Primer of Adaptive Finite Element Methods

Ricardo H. Nochetto and Andreas Veeser

Abstract Adaptive finite element methods (AFEM) are a fundamental numerical instrument in science and engineering to approximate partial differential equations. In the 1980s and 1990s a great deal of effort was devoted to the design of a posteriori error estimators, following the pioneering work of Babuška. These are computable quantities, depending on the discrete solution(s) and data, that can be used to assess the approximation quality and improve it adaptively. Despite their practical success, adaptive processes have been shown to converge, and to exhibit optimal cardinality, only recently for dimension $d > 1$ and for linear elliptic PDE. These series of lectures presents an up-to-date discussion of AFEM encompassing the derivation of upper and lower a posteriori error bounds for residual-type estimators, including a critical look at the role of oscillation, the design of AFEM and its basic properties, as well as a complete discussion of convergence, contraction property and quasi-optimal cardinality of AFEM.

Ricardo H. Nochetto
Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742, e-mail: rhn@math.umd.edu. Partially supported by NSF grant DMS-0807811.

Andreas Veeser
Dipartimento di Matematica, Università degli Studi di Milano, Via C. Saldini 50, I-20133 Milano, Italy, e-mail: andreas.veeser@unimi.it. Partially supported by Italian PRIN 2008 “Analisi e sviluppo di metodi numerici avanzati per EDP”.

1 Piecewise Polynomial Approximation

We start with a discussion of piecewise polynomial approximation in W_p^k Sobolev spaces and graded meshes in any dimension d. We first compare pointwise approximation over uniform and graded meshes for $d = 1$ in §1.1, which reveals the advantages of the latter over the former and sets the tone for the rest of the paper. We continue with the concept of Sobolev number in §1.2.

We explore the geometric aspects of mesh refinement for conforming meshes in §1.3 and nonconforming meshes in §1.7, but postpone a full discussion until §6. We include a statement about complexity of the refinement procedure, which turns out to be instrumental later.

We briefly discuss the construction of finite element spaces in §1.4, along with polynomial interpolation of functions in Sobolev spaces in §1.5. This provides local estimates adequate for comparison of quasi-uniform and graded meshes for $d > 1$. We exploit them in developing the so-called error equidistribution principle and the construction of suitably graded meshes via thresholding in §1.6. We conclude that graded meshes can deliver optimal interpolation rates for certain classes of singular functions, and thus supersede quasi-uniform refinement.

1.1 Classical vs Adaptive Pointwise Approximation

We start with a simple motivation in 1d for the use of adaptive procedures, due to DeVore [22]. Given $\Omega = (0,1)$, a partition $\mathscr{T}_N = \{x_i\}_{n=0}^N$ of Ω

$$0 = x_0 < x_1 < \cdots < x_n < \cdots < x_N = 1$$

and a continuous function $u : \Omega \to \mathbb{R}$, we consider the problem of *interpolating* u by a *piecewise constant* function U_N over $\mathscr{T}_N$. To quantify the difference between u and U_N we resort to the *maximum norm* and study two cases depending on the regularity of u.

Case 1: W_∞^1-Regularity. Suppose that u is Lipschitz in $[0,1]$. We consider the approximation

$$U_N(x) := u(x_{n-1}) \qquad \text{for all } x_{n-1} \le x < x_n.$$

Since

$$|u(x) - U_N(x)| = |u(x) - u(x_{n-1})| = \left| \int_{x_{n-1}}^{x} u'(t)dt \right| \le h_n \|u'\|_{L^\infty(x_{n-1},x_n)}$$

we conclude that

$$\|u - U_N\|_{L^\infty(\Omega)} \le \frac{1}{N} \|u'\|_{L^\infty(\Omega)}, \tag{1}$$

provided the local mesh-size h_n is about constant (*quasi-uniform* mesh), and so proportional to N^{-1} (the reciprocal of the number of degrees of freedom). Note that the same integrability is used on both sides of (1). A natural question arises: *Is it possible to achieve the same asymptotic decay rate N^{-1} with weaker regularity demands?*

