast the corresponding global and element stiffness matrices. We assume that global stiffness matrices are assembled by summing contributions from element stiffness matrices computed on each simplex $T \in \mathcal{T}$.

3.1 Size and Sparsity Structure

We begin by comparing the sizes of the element stiffness matrices for $V_p(T)$ and $W_{p+d}(T)$, as well as the amount of information which must be transferred to the global stiffness matrices in each case if static condensation is used locally to eliminate interior degrees of freedom. Letting $n = n(p,d)$ and $m = m(p,d)$ be the number of degrees of freedom associated with $V_p(T)$ and $W_{p+d}(T)$, respectively, and $\hat{n} = \hat{n}(p,d)$ and $\hat{m} = \hat{m}(p,d)$ denote the analogous quantities after interior degrees of freedom have been eliminated, we have

$$n = \binom{p+d}{d}, \quad m = \binom{p+d-1}{d} - \binom{p-1}{d} + (d+1)\left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1}\right), \tag{3.1}$$

$$\hat{n} = \binom{p+d}{d} - \binom{p-1}{d}, \quad \hat{m} = (d+1)\left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1}\right). \tag{3.2}$$

We note that n is a polynomial of degree d in p and m is a polynomial of degree $d-1$ in p, so it is clear that $n > m$ when p is large enough, for any fixed d. The polynomial degrees for $\hat{n}$ and $\hat{m}$ are of degrees $d-1$ and $d-2$ in p, respectively. In Table 1 we list values of the the four quantities (3.1)–(3.2) for $1 \le p \le 7$ and $d = 2, 3$.

Recall that $\mathcal{S}_j$ denotes the set of subsimplices of dimension j in $\mathcal{T}$, and $|\mathcal{S}_j|$ denotes its cardinality, $0 \le j \le d$. Without static condensation to eliminate the degrees of freedom associated with the interiors of each $T \in \mathcal{T}$, the sizes of the global stiffness matrices for V_p and W_{p+d} are, respectively,

$$N = \sum_{j=0}^{d} |\mathcal{S}_j| \binom{p-1}{j}, \quad M = |\mathcal{S}_d| \left(\binom{p+d-1}{d} - \binom{p-1}{d}\right) + |\mathcal{S}_{d-1}| \left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1}\right).$$

Table 1 Size of the local stiffness matrices for $V_p(T)$ and $W_{p+d}(T)$ with and without static condensation, for $d = 2, 3$

	$d=2$				$d=3$			
p	n	m	$\hat{n}$	$\hat{m}$	n	m	$\hat{n}$	$\hat{m}$
1	3	4	3	3	4	5	4	4
2	6	6	6	3	10	16	10	12
3	10	8	9	3	20	30	20	20
4	15	10	12	3	35	47	34	28
5	21	12	15	3	56	67	52	36
6	28	14	18	3	84	90	74	44
7	36	16	21	3	120	116	100	52

When the interior degrees of freedom are eliminated, the sizes become

$$\hat{N} = \sum_{j=0}^{d-1} |\mathcal{S}_j| \binom{p-1}{j}, \quad \hat{M} = |\mathcal{S}_{d-1}| \left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1} \right).$$

The formulas count degrees of freedom on Γ_D, though these are not truly unknowns in the problem, because many practical implementations proceed in this way when assembling global matrices, and encode Dirichlet boundary conditions in the system as a final step. Recognizing that $\binom{p-1}{d}$ and $\binom{p-1}{d-1}$ are polynomials of degree d and $d-1$ in p, respectively, we see again that, for any fixed d, $N > M$ and $\hat{N} > \hat{M}$ for sufficiently large p. To illustrate this, consider a standard uniform triangulation of the unit square by isosceles right triangles (half-squares) with side-length $1/s$. For such triangulations, $N = (p(s-1)+1)^2$ and $M = 4p(s-1)^2 + (s^2-1)$, so $N > M$ for all $s \geq 2$ when $p \geq 5$. For such triangulations we also have $\hat{N} = (3p-2)s^2 - 4(p-1)s + p - 1$ and $\hat{M} = 3s^2 - 4s + 1$, so $\hat{N} > \hat{M}$ for all $s \geq 2$ when $p \geq 2$.

We now turn to the discussion of sparsity for the global matrices for V_p and W_{p+d}. Given $S \in \mathcal{S}_j$, let $\mathcal{T}_S$ be the set of simplices $T \in \mathcal{T}$ which have S as a sub-simplex. We also define $\mathcal{S}_i(\mathcal{T}_S) = \cup_{T \in \mathcal{T}_S} \mathcal{S}_i(T)$ and denote its cardinality by $|\mathcal{S}_i(\mathcal{T}_S)|$. If $S = T \in \mathcal{S}_d = \mathcal{T}$, then $\mathcal{T}_S = \{T\}$ and $|\mathcal{S}_i(\mathcal{T}_S)| = \binom{d+1}{i+1}$. If $S = F \in \mathcal{S}_{d-1}$, then $\mathcal{T}_S$ consists of the one or two simplices which have F as a face; in the first case $|\mathcal{S}_i(\mathcal{T}_S)| = \binom{d+1}{i+1}$ as before, and in the second $|\mathcal{S}_i(\mathcal{T}_S)| = 2\binom{d+1}{i+1} - \binom{d}{i+1}$. For $j < d-1$, the cardinalities of $\mathcal{T}_S$ and $\mathcal{S}_i(\mathcal{T}_S)$ for a given $S \in \mathcal{S}_j$ cannot be determined *a priori* for general unstructured meshes. To compute the sparsity structure of the global stiffness matrix, the sets $\mathcal{S}_i(\mathcal{T}_S)$ for each $S \in \mathcal{S}_j$ (and each i and j) must be determined, at least indirectly. For W_{p+d} this task is greatly simplified by the fact that we need only consider $\mathcal{S}_i(\mathcal{T}_S)$ for each $S \in \mathcal{S}_j$ with $i, j \in \{d-1, d\}$—these are the two cases which are easiest to resolve! More specifically, let $\phi \in W_{p+d}$ be a basis function, and let S be the (sub-)simplex of minimal dimension $j \in \{d-1, d\}$ on which ϕ does not vanish identically. The number of possible non-zeros in the the row of the matrix corresponding to ϕ is

$$|\mathcal{S}_{d-1}(\mathcal{T}_S)| \left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1} \right) + |\mathcal{S}_{d-1}(\mathcal{T}_S)| \left(\binom{p+d-1}{d} - \binom{p-1}{d} \right).$$

If static condensation is used, the number of non-zeros in a row for ϕ associated with an interior face is

$$(2d+1) \left(\binom{p+d-2}{d-1} - \binom{p-1}{d-1} \right).$$

For a boundary face, $2d+1$ is replaced by $d+1$. We see that, in the case of W_{p+d}, the sparsity structure is known in advance. For example, when $d = 2$ and static condensation is used, the number of non-zeros in any row does not exceed 5, regardless of p and the mesh topology. When $d = 3$ and static condensation is used, the number of non-zeros in any row does not exceed $7(2p-1)$.

For comparison, we briefly discuss the situation for V_p. Let $\phi \in V_p$ be a basis function, and let S be the (sub-)simplex of minimal dimension j on which ϕ does not vanish identically. The number of possible non-zeros in the the row of B corresponding to ϕ, and the total number of possible non-zeros are, respectively,

$$\sum_{i=0}^{d} |\mathcal{S}_i(\mathcal{T}_S)| \binom{p-1}{i}, \quad \sum_{j=0}^{d} \sum_{S\in\mathcal{S}_j} \sum_{i=0}^{d} |\mathcal{S}_i(\mathcal{T}_S)| \binom{p-1}{i}.$$

If static condensation is used to eliminate interior degrees of freedom, the sums are terminated at $d-1$ instead instead of d.

3.2 Spectral Behavior of the Stiffness Matrix for W_{p+d}

We argue in Theorem 1.5 and Remark 3.4 below that the spectral behavior of the global stiffness matrix for W_{p+d}, with or without static condensation, makes it amenable to solution techniques which are simpler/faster than those for V_p. In brief, the conditioning of the stiffness matrix for W_{p+d}, perhaps after simple diagonal rescaling, does not deteriorate as the triangulation is refined, unlike that for V_p. Before specifically commenting on the spectral behavior of the global stiffness matrix for W_{p+d}, we first make comparison with matrices arising from the H^1-inner-product for a general finite dimensional subspace $X \subset H^1_{0,D}(\Omega)$, having basis $\{\phi_i : 1 \le i \le N\}$. We define the stiffness matrices

$$B_{ij} = B(\phi_j, \phi_i), \quad \hat{B}_{ij} = \big(\phi_j, \phi_i\big)_1 = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i + \phi_j\phi_i.$$

Making the obvious identification between $\boldsymbol{v} \in \mathbb{R}^N$ and $v \in X$, we see that

$$B(v, w) = \boldsymbol{w}^t B \boldsymbol{v}, \quad (v, w)_1 = \boldsymbol{w}^t \hat{B} \boldsymbol{v}.$$

Stated in terms of the matrices B and $\hat{B}$, the continuity and coercivity of the bilinear form B are

$$|\boldsymbol{w}^t B \boldsymbol{v}| \le \mathfrak{C} \left(\boldsymbol{v}^t \hat{B} \boldsymbol{v}\right)^{1/2} \left(\boldsymbol{w}^t \hat{B} \boldsymbol{w}\right)^{1/2}, \quad \boldsymbol{v}^t B \boldsymbol{v} \ge \mathfrak{c} \boldsymbol{v}^t \hat{B} \boldsymbol{v} \quad \forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^N.$$