Case 2: W_1^1-Regularity. To answer this question, we suppose $\|u'\|_{L^1(\Omega)} = 1$ and consider the non-decreasing function

$$\phi(x) := \int_0^x |u'(t)|dt$$

which satisfies $\phi(0) = 0$ and $\phi(1) = 1$. Let $\mathscr{T}_N = \{x_i\}_{n=0}^N$ be the partition given by

$$\int_{x_{n-1}}^{x_n} |u'(t)|dt = \phi(x_n) - \phi(x_{n-1}) = \frac{1}{N}.$$

Then, for $x \in [x_{n-1}, x_n]$,

$$|u(x) - u(x_{n-1})| = \Big| \int_{x_{n-1}}^x u'(t)dt \Big| \le \int_{x_{n-1}}^x |u'(t)|dt \le \int_{x_{n-1}}^{x_n} |u'(t)|dt = \frac{1}{N},$$

whence

$$\|u - U_N\|_{L^\infty(\Omega)} \le \frac{1}{N} \|u'\|_{L^1(\Omega)}. \tag{2}$$

We thus conclude that we could achieve the same rate of convergence N^{-1} for rougher functions with just $\|u'\|_{L^1(\Omega)} < \infty$. The following comments are in order for Case 2.

Remark 1 (Equidistribution). The optimal mesh $\mathscr{T}_N$ *equidistributes* the max-error. This mesh is graded instead of uniform but, in contrast to a uniform mesh, such a partition may not be adequate for another function with the same basic regularity as u. It is instructive to consider the singular function $u(x) = x^\gamma$ with $\gamma = 0.1$ and error tolerance 10^{-2} to quantify the above computations: if N_1 and N_2 are the number of degrees of freedom with uniform and graded partitions, we obtain $N_1/N_2 = 10^{18}$.

Remark 2 (Nonlinear Approximation). The regularity of u in (2) is measured in $W_1^1(\Omega)$ instead of $W_\infty^1(\Omega)$ and, consequently, the fractional γ regularity measured in $L^\infty(\Omega)$ increases to one full derivative when expressed in $L^1(\Omega)$. This exchange of integrability between left and right-hand side of (2), and gain of differentiability, is at the heart of the matter and the very reason why suitably graded meshes achieve optimal asymptotic error decay for singular functions. By those we mean functions which are not in the usual linear Sobolev scale, say $W_\infty^1(\Omega)$ in this example, but rather in a nonlinear scale [22]. We will get back to this issue in §7.

1.2 The Sobolev Number: Scaling and Embedding

In order to make Remark 2 more precise, we introduce the Sobolev number. Let $\Omega \subset \mathbb{R}^d$ with $d > 1$ be a Lipschitz and bounded domain, and let $k \in \mathbb{N}, 1 \leq p \leq \infty$. The Sobolev space $W_p^k(\Omega)$ is defined by

$$W_p^k(\Omega) := \{v : \Omega \to \mathbb{R} |\ D^\alpha v \in L^p(\Omega)\ \ \forall |\alpha| \leq k\}.$$

If $p = 2$ we set $H^k(\Omega) = W_2^k(\Omega)$ and note that this is a Hilbert space. The *Sobolev number* of $W_p^k(\Omega)$ is given by

$$\mathrm{sob}(W_p^k) := k - \frac{d}{p}. \tag{3}$$

This number governs the scaling properties of the semi-norm

$$|v|_{W_p^k(\Omega)} := \Big(\sum_{|\alpha|=k} \|D^\alpha v\|_{L^p(\Omega)}^p \Big)^{1/p},$$

because rescaling variables $\hat{x} = \frac{1}{h} x$ for all $x \in \Omega$, transforms Ω into $\hat{\Omega}$ and v into $\hat{v}$, while the corresponding norms scale as

$$|\hat{v}|_{W_p^k(\hat{\Omega})} = h^{\mathrm{sob}(W_p^k)} |v|_{W_p^k(\Omega)}.$$