Proposition 3.1 *Let $\mu = \mu_1 + i\mu_2$, $\mu_1, \mu_2 \in \mathbb{R}$, be an eigenvalue of B. Then*

$$\mathfrak{c}\lambda_{\min}(\hat{B}) \le \mu_1 \le \mathfrak{C}\lambda_{\max}(\hat{B}), \quad |\mu_2| \le \mathfrak{C}\lambda_{\max}(\hat{B}).$$

Proof Let $\boldsymbol{v} = \boldsymbol{v}_1 + i\boldsymbol{v}_2$, $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^N$, be an eigenvector for μ; and assume, without loss of generality, that $\|\boldsymbol{v}\| = 1$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{C}^N$. It is straightforward to show that

$$\boldsymbol{v}_1^t B \boldsymbol{v}_1 + \boldsymbol{v}_2^t B \boldsymbol{v}_2 = \mu_1, \quad \boldsymbol{v}_1^t B \boldsymbol{v}_2 - \boldsymbol{v}_2^t B \boldsymbol{v}_1 = \mu_2.$$

So we see that

$$\mathfrak{c}\lambda_{\min}(\hat{B}) \le \mathfrak{c}\left(\boldsymbol{v}_1^t \hat{B} \boldsymbol{v}_1 + \boldsymbol{v}_2^t \hat{B} \boldsymbol{v}_2\right) \le \mu_1 \le \mathfrak{C}\left(\boldsymbol{v}_1^t \hat{B} \boldsymbol{v}_1 + \boldsymbol{v}_2^t \hat{B} \boldsymbol{v}_2\right) \le \mathfrak{C}\lambda_{\max}(\hat{B}).$$

Furthermore,

$$|\mu_2| \le 2|\boldsymbol{v}_2^t B \boldsymbol{v}_1| \le 2\mathfrak{C}\left(\boldsymbol{v}_1^t \hat{B} \boldsymbol{v}_1\right)^{1/2} \left(\boldsymbol{v}_2^t \hat{B} \boldsymbol{v}_2\right)^{1/2} \le 2\mathfrak{C}\lambda_{\max}(\hat{B})\|\boldsymbol{v}_1\|\|\boldsymbol{v}_2\| \le \mathfrak{C}\lambda_{\max}(\hat{B}).$$

We have used the Cauchy Inequality, $2ab \le a^2 + b^2$ for the final inequality above. □

To get a better handle on the spectral properties of $\hat{B}$, we we now consider element stiffness matrices for X. Given $T \in \mathcal{T}$, let $I(T) = \{j : T \cap \text{supp}(\phi_j) \neq \emptyset\}$ and $n = n(T) = |I(T)|$. We define $\hat{B}_T \in \mathbb{R}^{n\times n}$ via

$$(\hat{B}_T)_{ij} = (\phi_j, \phi_i)_{1,T} = \int_T \nabla\phi_j \cdot \nabla\phi_i + \phi_j\phi_i \text{ for } i, j \in I(T).$$

Again making the obvious association between $\boldsymbol{v} \in \mathbb{R}^N$ and $v \in X$, we define $\boldsymbol{v}_T \in \mathbb{R}^n$ such that $\|v\|^2_{1,T} = \boldsymbol{v}_T^t \hat{B}_T \boldsymbol{v}_T$; it is clear that $\boldsymbol{v}_T$ consists of the components of $\boldsymbol{v}$ whose indices are in $I(T)$. We also define $\hat{D} = \text{diag}(\hat{B})$ and $\hat{D}_T = \text{diag}(\hat{B}_T)$. It is apparent from these definitions that

$$\boldsymbol{v}^t \hat{B} \boldsymbol{v} = \sum_{T\in\mathcal{T}} \boldsymbol{v}_T^t \hat{B}_T \boldsymbol{v}_T, \quad \boldsymbol{v}^t \hat{D} \boldsymbol{v} = \sum_{T\in\mathcal{T}} \boldsymbol{v}_T^t \hat{D}_T \boldsymbol{v}_T.$$

The next result follows immediately from this discussion.

Proposition 3.2 *Suppose there are constants $c, C > 0$ such that $c \leq \frac{\boldsymbol{w}^t \hat{B}_T \boldsymbol{w}}{\boldsymbol{w}^t \hat{D}_T \boldsymbol{w}} \leq C$ for all non-zero $\boldsymbol{w} \in \mathbb{R}^n$. Then $c \leq \frac{\boldsymbol{w}^t \hat{B} \boldsymbol{w}}{\boldsymbol{w}^t \hat{D} \boldsymbol{w}} \leq C$ for all non-zero $\boldsymbol{w} \in \mathbb{R}^N$. As a consequence, the spectrum of $\hat{D}^{-1/2} \hat{B} \hat{D}^{-1/2}$ is contained in $[c, C]$.*

Suppose $X = W_{p+d}$ and we use a hierarchical basis (cf. Remark 2.4). Fixing $T \in \mathcal{T}$ and using the corresponding basis for $W_{p+d}(T)$, we may use simple scaling arguments to see that that $\hat{B}_T$ can be expressed in the form

$$\hat{B}_T = h_T^{d-2} B_1 + h_T^d B_2,$$

where B_1, B_2 depend only on p, d and the shape-regularity of T. The matrix B_1, whose entries are $h_T^{2-d} \int_T \nabla\phi_j \cdot \nabla\phi_i$, has full-rank because $(\cdot,\cdot)_{1,T}$ is an inner-product on $W_{p+d}(T)$. The matrix B_2, whose entries are $h_T^{-d} \int_T \phi_j\phi_i$ is clearly a full-rank Gram matrix. This implies that there are constants $c_T, C_T > 0$ depending only on p, d and the shape-regularity of T for which

$$c_T \boldsymbol{w}^t \hat{D}_T \boldsymbol{w} \leq \boldsymbol{w}^t \hat{B}_T \boldsymbol{w} \leq C_T \boldsymbol{w}^t \hat{D}_T \boldsymbol{w} \text{ for all } \boldsymbol{w} \in \mathbb{R}^n. \tag{3.3}$$

Invoking the shape-regularity of the family $\{\mathcal{T}\}$, we can replace the local constants c_T, C_T with universal constants c, C and apply Proposition 3.2. We are now ready to prove Theorem 1.5 our key result concerning the spectral properties of B for W_{p+d}:

Proof of Theorem 1.5 Letting $\hat{B}$ and $\hat{D}$ be as in the discussion above, and D be the diagonal of B, we have already seen in Proposition 3.1 that B and $\hat{B}$ are spectrally equivalent to each other. It is trivial to see that D and $\hat{D}$ are spectrally equivalent to each other. So, to prove that B and D are spectrally equivalent to each other, we need merely show that $\hat{B}$ and $\hat{D}$ are spectrally equivalent to each other. But this was established by Proposition 3.2 and the discussion that followed. □

Remark 3.3 Had we chosen $X = V_p$, the corresponding matrix B_1 has a one-dimensional nullspace spanned by the vector $\boldsymbol{v} \in \mathbb{R}^n$ that corresponds to the constant function $v = 1$ in $V_p(T)$. We deduce that $\boldsymbol{v}^T \hat{B}_T \boldsymbol{v} = \int_T 1 = |T| \sim h_T^d$, whereas $\boldsymbol{v}^T \hat{D}_T \boldsymbol{v} \sim h_T^{d-2}$. For any other non-zero $\boldsymbol{w} \in \mathbb{R}^n$, $\boldsymbol{w}^T \hat{B}_T \boldsymbol{w}$ and $\boldsymbol{w}^T \hat{D}_T \boldsymbol{w}$ scale in precisely the same way, so there are no scale-invariant c_T, C_T for which (3.3) holds. Therefore, Proposition 3.2 cannot be applied.

Remark 3.4 (**Effect of static condensation**) To analyze the effect of static condensation on the global stiffness matrix B for W_{p+d}, we split the space as $W_{p+d} = W_{p+d,1} \oplus W_{p+d,2}$, where $W_{p+d,1}$ is spanned by the "interior" basis functions—those supported on a single element. This splitting of the space induces the natural 2×2 block structure on B

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

and we must investigate the spectral properties of the Schur complement $S = B_{22} - B_{21} B_{11}^{-1} B_{12}$. Given $\boldsymbol{z} \in \mathbb{R}^{\hat{M}}$, we extend it to a vector $\tilde{\boldsymbol{z}} \in \mathbb{R}^{M}$ by appending it to the vector $-B_{11}^{-1} B_{12} \boldsymbol{z} \in \mathbb{R}^{M-\hat{M}}$. For $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{\hat{M}}$, we have $\boldsymbol{w}^t S \boldsymbol{v} = \tilde{\boldsymbol{w}}^t B \tilde{\boldsymbol{v}}$, so our analysis above suffices to show that the spectral properties of S cannot be worse than those of B.

4 Numerical Experiments

We recall that our results concerning the reliability and computational cost of our estimator were obtained for fixed p and (adaptive) h-refinement on simplicial meshes, with reliability shown in the H^1-norm. In the first subsection, we numerically illustrate these results on a standard test problem in $\mathbb{R}^2$ for modest p. The second subsection is devoted to extensive testing of the robustness of the estimator with respect to polynomial degree. Here we consider the behavior of the error estimator under uniform p-refinement on fixed (adapted) meshes of quadrilaterals/bricks and/or simplices for several different types of problems, one of which is in $\mathbb{R}^3$. We did not use static condensation for any of the linear systems.

A key measure of the quality of the estimator is its *effectivity* in a norm of interest,

$$\text{EFF} = \|\varepsilon\| / \|u - \hat{u}\|.$$

In most cases, we will report effectivities in the global appropriate global energy norm, because our theory deals with such cases. But as a matter of interest, for the h-refinement study we also report global L^2-effectivity and local H^1-effectivity—the latter of which provides a good measure of the efficiency of local indicators $\|\varepsilon\|_{1,T}$ for driving an adaptive algorithm. The global error estimates are typically within a factor of two of the actual errors, indicating good effectivity of the estimator. In the case of lower-order elements and h-refinement, the experiments suggest asymptotic exactness (i.e. EFF$\to$ 1) in some cases, which is consistent with analysis presented in [28] for a similar estimator, and can be seen for a variety of recovery-type estimators for lower-order elements. We are not aware of any asypomptotic exactness results for any type of estimator in the p-version or hp-versions of finite elements, and numerical results in [7,9] suggest that such results should not be expected at least for recovery-type estimators. Even the p-robust family of estimators discussed in [21] never claims effectivities very near one.

4.1 Verification of Properties of the Estimator Under Adaptive h-Refinement

We study the performance of the a posteriori error estimator with an h-refinement algorithm for fixed $p = 1, 2, 3$ on a few model problems that exhibit typical challenges, points-singularities and internal layers or boundary layers. Starting from a coarse mesh, triangles are marked for refinement as follows: denoting by $T_{\max} \in \mathcal{T}$ the simplex with the largest estimated error, i.e. $\|\varepsilon\|_{H^1(T_{\max})} \geq \|\varepsilon\|_{H^1(T)}\ \forall T \in \mathcal{T}$, we mark T for refinement if

$\|\varepsilon\|_{H^1(T)} \geq \gamma \|\varepsilon\|_{H^1(T_{\max})}$, where $\gamma \in [0, 1]$ is some user-defined parameter. In all of the tests in this subsection we take $\gamma = 0.3$.

The problems having boundary or internal layers are for moderate convection-dominance,

$$-\epsilon \Delta u + \boldsymbol{b} \cdot \nabla u + cu = f,$$

where $\|\boldsymbol{b}\|_{L^\infty} \approx 1$ and $\epsilon = 10^{-3}$. We apply the same adaptivity strategy described above but replace the local H^1-norm by the corresponding local energy norm for marking elements. We remark that we have not proved that our estimator is robust with respect to ϵ; this will merely be demonstrated in one of the examples below. A proper treatment of convection-dominated problems in the context of a posteriori error estimation typically involves different (stabilized) discretization schemes (cf. [3,36,37]); this is beyond the scope of the current paper. The purpose of these examples is to illustrate that no change in our error estimation approach is needed for such problems if ϵ is within a modest range that nonetheless allows for challenging layer-phenomena.

4.1.1 Corner Singularity, the L-shaped Domain

We consider a prototypical problem on the L-shaped domain

$$-\Delta u = f \text{ in } \Omega = (-1, 1)^2 \backslash (0, 1) \times (-1, 0), \quad u = 0 \text{ on } \partial\Omega,$$

with f chosen so that the exact solution is given by $u = r^{2/3} \sin(\frac{2}{3}\theta)(x_1^2 - 1)(x_2^2 - 1)$. This solution exhibits the typical singular behavior at the origin for generic f. We note that, in this case, the oscillation term in the reliability bound reduces to purely data oscillation which has the local form $\operatorname{osc}(R, r, T) = h_T \inf_{\kappa \in \mathcal{P}_{p-1}(T)} \|f - \kappa\|_{0,T}$.

Starting from a uniform mesh with $h = 1/8$, relevant data was collected for a sequence of 20 nested meshes obtained by the adaptive scheme described above for each p. Global H^1 and L^2 effectivities are given in Fig. 1. The global effectivities in both norms are quite good, with some indication of asymptotic exactness (or at least effectivities very near 1) in H^1 for each p, and in L^2 when $p \geq 2$. In terms of local H^1 effectivities, we observe that the maximum local effectivities range from $[1.02, 2.17]$ for all tested polynomial degrees and for all meshes, which bodes well for their efficiency as local indicators for driving adaptive refinement. The H^1 error against the (total) degrees of freedom are given in Fig. 2. In the cases $p = 1, 2$ we observe the optimal convergence rate $|\dim V_p + \dim W_{p+d}|^{-p/2}$. In the cubic case ($p = 3$), convergence seems better than the optimal rate, which indicates that we are still in the pre-asymptotic regime at this stage.

As a verification of the claims of Sect. 3 we briefly summarize the ratios $\dim W_{p+d} / \dim V_p$ and the condition numbers of the diagonally-rescaled stiffness matrices for W_{p+d}, $B \to D^{-1/2} B D^{-1/2}$. In all cases the largest value of the dimension ratio corresponds to the coarsest mesh, and the smallest ratio to the finest mesh. For $p = 1$, we have $\dim W_{p+d} / \dim V_p \in [5.09, 5.77]$, for $p = 2$ the ratios were in $[2.32, 2.41]$, and for $p = 3$ the ratios were in $[1.49, 1.51]$. Again, for each p and all meshes the computed condition numbers for W_{p+d} remained in a relatively narrow range, neither monotonically increasing nor decreasing as the mesh was refined. For $p = 1$ the range of condition numbers was $[28.4, 37.3]$, for $p = 2$ this range was $[15.0, 16.9]$, and for $p = 3$ this range was $[30.1, 33.2]$.

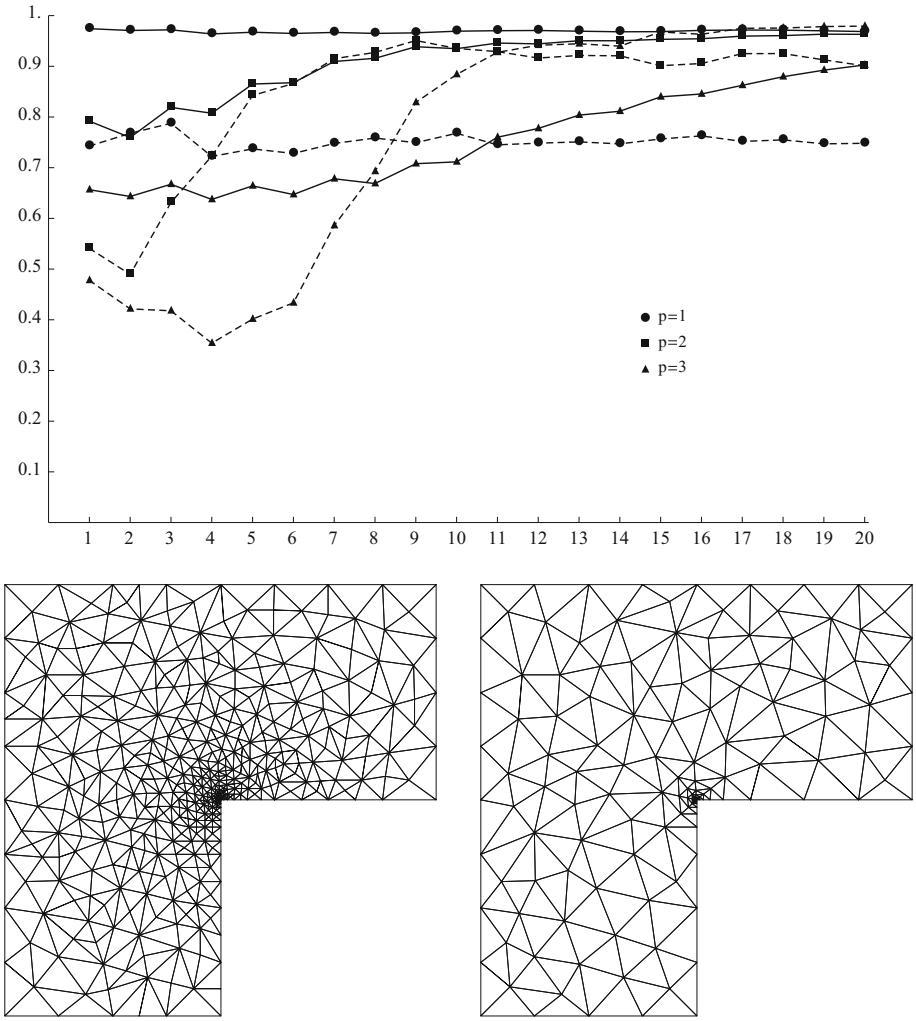

Fig. 1 Global H^1 (*solid*) and L^2 (*dashed*) effectivities for the L-shaped domain on a sequence of 20 adaptively-refined meshes. Below are h-adapted meshes when $p = 1$ (*left*) and $p = 3$ (*right*)

4.1.2 Single Boundary Layer

Let $\Omega = (0, 1) \times (0, 1)$, $\partial\Omega_D$ consist of the left and right edges, and $\partial\Omega_N$ consist of the top and bottom edges. We consider the problem

$$-\epsilon\Delta u + \frac{\partial u}{\partial x} + u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega_D, \quad \frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \partial\Omega_N.$$

Here, we take f so that the solution is

$$u = x - \frac{e^{-(1-x)/\epsilon} - e^{-1/\epsilon}}{1 - e^{-1/\epsilon}},$$

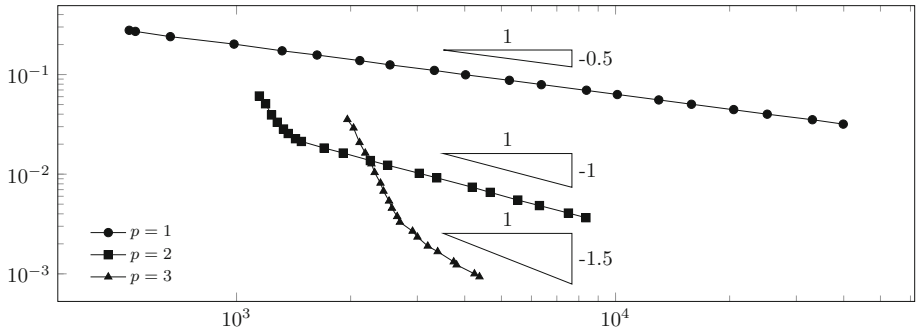

Fig. 2 Test problem in Sect. 4.1.1. Energy errors versus the total degrees of freedom on a sequence of 20 adaptively-refined meshes and reference slopes