In addition, we have the following *compact embedding*: if $m > k$ and $\mathrm{sob}(W_q^m) > \mathrm{sob}(W_p^k)$, then

$$W_q^m(\Omega) \subset W_p^k(\Omega).$$

We say that two Sobolev spaces are in the same nonlinear Sobolev scale if they have the same Sobolev number. We note that for compactness the space $W_q^m(\Omega)$ must be above the Sobolev scale of $W_p^k(\Omega)$. A relevant example for $d = 2$ are the pair $H^1(\Omega)$ and $L^\infty(\Omega)$ which have the same Sobolev number, in fact $\mathrm{sob}(H^1) = \mathrm{sob}(L^\infty) = 0$, but the former is not even contained in the latter: in fact

$$v(x) = \log\log \frac{|x|}{2} \in H^1(\Omega) \backslash L^\infty(\Omega)$$

in the unit ball. This is a source of difficulties for polynomial interpolation theory and the need for quasi-interpolation operators. This is discussed in §1.5.

We conclude with a comment about Remark 2. We see that $d = 1$ and $\mathrm{sob}(W_1^1) = \mathrm{sob}(L^\infty) = 0$ but $W_1^1(\Omega)$ is compactly embedded in $L^\infty(\Omega)$ in this case. This shows that these two spaces are in the same nonlinear Sobolev scale and that the above inequality between Sobolev numbers for a compact embedding is only sufficient.

1.3 Conforming Meshes: The Bisection Method

In order to approximate functions in $W_p^k(\Omega)$ by piecewise polynomials, we decompose Ω into simplices. We briefly discuss the *bisection* method, the most elegant and successful technique for subdividing Ω in any dimension into a conforming mesh. We also discuss briefly nonconforming meshes in §1.7. We present complete proofs, especially of the complexity of bisection, later in §6.

We focus on $d = 2$ and follow Binev, Dahmen, and DeVore [7], but the results carry over to any dimension $d > 1$ (see Stevenson [53]). We refer to Nochetto, Siebert, and Veeser [45] for a rather complete discussion for $d > 1$.

Let $\mathscr{T}$ denote a *mesh* (triangulation or grid) made of simplices T, and let $\mathscr{T}$ be *conforming* (edge-to-edge). Each element is labeled, namely it has an edge $E(T)$ assigned for refinement (and an opposite vertex $v(T)$ for $d = 2$); see Figure 1.

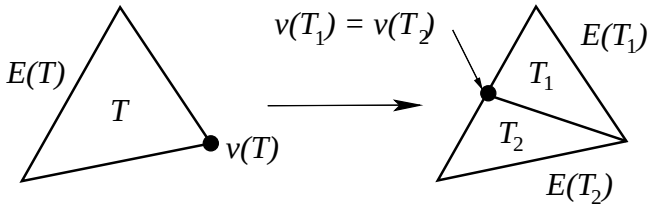


Fig. 1 Triangle $T \in \mathscr{T}$ with vertex $v(T)$ and opposite refinement edge $E(T)$. The bisection rule for $d = 2$ consists of connecting $v(T)$ with the midpoint of $E(T)$, thereby giving rise to children T_1, T_2 with common vertex $v(T_1) = v(T_2)$, the newly created vertex, and opposite refinement edges $E(T_1), E(T_2)$.

The bisection method consists of a suitable *labeling* of the initial mesh $\mathscr{T}_0$ and a rule to assign the refinement edge to the two children. For $d = 2$ we consider the *newest vertex bisection* as depicted in Figure 1. For $d > 2$ the situation is more complicated and one needs the concepts of type and vertex order [45, 53].

Bisection creates a *unique* master forest $\mathbb{F}$ of binary trees with infinite depth, where each node is a simplex (triangle in 2d), its two successors are the two children created by bisection, and the roots of the binary trees are the elements of the initial conforming partition $\mathscr{T}_0$. It is important to realize that, no matter how an element arises in the subdivision process, its associated newest vertex is unique and only depends on the labeling of $\mathscr{T}_0$: so $v(T)$ and $E(T)$ are independent of the order of the subdivision process for all $T \in \mathbb{F}$; see Lemma 16 in §6. Therefore, $\mathbb{F}$ is unique.