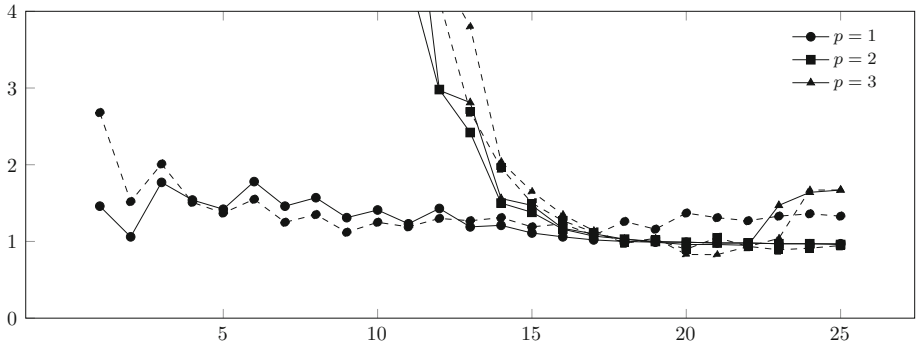

Fig. 3 Test problem in Sect. 4.1.2. Global energy effectivities (*solid*) and L^2 effectivities (*dashed*) for the single boundary layer problem on a sequence of 25 adaptivity-refined meshes

for $\epsilon = 10^{-3}$ (modest convection dominance). This is a typical model problem that is considered, for example, in [3].

For this problem the energy norm is defined by $\|v\|^2 = \epsilon|v|_1^2 + \|v\|_0^2$. We start with a uniform mesh with $h = 1/5$ and calculate the energy and L^2 errors, estimated errors, and effectivities for $p = 1, 2, 3$ on 25 refined meshes obtained by the adaptive strategy. The global effectivities with respect to the energy norm and the L^2 norm are given in Figs. 3 and 4. The plots show, that, at least asymptotically, the effectivities are near 1 in both the energy and L^2 noms. For example, the energy effectivities for refinement levels 14 through 25 fall in the range $[0.97, 1.21]$ for $p = 1$, $[0.96, 1.5]$ for $p = 2$, and $[0.85, 1.67]$ for $p = 3$. Finally we plot the energy errors against the total degrees of freedom in Fig. 5. Again, for refinement level sufficiently large, we observe, approximately, the optimal rate of convergence $|\dim V_p + \dim W_{p+d}|^{-p/2}$ for $p = 1, 2, 3$.

4.1.3 Curved Internal Layer

Letting $\Omega = (0, 1) \times (0, 1)$, $\partial\Omega_0$ consist of the left and bottom edges, and $\partial\Omega_1$ consist of the top and right edges, we consider the problem

$$-\epsilon\Delta u + (-y, x) \cdot \nabla u = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega_0, \quad u = 1 \text{ on } \partial\Omega_1. \tag{4.1}$$

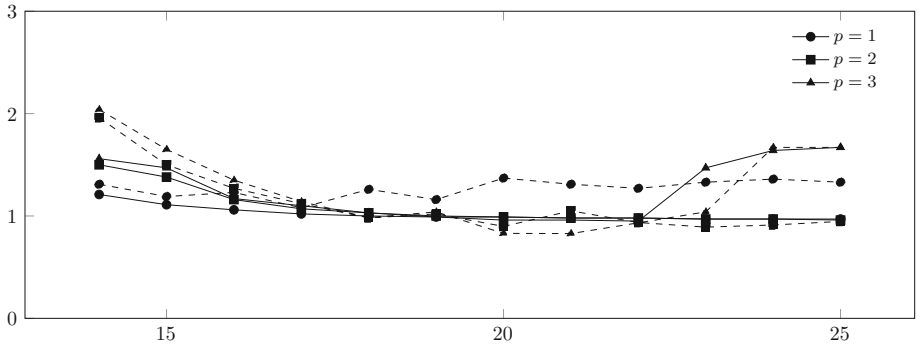

Fig. 4 The effectivities for the single boundary layer problem restricted to refinement levels 14 through 25

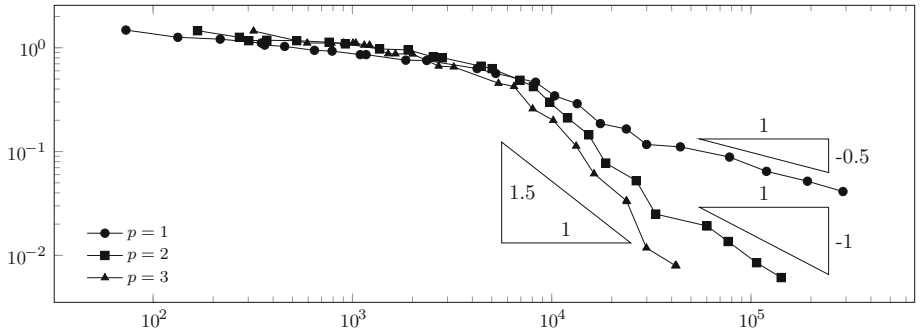

Fig. 5 Energy errors versus the total degrees of freedom for the single boundary layer problem on a sequence of 25 adaptively-refined meshes and reference slopes

Formally, in the limiting case $\epsilon = 0$, the weak solution is piecewise constant, having value 0 when $x^2+y^2 < 1$, and value 1 when $x^2+y^2 > 1$. Note that, since the Dirichlet boundary data is not $H^{1/2}(\partial\Omega)$ (it is in $H^{1/2-s}(\Omega)$ for all $s > 0$), problem (4.1) is ill-posed on $H^1(\Omega)$. Such convection-dominated problems having discontinuous Dirichlet data are often motivated by (lid-)driven cavity problems, and are not uncommon in finite element literature for convection-dominated problems (cf. [14,29,33,36]). After interpolating the boundary conditions (which requires an arbitrary choice of u at the corners of $\partial\Omega$), we obtain a discrete, continuous function on $\partial\Omega$, thus leading to a well-posed problem for each fixed mesh. However, it is expected, and numerically confirmed, that $\|\hat{u}\|_1 \to \infty$ as the mesh is refined (Fig. 6).

As ϵ decreases, an internal layer forms near the curve $x^2 + y^2 = 1$. Furthermore, due to the incompatible Dirichlet boundary conditions, there are severe singularities at the two corners (1, 0) and (0, 1), and these singularities tend to "distract" the adaptive algorithm into refining nearly exclusively around these two corners. We modify the adaptive strategy by omitting the largest 1% of the local error indicators in our marking strategy. Such a strategy of omitting some percentage of the largest local error indicators for this kind of problem has been suggested in [38].

Since the solution to (4.1) is undefined in $H^1(\Omega)$, we do not attempt convergence and effectivity estimates based on a reference solution, but instead provide qualitative results that illustrate the how the local indicators guide the refinement. We plot in Fig. 7 the resulting meshes for polynomial degrees $p = 1$ and $p = 3$ after 20 refinements and with $\epsilon = 10^{-3}$.

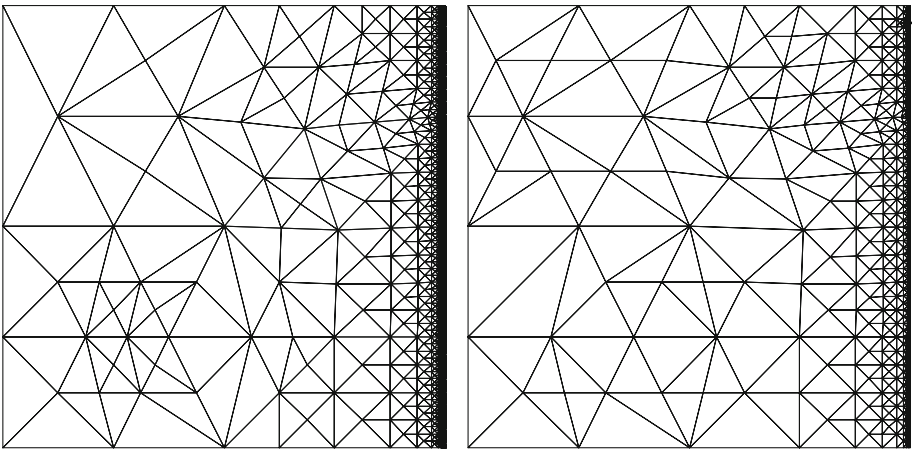

Fig. 6 h-adapted meshes for the single boundary layer problem when $p = 1$ (*left*) and $p = 3$ (*right*)

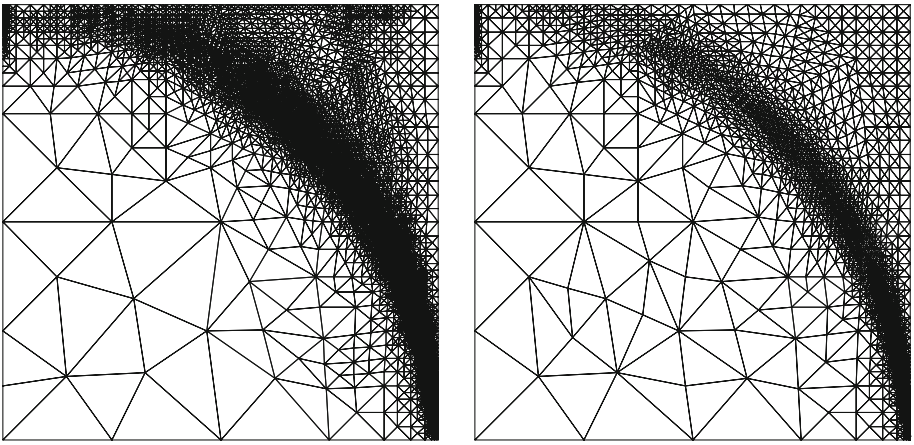

Fig. 7 h-adapted meshes for the curved internal layer problem when $p = 1$ (*left*) and $p = 3$ (*right*)

The figure clearly shows that the error estimators and adaptive strategy refine in the vicinity of the boundary layer $x^2 + y^2 = 1$ and the two corners $(1, 0)$ and $(0, 1)$.

4.2 Investigation of Properties of the Estimator Under Uniform p-Refinement

In the experiments that follow, we investigate the behavior of our estimator with respect to uniform p-refinement on fixed (adapted) meshes which may consist of quadrilaterals (or bricks), triangles, or a combination of the two. In the case of quadrilateral or brick elements, we use the full tensor-product space indexed by maximal degree in each variable, not a reduced space indexed by total degree. The choice of full tensor-product space more naturally fits with our theoretical development of the error estimator, and it provides better convergence for some of the more challenging problems below. The auxiliary space W_{p+d} for the tensor elements still consists of the interior bubbles of degree up to $p+d$ and face bubbles of degree up to $p + d - 1$ which were not already present in V_p. The problems are chosen to illustrate

the behavior of the estimator in a variety of situations in which certain problem-dependent parameters might reasonably affect performance.

In nearly all cases, we observe that the error estimates stay within a factor of two of the actual errors as p increases. The conditioning varied widely between problems due to problem parameters and the use of an integrated Legendre basis for tensor elements versus a standard Legendre basis for triangular elements, but the ratio of condition numbers (W_{p+d} over V_p) indicates that the cost of computing ε is acceptable. For example, for all choices of β in Sect. 4.2.1, the condition number ratios for rectangular elements remained $\mathcal{O}(1)$ and the condition numbers themselves remained $\mathcal{O}(10)$ for all p. For the same problem on triangular elements this ratio decreased steadily to reach $\mathcal{O}(10^{-4})$ when $p = 8$, with the condition number for W_{p+d} at $\mathcal{O}(100)$. The size of the stiffness matrix and number of non-zeros for W_{p+d} tended to drop below that for V_p at either $p = 4$ or $p = 5$ for all 2D problems—static condenstation was not used.

Because the focus of this set of experiments is to illustrate robustness of the estimator with respect to polynomial degee p, we provide the effectivity and convergence plots with respect to p, instead of the number of degrees of freedom. Because a priori knowledge of the solution is used to expertly craft a mesh, exponential convergence is expected (cf. [31] and references therein), and this is what is observed in the convergence plots. To make the connection between p and the number of degrees of freedom clearer, recall that the mesh is fixed, so the number of degrees of freedom is given by a polynomial of degree d in p. Since the convergence plots have a logarithmic scale on the error axis and a linear scale on the p axis, convergence curves that are roughly linear (or can be bounded above by a line) in these semi-log plots indicate convergence that is exponential with respect to some fractional power of the number of degrees of freedom, which is consistent with the theory.

4.2.1 Discontinuous and Anisotropic Diffusion on the Square

Letting $\Omega = (-1, 1) \times (-1, 1)$, we consider problems of the form

$$-\nabla \cdot (A\nabla u) = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \quad A = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix}, \quad \alpha = \begin{cases} 1 & x < 0 \\ \beta & x > 0 \end{cases},$$

for various choices of $\beta > 1$. Because the jump discontinuity in the diffusion matrix happens along a straight line, one does not expect singularities in u for generic f. This allows us to isolate potential effects of varying β on the effectivity of the estimator from those which might arise due to singularities in u—singular solutions are considered in the two subsequent problems. The function f is chosen so that the solution is given by

$$u = \cos(\pi y/2) \begin{cases} \left(e^{-1} - e^{x} + \frac{(e-1)(\beta+e)}{e(\beta+1)}(x+1)\right) & x < 0 \\ \beta^{-1}\left(e - e^{x} + \frac{(e-1)(\beta+e)}{e(\beta+1)}(x-1)\right) & x > 0 \end{cases}.$$

We note that $u = \cos(\pi y/2)w(x)$, where w is the solution of the 1D problem $-(\alpha w')' = e^x$ in $(-1, 1)$ with $w(-1) = w(1) = 0$, so u exhibits the typical behavior of having relatively small magnitude where β is large.

We report convergence and effectivity for $\beta = 10, 100, 1000$ on two different meshes—the first consisting of two rectangles obtained by dividing the domain along the line $x = 0$, and the second consisting of four triangles obtained by dividing the two rectangles along their diagonals. Convergence and effectivity plots, for the energy norm, are provided in Fig. 8 for both types of elements. To save space, each of the four plots contain graphs for all three

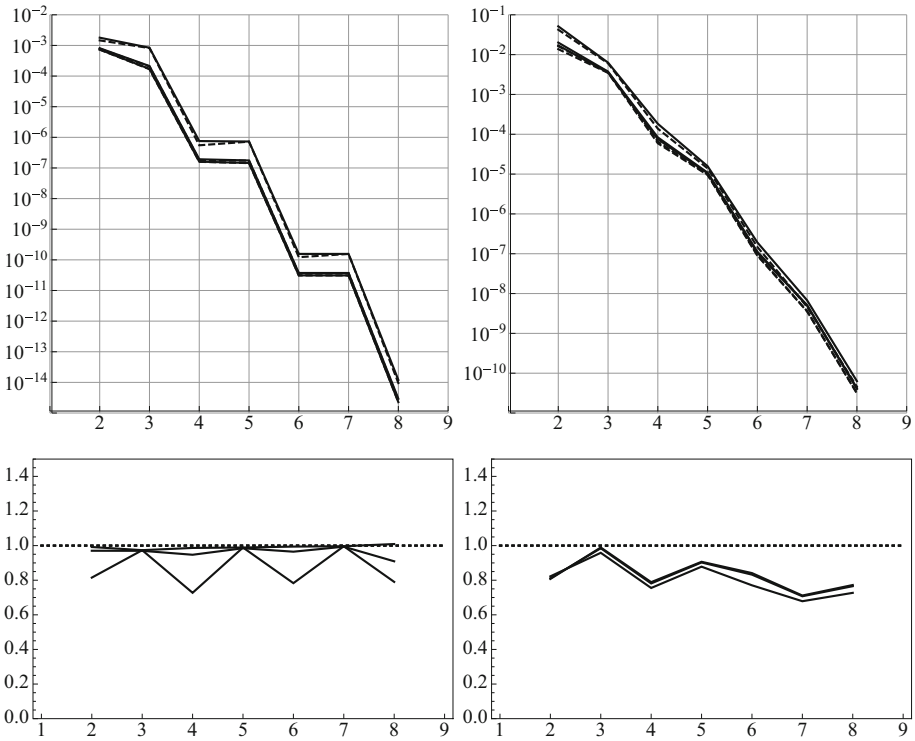

Fig. 8 Convergence of the error (*solid*) and error estimates (*dashed*) with respect to p for the discontinuous and anisotropic diffusion problem, for both rectangular elements (*left*) and triangular elements (*right*) and $\beta = 10, 100, 1000$. Global effectivities (*solid*) for both types of elements are given below their respective convergence plots

values of β. The convergence and effectivity behavior for $\beta = 100$ and $\beta = 1000$ is nearly identical, so their graphs are almost indistinguishable—the case $\beta = 10$ is more clearly distinguishable from the other for both types of elements. The effectivities stay within the range $[0.7, 1)$ in all cases.

4.2.2 Slit Disk

Let Ω be the unit disk with a slit along the positive x-axis, with Γ_1 consisting of the boundary of the disk ($r = 1$) and the top of the slit ($\theta = 0^+$, $0 \leq r \leq 1$), and Γ_2 consisting of the bottom of the slit ($\theta = 2\pi^-$, $0 < r < 1$); see Fig. 9. We consider the problem

$$-\Delta u = f = (4 - \sigma^2)\sin(\sigma\theta) \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma_1, \quad \text{condition on } \Gamma_2,$$

for two choices of σ. If $u = 0$ on Γ_2, we take $\sigma = 1/2$ and refer to the problem as the *Dirichlet–Dirichlet* slit; and if $\partial u/\partial n = 0$ on Γ_2 we take $\sigma = 1/4$ call this the *Dirichlet-Neumann* slit. In both cases, the solution is given by

$$u = (r^\sigma - r^2)\sin(\sigma\theta),$$

and it exhibits the typical singularities present for generic f. In Fig. 9 we see the mesh and a close-up of the central portion of the mesh. It is clear from these images that the mesh

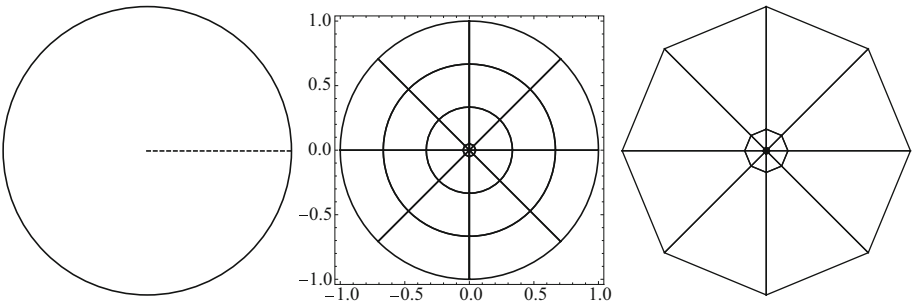

Fig. 9 The slit disk, together with its mesh and a close-up of the central portion