A finite subset $\mathscr{F} \subset \mathbb{F}$ is called a *forest* if $\mathscr{T}_0 \subset \mathscr{F}$ and the nodes of $\mathscr{F}$ satisfy

- all nodes of $\mathscr{F} \setminus \mathscr{T}_0$ have a predecessor;
- all nodes in $\mathscr{F}$ have either two successors or none.

Any node $T \in \mathscr{F}$ is thus uniquely connected with a node T_0 of the initial triangulation $\mathscr{T}_0$, i.e. T belongs to the infinite tree $\mathbb{F}(T_0)$ emanating from T_0. Furthermore, any forest may have *interior nodes*, i.e. nodes with successors, as well as *leaf nodes*, i.e.

nodes without successors. The set of leaves corresponds to a mesh (or triangulation, grid, partition) $\mathscr{T} = \mathscr{T}(\mathscr{F})$ of $\mathscr{T}_0$ which may not be conforming or edge-to-edge.

We thus introduce the set $\mathbb{T}$ of all conforming refinements of $\mathscr{T}_0$:

$$\mathbb{T} := \{\mathscr{T} = \mathscr{T}(\mathscr{F}) \mid \mathscr{F} \subset \mathbb{F} \text{ is finite and } \mathscr{T}(\mathscr{F}) \text{ is conforming}\}.$$

If $\mathscr{T}_* = \mathscr{T}(\mathscr{F}_*) \in \mathbb{T}$ is a conforming refinement of $\mathscr{T} = \mathscr{T}(\mathscr{F}) \in \mathbb{T}$, we write $\mathscr{T}_* \geq \mathscr{T}$ and understand this inequality in the sense of trees, namely $\mathscr{F} \subset \mathscr{F}_*$.

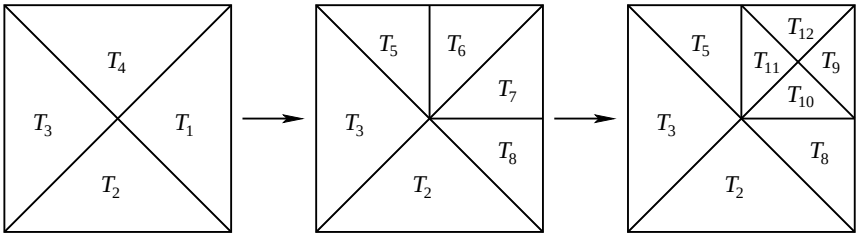
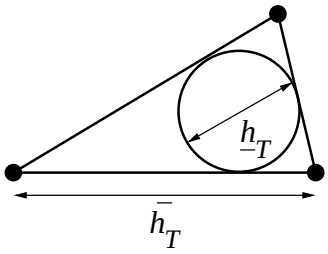


Fig. 2 Sequence of bisection meshes $\{\mathscr{T}_k\}_{k=0}^2$ starting from the initial mesh $\mathscr{T}_0 = \{T_i\}_{i=1}^4$ with longest edges labeled for bisection. Mesh $\mathscr{T}_1$ is created from $\mathscr{T}_0$ upon bisecting T_1 and T_4, whereas mesh $\mathscr{T}_2$ arises from $\mathscr{T}_1$ upon refining T_6 and T_7. The bisection rule is described in Figure 1.

Fig. 3 Forest $\mathscr{F}_2$ corresponding to the grid sequence $\{\mathscr{T}_k\}_{k=0}^2$ of Figure 2. The roots of $\mathscr{F}_2$ form the initial mesh $\mathscr{T}_0$ and the leaves of $\mathscr{F}_2$ constitute the conforming bisection mesh $\mathscr{T}_2$. Moreover, each level of $\mathscr{F}_2$ corresponds to all elements with generation equal to the level.