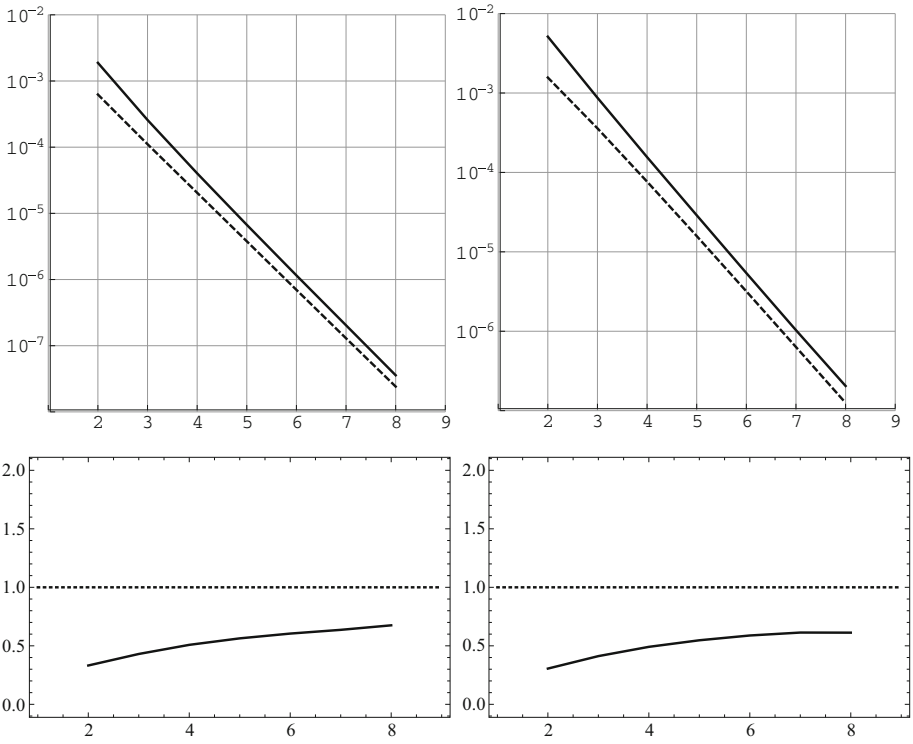

Fig. 10 Convergence of the error (*solid*) and error estimates (*dashed*) with respect to p for both the Dirichlet–Dirichlet case (*left*) and the Dirichlet–Neumann case (*right*). Global effectivities (*solid*) for both problems are given below their respective convergence plots

includes both curved and straight quadrilaterals, but it also includes triangles touching the origin. Despite the difference in singularity strength for the two types of boundary conditions, the same mesh is used in both cases. At every refinement step the elements touching the singularity are refined using an edge split ratio $\alpha = 3/20$. For p-convergence tests the hierarchic refinement process is repeated 20 times.

Convergence and effectivity plots are given in Fig. 10 for both problems, with respect to polynomial degree p. We emphasize that the effectivities in both cases do not deteriorate

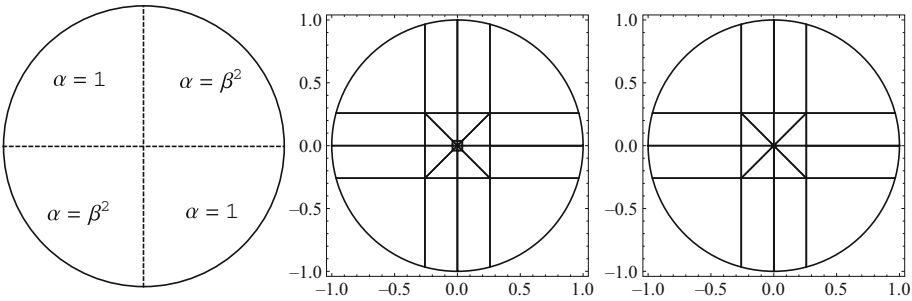

Fig. 11 The Kellogg problem, together with its mesh for $\beta = 5$ (*center*) and $\beta = 10$ (*right*)

with p, and indicate that the error estimate is generally within a factor of two of the actual error.

4.2.3 Kellogg Problem

Let Ω be the unit disk and $\beta > 1$, and define $\sigma = \arctan(\beta^{-1})/(\pi/4)$. We consider the problem

$$-\nabla \cdot (\alpha \nabla u) = f = (4 - \sigma^2)\, \alpha\, g \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega,$$

where

$$g(\theta) = \begin{cases} -\cos(\sigma(\pi/4 - \theta))/\beta, & \theta \in [0, \pi/2) \\ -\sin(\sigma(3\pi/4 - \theta)), & \theta \in [\pi/2, \pi) \\ \cos(\sigma(5\pi/4 - \theta))/\beta, & \theta \in [\pi, 3\pi/2) \\ \sin(\sigma(7\pi/4 - \theta)), & \theta \in [3\pi/2, 2\pi) \end{cases}, \quad \alpha(\theta) = \begin{cases} \beta^2, & \theta \in [0, \pi/2) \cup [\pi, 3\pi/2) \\ 1, & \theta \in [\pi/2, \pi) \cup [3\pi/2, 2\pi) \end{cases},$$

and we require that both u and $\alpha\, \partial u/\partial n$ are continuous across the interfaces between the four quadrants (see Fig. 11). We may naturally think of α and g as functions on $\mathbb{R}$ via 2π-periodic extension. The solution is given by

$$u = (r^\sigma - r^2) g(\theta),$$

and it exhibits the typical leading singularity present for generic f. By increasing β, we can make $\sigma > 0$ as small as we like, thereby generating an increasingly strong singularity at the origin. For our experiments we consider the cases $\beta = 5$ and $\beta = 10$, for which the solution has leading singularities $r^{0.251332}$ and $r^{0.126902}$, respectively. As in the slit problem above, the meshes have a mix of curved and straight triangles and quadrilaterals, as seen in Fig. 11, and same refinement strategies are employed. To accomodate the stronger singularity in $\beta = 10$, more aggressive ratio $\alpha = 1/20$ is used. The resulting problem sizes are summarised in Table 3. Convergence and effectivity plots are given in Fig. 12 for both $\beta = 5$ and $\beta = 10$, with respect to polynomial degree p. As before, we see that the effectivities in both cases do not deteriorate with p, and indicate that the error estimate is generally within a factor of two of the actual error.

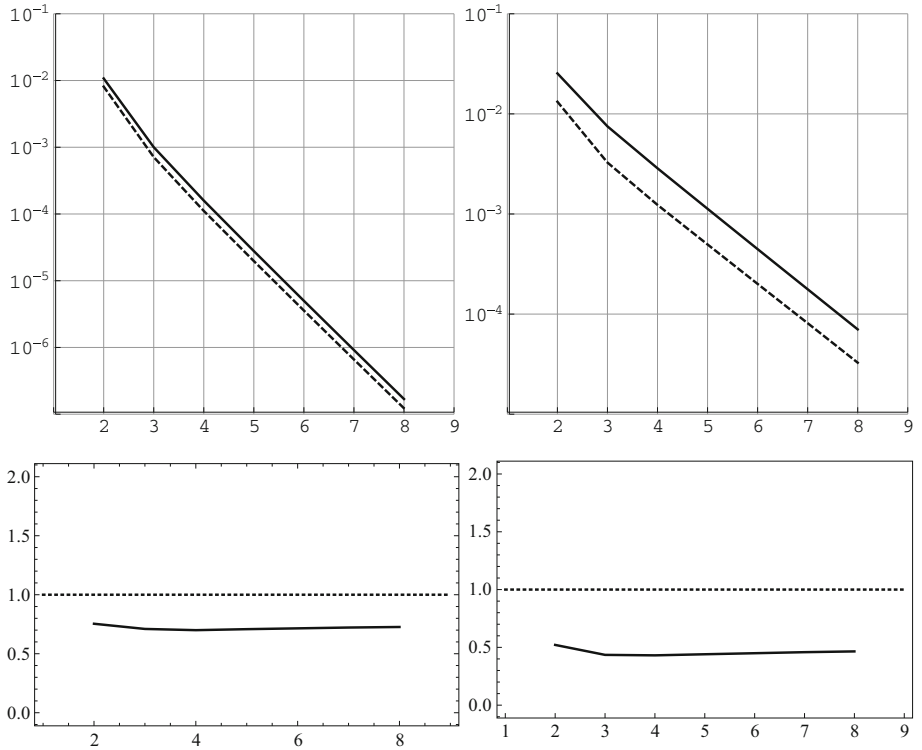

Fig. 12 Convergence of the error (*solid*) and error estimates (*dashed*) with respect to p for the Kellogg problem with $\beta = 5$ (*left*) and $\beta = 10$ (*right*). Global effectivities (*solid*) in the energy norm for both problems are given below their respective convergence plots

4.2.4 Boundary Layers

Letting Ω be either the unit square or the unit cube, we consider the problem

$$-\epsilon \Delta u + \frac{\partial u}{\partial x} + 2u/\alpha = 1 \text{ in } \Omega,$$

with homogenous Dirichlet conditions at $x = 0$ and $x = 1$, and homogeneous Neumann conditions on the rest of the boundary. The solution is given by

$$u = \frac{\alpha}{2}\left(1 + \left(\frac{e^{r^-} - 1}{e^{r^+} - e^{r^-}}\right)e^{r^+ x} - \left(\frac{e^{r^+} - 1}{e^{r^+} - e^{r^-}}\right)e^{r^- x}\right), \quad r^{\pm} = \frac{1 \pm \sqrt{1 + 8\epsilon/\alpha}}{2\epsilon}.$$

Such solutions exhibit boundary layers near both $x = 0$ and $x = 1$ when $0 < \varepsilon \ll 1$ and $0 < \alpha \ll 1$. The quadrilateral meshes for $\epsilon = \alpha = 10^{-1}$ and $\epsilon = \alpha = 10^{-4}$ are given in Fig. 13.