Example: Consider $\mathscr{T}_0 = \{T_i\}_{i=1}^4$ and the longest edge to be the refinement edge. Figure 2 displays a sequence of conforming meshes $\mathscr{T}_k \in \mathbb{T}$ created by bisection. Each element T_i of $\mathscr{T}_0$ is a root of a finite tree emanating from T_i, which together form the forest $\mathscr{F}_2$ corresponding to mesh $\mathscr{T}_2 = \mathscr{T}(\mathscr{F}_2)$. Figure 3 displays $\mathscr{F}_2$, whose leaf nodes are the elements of $\mathscr{T}_2$.

Properties of Bisection. We now discuss several crucial geometric properties of bisection. We start with the concept of shape regularity. For any $T \in \mathscr{T}$, we define



$$\bar{h}_T = \operatorname{diam}(T)$$
$$h_T = |T|^{1/d}$$
$$\underline{h}_T = 2\sup\{r > 0 \,|\, B(x,r) \subset T \text{ for } x \in T\}.$$

Then

$$\underline{h}_T \le h_T \le \bar{h}_T \le \sigma \underline{h}_T \qquad \forall T \in \mathscr{T},$$

where $\sigma > 1$ is the shape regularity constant. We say that a sequence of meshes is *shape regular* if σ is uniformly bounded, or in other words that the element shape does not degenerate with refinement. The next lemma guarantees that bisection keeps σ bounded.

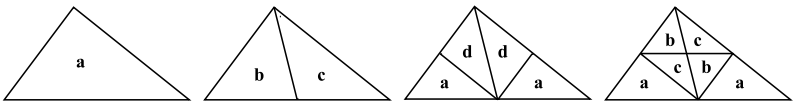


Fig. 4 Bisection produces at most 4 similarity classes for any triangle.

Lemma 1 (Shape Regularity). *The partitions $\mathscr{T}$ generated by newest vertex bisection satisfy a uniform minimal angle condition, or equivalently σ is uniformly bouunded, only depending on the initial partition $\mathscr{T}_0$.*

Proof. Each $T \in \mathscr{T}_0$ gives rise to a fixed number of similarity classes, namely 4 for $d = 2$ according to Figure 4. This, combined with the fact that $\#\mathscr{T}_0$ is finite, yields the assertion. □

We define the *generation (or level)* $g(T)$ of an element $T \in \mathscr{T}$ as the number of bisections needed to create T from its ancestor $T_0 \in \mathscr{T}_0$. Since bisection splits an element into two children with equal measure, we realize that

$$h_T = 2^{-g(T)/2} h_{T_0} \qquad \forall T \in \mathscr{T}. \tag{4}$$

Referring to Figure 3 we observe that the leaf nodes $T_9, T_{10}, T_{11}, T_{12}$ have generation 2, whereas T_5, T_8 have generation 1 and T_2, T_3 have generation 0.

The following geometric property is a simple consequence of (4).

Lemma 2 (Element Size vs Generation). *There exist constants $0 < D_1 < D_2$, only depending on $\mathscr{T}_0$, such that*

$$D_1 2^{-g(T)/2} \le h_T < \bar{h}_T \le D_2 2^{-g(T)/2} \qquad \forall T \in \mathscr{T}. \tag{5}$$

Labeling and Bisection Rule. Whether the recursive application of bisection does not lead to inconsistencies depends on a suitable initial labeling of edges and a bisection rule. For $d = 2$ they are simple to state [7], but for $d > 2$ we refer to Condition (b) of Section 4 of [53]. Given $T \in \mathscr{T}$ with generation $g(T) = i$, we assign the label $(i+1, i+1, i)$ to T with i corresponding to the refinement edge $E(T)$. The following rule dictates how the labeling changes with refinement: the side i is bisected and both new sides as well as the bisector are labeled $i+2$ whereas the remaining labels do not change. To guarantee that the label of an edge is independent of the elements sharing this edge, we need a special labeling for $\mathscr{T}_0$ [7]:

$$\textit{edges of } \mathscr{T}_0 \textit{ have labels } 0 \textit{ or } 1 \textit{ and all elements } T \in \mathscr{T} \textit{ have exactly two edges with label } 1 \textit{ and one with label } 0. \tag{6}$$

It is not obvious that such a labeling exists, but if it does then all elements of $\mathscr{T}_0$ can be split into pairs of compatibly divisible elements. We refer to Figure 5 for an example of initial labeling of $\mathscr{T}_0$ satisfying (6) and the way it evolves for two successive refinements $\mathscr{T}_2 \geq \mathscr{T}_1 \geq \mathscr{T}_0$ corresponding to Figure 2.

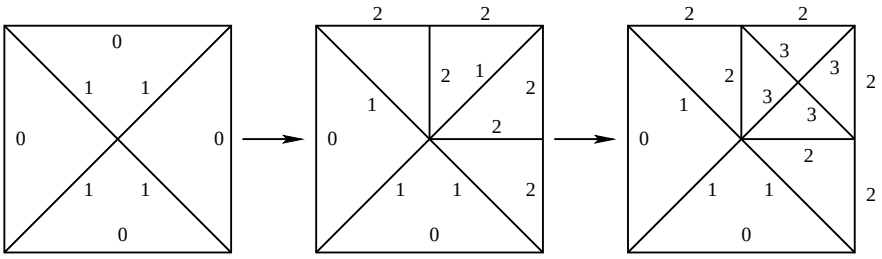


Fig. 5 Initial labeling and its evolution for the sequence of conforming refinements $\mathscr{T}_0 \leq \mathscr{T}_1 \leq \mathscr{T}_2$ of Figure 2.

To guarantee (6) we can proceed as follows: given a coarse mesh of elements T we can bisect twice each T and label the 4 grandchildren, as indicated in Figure 6 for the resulting mesh $\mathscr{T}_0$ to satisfy the initial labeling [7]. A similar, but much trickier,

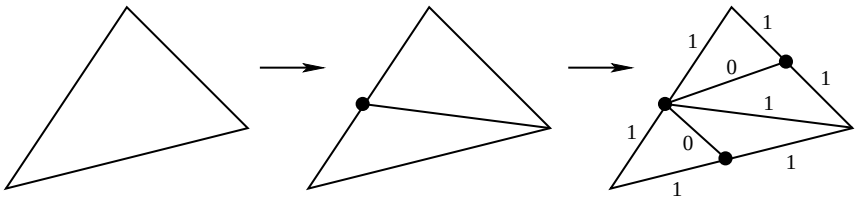


Fig. 6 Bisecting each triangle of $\mathscr{T}_0$ twice and labeling edges in such a way that all boundary edges have label 1 yields an initial mesh satisfying (6).

construction can be made in any dimension $d > 2$ (see Stevenson [53]). For $d = 3$

the number of elements increases by an order of magnitude, which indicates that (6) is a severe restriction in practice. Finding alternative, more practical, conditions is an open and important problem.

The Procedure REFINE. Given $\mathscr{T} \in \mathbb{T}$ and a subset $\mathscr{M} \subset \mathscr{T}$ of marked elements, the procedure

$$\mathscr{T}_* = \mathsf{REFINE}(\mathscr{T}, \mathscr{M})$$

creates a new conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ by bisecting all elements of $\mathscr{M}$ at least once and perhaps additional elements to keep conformity.

Conformity is a constraint in the refinement procedure that prevents it from being completely local. The propagation of refinement beyond the set of marked elements $\mathscr{M}$ is a rather delicate matter, which we discuss later in §6. For instance, we show that a naive estimate of the form

$$\#\mathscr{T}_* - \#\mathscr{T} \leq \Lambda_0 \, \#\mathscr{M}$$

is *not* valid with an absolute constant Λ_0 independent of the refinement level. This can be repaired upon considering the cumulative effect for a sequence of conforming bisection meshes $\{\mathscr{T}_k\}_{k=0}^{\infty}$. This is expressed in the following crucial complexity result due to Binev, Dahmen, and DeVore [7] for $d = 2$ and Stevenson [53] for $d > 2$. We present a complete proof later in §6.