The convergence and effectivity plots for these problems are given in Fig. 14. These are given in the energy-norm,

$$\|v\|^2 = \epsilon |v|_1^2 + 2\|v\|_0^2/\alpha,$$

derived from the symmetric part of the associated bilinear form.

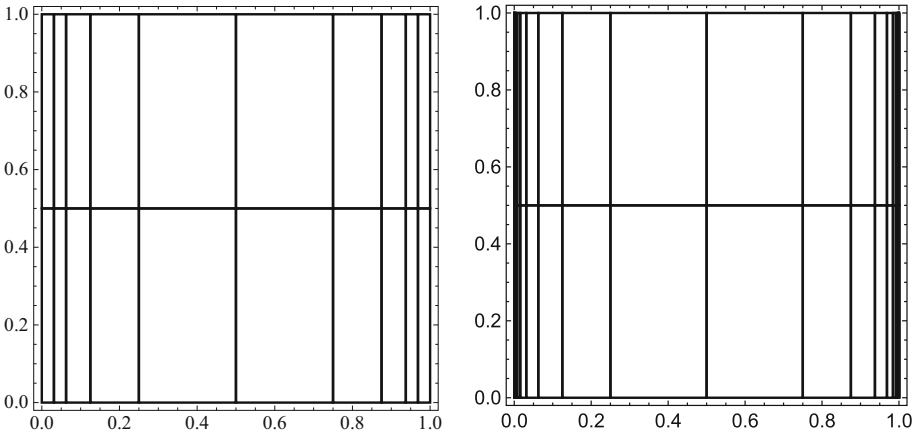

Fig. 13 Rectangular meshes for the cases $\epsilon = \alpha = 10^{-1}$ and $\epsilon = \alpha = 10^{-4}$ of the Boundary Layer problem

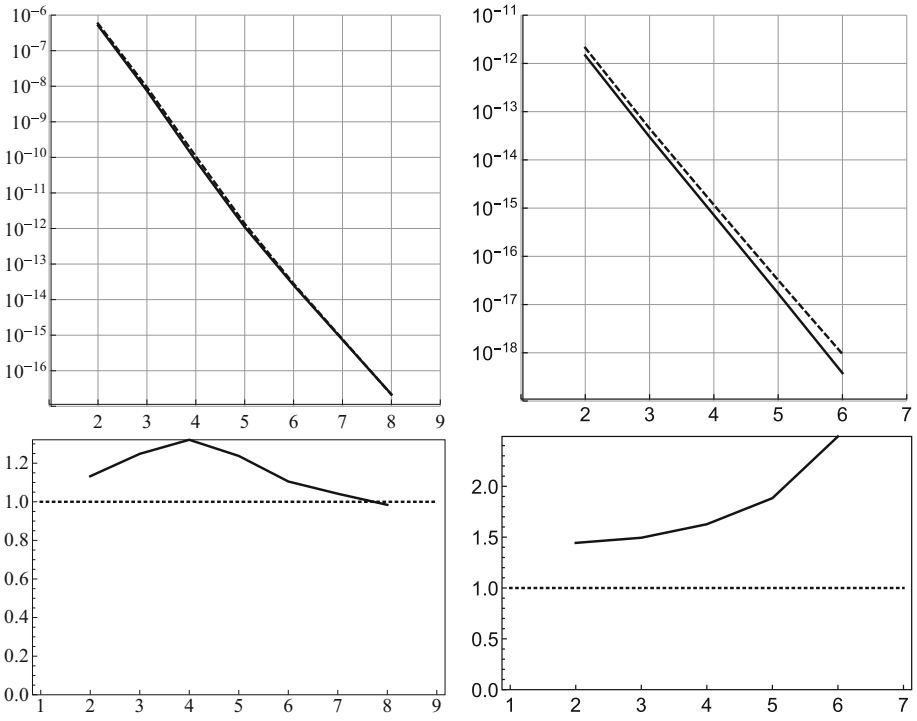

Fig. 14 Convergence of the error (*solid*) and error estimates (*dashed*) with respect to p for the Boundary Layer problem with $\epsilon = \alpha = 10^{-1}$ (*left*) and $\epsilon = \alpha = 10^{-4}$ (right). Global effectivities (*solid*) for both problems are given below their respective convergence plots

Since the errors are near machine-precision for $p \geq 7$, the reported effectivities may not be as accurate in that range. For comparison, we also provide convergence and effectivity plots in the H^1 norm for the case $\epsilon = \alpha = 10^{-4}$ in Fig. 15.

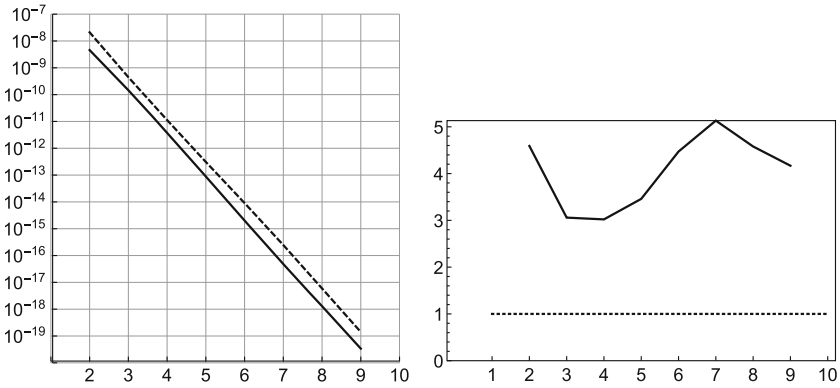

Fig. 15 Convergence of the error (*solid*) and error estimates (*dashed*) with respect to p for the Boundary Layer problem with $\epsilon = \alpha = 10^{-4}$. Global effectivities (*solid*) are given at right. In this case, errors and error estimates are measured in the H^1-norm

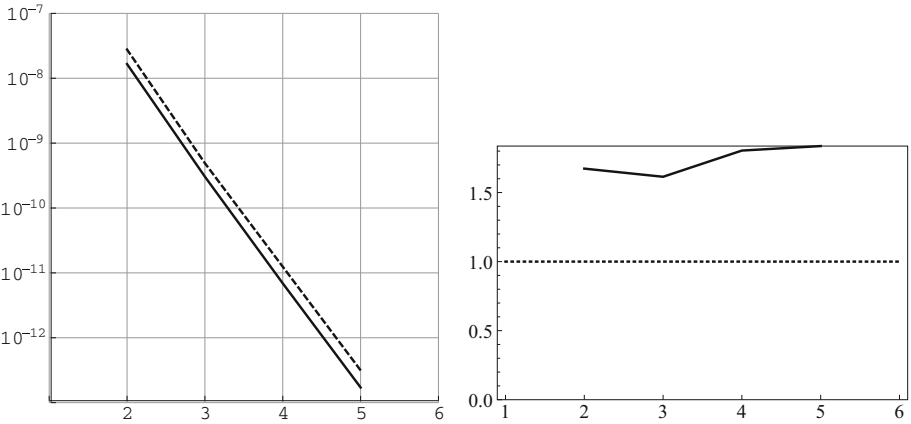

Fig. 16 At left, convergence of the error (*solid*) and error estimates (*dashed*) with respect to p for the 3D Boundary Layer problem with $\epsilon = \alpha = 10^{-2}$. Global effectivities (*solid*) are given at right

Finally, we consider the case $\epsilon = \alpha = 10^{-2}$ in 3D with hexahedral bricks with an appropriate x-grading, and whose yz-aspect ratio is 1 for each brick. The convergence and effectivity information are given in Fig. 16.

4.3 Investigation of Properties of the Estimator Under *hp*-Refinement

We briefly revisit two of the examples from Sect. 4.2, using an a priori hp-refinement strategy instead of the fixed a priori h-refinement strategy with uniformly-increasing p used in that section. In both examples, the elementwise polynomial degrees (the p-vector) are set as the elemental graph distance from the singularity (cf. [31]) in each h-refined mesh. The resulting problem sizes, including comparisons with their counterparts for the p method, are summarized in Tables 2 and 3.

Table 2 Summary of the problem sizes for the Slit Disk problem

Level:	1	2	3	4	5	6	7	8
p		753	1651	2897	4491	6433	8723	11,361
hp	99	223	418	700	1085	1589	2228	3018

Table 3 Summary of the problem sizes for the Kellogg problem

Level:	1	2	3	4	5	6	7	8
p		709	1585	2809	4381	6301	8569	11,185
hp	133	277	497	809	1229	1773	2457	3297

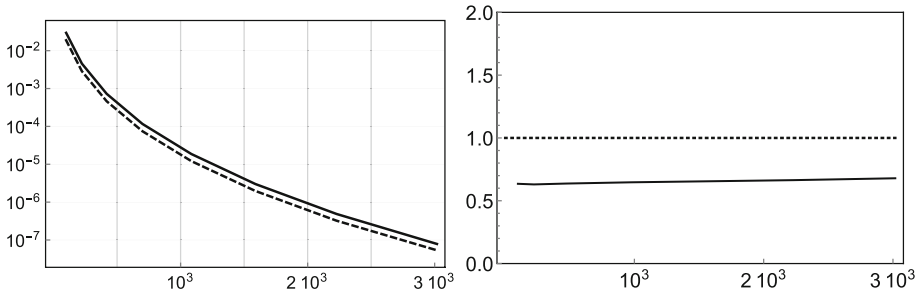

Fig. 17 hp-Convergence of the error (*solid*) and error estimates (*dashed*) with respect to the number of d.o.f for the Dirichlet–Dirichlet case (*left*) and global effectivity (*solid*) in the energy norm (*right*)

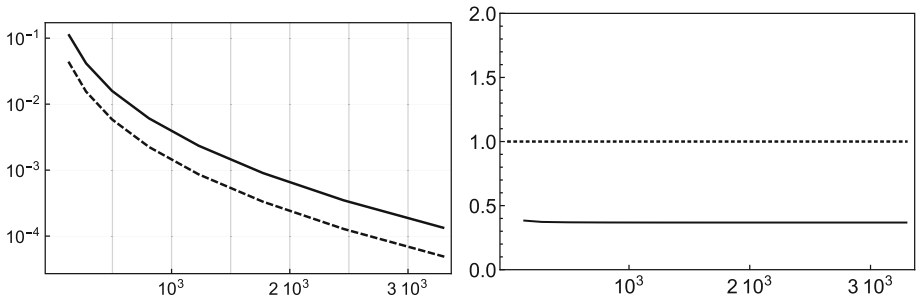

Fig. 18 hp-Convergence of the error (*solid*) and error estimates (*dashed*) with respect to the number of d.o.f for the Kellogg problem with $\beta = 5$ (*left*) and global effectivity (*solid*) in the energy norm (*right*)

4.3.1 Slit Disk, hp-Version

We revisit the Slit Disk problem of Sect. 4.2.2, focusing on the version having Dirichlet boundary conditions on all edges (DD). The h-refinement in this case is identical to that used in Sect. 4.2.2, and p increases away from the singularity as discussed above. In Fig. 17, we see both the convergence of the error and error estimate with respect to the number of degrees of freedom, as well as the corresponding effectivities. The convergence is exponential, as expected, and the effectivities remain uniformly good throughout the refinement procedure.

4.3.2 Kellogg Problem, hp-Version

We consider an hp-version of the Kellogg problem (Sect. 4.2.3), with $\beta = 5$. The h-refinement in this case is the same as that described in that section. As expected, the error and error estimates converge exponentially with respect to the number of degrees of freedom; see Fig. 18. Again, the effectivities are essentially uniform during the procedure, though they are slightly worse than their counterparts in the p-refinement case.

Acknowledgements This work was supported in part by the National Science Foundation through grants DMS-1417980 (Neilan) and DMS-1414365 (Ovall).

References

1. Adjerid, S., Aiffa, M., Flaherty, J.E.: Hierarchical finite element bases for triangular and tetrahedral elements. Comput. Methods Appl. Mech. Eng. **190**(22–23), 2925–2941 (2001)
2. Ainsworth, M., Oden, J.T.: A posteriori error estimation in finite element analysis. In: *Pure and Applied Mathematics*. Wiley, New York (2000)
3. Araya, R., Poza, A.H., Stephan, E.P.: A hierarchical a posteriori error estimate for an advection–diffusion–reaction problem. Math. Models Methods Appl. Sci. **15**(7), 1119–1139 (2005)
4. Arnold, D.N.: Spaces of finite element differential forms. In: Brezzi, F., Colli-Franzone, P., Gianazza, U.P., Gilardi, G. (eds.) *Analysis and Numerics of Partial Differential Equations*, volume 4 of Springer INdAM Series, pp. 117–140. Springer, Milan (2013)
5. Bank, R.E.: Hierarchical bases and the finite element method. In *Acta numerica, 1996*, volume 5 of Acta numerica, pp. 1–43. Cambridge University Press, Cambridge (1996)
6. Bank, R.E., Grubišić, L., Ovall, J.S.: A framework for robust eigenvalue and eigenvector error estimation and Ritz value convergence enhancement. Appl. Numer. Math. **66**, 1–29 (2013)
7. Bank, R.E., Nguyen, H.: *hp* adaptive finite elements based on derivative recovery and superconvergence. Comput. Vis. Sci. **14**(6), 287–299 (2011)
8. Bank, R.E., Smith, R.K.: A posteriori error estimates based on hierarchical bases. SIAM J. Numer. Anal. **30**(4), 921–935 (1993)
9. Bank, R.E., Xu, J., Zheng, B.: Superconvergent derivative recovery for Lagrange triangular elements of degree p on unstructured grids. SIAM J. Numer. Anal. **45**(5), 2032–2046 (2007) **(electronic)**
10. Bernardi, C., Verfürth, R.: Adaptive finite element methods for elliptic equations with non-smooth coefficients. Numer. Math. **85**(4), 579–608 (2000)
11. Beuchler, S., Schöberl, J.: New shape functions for triangular p-FEM using integrated Jacobi polynomials. Numer. Math. **103**(3), 339–366 (2006)
12. Bornemann, F.A., Erdmann, B., Kornhuber, R.: A posteriori error estimates for elliptic problems in two and three space dimensions. SIAM J. Numer. Anal. **33**(3), 1188–1204 (1996)
13. Braess, D., Pillwein, V., Schöberl, J.: Equilibrated residual error estimates are p-robust. Comput. Methods Appl. Mech. Eng. 198(1314):1189–1197 (2009). *HOFEM07 International Workshop on High-Order Finite Element Methods* (2007)
14. Cai, Z., Wang, Y.: Pseudostress-velocity formulation for incompressible Navier–Stokes equations. Int. J. Numer. Methods Fluids **63**, 341–356 (2010)
15. Carnevali, P., Morris, R.B., Tsuji, Y., Taylor, G.: New basis functions and computational procedures for p-version finite element analysis. Int. J. Numer. Methods Eng. **36**(22), 3759–3779 (1993)
16. Craig, A.W., Zhu, J.Z., Zienkiewicz, O.C.: A posteriori error estimation, adaptive mesh refinement and multigrid methods using hierarchical finite element bases. In: *The Mathematics of Finite Elements and Applications. V (Uxbridge, 1984)*, pp. 587–594. Academic Press, London (1985)
17. Deuflhard, P., Leinen, P., Yserentant, H.: Concepts of an adaptive hierarchical finite element code. IMPACT Comput. Sci. Eng. **1**(1), 3–35 (1989)
18. Dörfler, W., Nochetto, R.H.: Small data oscillation implies the saturation assumption. Numer. Math. **91**(1), 1–12 (2002)
19. Eijkhout, V., Vassilevski, P.: The role of the strengthened Cauchy–Buniakowskiĭ–Schwarz inequality in multilevel methods. SIAM Rev. **33**(3), 405–419 (1991)
20. Ern, A., Guermond, J.-L.: *Theory and Practice of Finite Elements*, volume 159 of Applied Mathematical Sciences. Springer, New York (2004)

21. Ern, A., Vohralík, M.: Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. SIAM J. Numer. Anal. **53**(2), 1058–1081 (2015)
22. Holst, M., Ovall, J.S., Szypowski, R.: An efficient, reliable and robust error estimator for elliptic problems in R^3. Appl. Numer. Math. **61**(5), 675–695 (2011)
23. Huang, W., Kamenski, L., Lang, J.: A new anisotropic mesh adaptation method based upon hierarchical a posteriori error estimates. J. Comput. Phys. **229**(6), 2179–2198 (2010)
24. Kamenski, L.: A study on using hierarchical basis error estimates in anisotropic mesh adaptation for the finite element method. Eng. Comput. **28**(4), 451–460 (2012)
25. Kreuzer, C., Siebert, K.G.: Decay rates of adaptive finite elements with Dörfler marking. Numer. Math. **117**(2), 679–716 (2011)
26. Li, H., Ovall, J.: A posteriori error estimation of hierarchical type for the Schrödinger operator with inverse square potential. Numer. Math. **128**(4), 707–740 (2014)
27. Melenk, J.M., Wohlmuth, B.I.: On residual-based a posteriori error estimation in *hp*-FEM. Adv. Comput. Math. **15**(1–4), 311–331 (2002). 2001. A posteriori error estimation and adaptive computational methods
28. Ovall, J.S.: Function, gradient, and Hessian recovery using quadratic edge-bump functions. SIAM J. Numer. Anal. **45**(3), 1064–1080 (2007) **(electronic)**
29. Papastavrou, A., Verfürth, R.: A posteriori error estimators for stationary convection–diffusion problems: a computational comparison. Comput. Methods Appl. Mech. Eng. **189**, 449–462 (2000)
30. Petzoldt, M.: A posteriori error estimators for elliptic equations with discontinuous coefficients. Adv. Comput. Math. **16**(1), 47–75 (2002)
31. Schwab, C.: *p*- and *hp*-finite element methods. *Numerical Mathematics and Scientific Computation*. The Clarendon Press, Oxford University Press, New York (1998). Theory and applications in solid and fluid mechanics
32. Scott, L.R., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. **54**(190), 483–493 (1990)
33. Skalický, T., Roos, H.-G.: Anisotropic mesh refinement for problems with internal and boundary layers. Int. J. Numer. Methods Eng. **46**, 1933–1953 (1999)
34. Szabó, B., Babuška, I.: Finite Element Analysis. Wiley, New York (1991)
35. Verfürth, R.: A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Wiley, New York (1996)
36. Verfürth, R.: A posteriori error estimators for convection–diffusion equations. Numer. Math. **80**(4), 641–663 (1998)
37. Verfürth, R.: Robust a posteriori error estimates for stationary convection–diffusion equations. SIAM J. Numer. Anal. **43**(4), 1766–1782 (2005). (electronic)
38. Verfürth, R.: A posteriori error estimation techniques for finite element methods. In: *Numerical Mathematics and Scientific Computation*. Oxford University Press, Oxford (2013)
39. Zienkiewicz, O.C., Kelly, D.W., Gago, J., Babuška, I.: Hierarchical finite element approaches, error estimates and adaptive refinement. In: *The Mathematics of Finite Elements and Applications IV (Uxbridge, 1981)*, pp. 313–346. Academic Press, London (1982)