Theorem 1 (Complexity of REFINE). *If $\mathscr{T}_0$ satisfies the initial labeling* (6) *for $d = 2$, or that in [53, Section 4] for $d > 2$, then there exists a constant $\Lambda_0 > 0$ only depending on $\mathscr{T}_0$ and d such that for all $k \geq 1$*

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathscr{M}_j.$$

If elements $T \in \mathscr{M}$ are to be bisected $b \geq 1$ times, then the procedure REFINE can be applied recursively, and Theorem 1 remains valid with Λ_0 also depending on b.

1.4 Finite Element Spaces

Given a conforming mesh $\mathscr{T} \in \mathbb{T}$ we define the finite element space of continuous piecewise polynomials of degree $n \geq 1$

$$\mathbb{S}^{n,0}(\mathscr{T}) := \{v \in C^0(\overline{\Omega}) |\ \ v|_T \in \mathbb{P}_n(T) \ \ \forall T \in \mathscr{T}\};$$

note that $\mathbb{S}^{n,0}(\mathscr{T}) \subset H^1(\Omega)$. We refer to Braess [10], Brenner-Scott [11], Ciarlet [19] and Siebert [50] for a discussion on the local construction of this space along with its properties.

We focus on the piecewise linear case $n = 1$ (Courant elements). Global continuity can be simply enforced by imposing continuity at the vertices z of $\mathscr{T}$, the so-called *nodal values*. We denote by $\mathscr{N}$ the set of vertices z of $\mathscr{T}$.

However, the following local construction leads to global continuity. If T is a generic simplex of $\mathscr{T}$, namely the convex hull of $\{z_i\}_{i=0}^d$, then we associate to each vertex z_i a *barycentric coordinate* λ_i^T, which is the linear function in T with nodal value 1 at z_i and 0 at the other vertices of T. Upon pasting together the barycentric coordinates λ_z^T of all simplices T containing vertex $z \in \mathscr{N}$, we obtain a continuous piecewise linear function $\phi_z \in \mathbb{S}^{1,0}(\mathscr{T})$ as depicted in Figure 7 for $d = 2$: The set

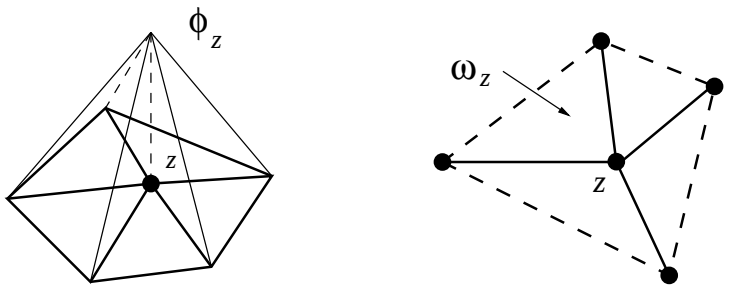


Fig. 7 Piecewise linear basis function ϕ_z corresponding to interior node z, support ω_z of ϕ_z and scheleton γ_z, the latter being composed of all sides within the interior of ω_z.

$\{\phi_z\}_{z\in\mathscr{N}}$ of all such functions is the nodal basis of $\mathbb{S}^{1,0}(\mathscr{T})$, or Courant basis. We denote by $\omega_z := \operatorname{supp}(\phi_z)$ the support of ϕ_z, from now on called *star* associated to z, and by γ_z the scheleton of ω_z, namely all the sides containing z.

We denote functions in $\mathbb{S}^{n,0}(\mathscr{T})$ with capital letters. In view of the definition of ϕ_z, we have the following unique representation of any function $V \in \mathbb{S}^{n,0}(\mathscr{T})$

$$V(x) = \sum_{z\in\mathscr{N}} V(z)\phi_z(x).$$

If we further impose $V(z) = 0$ for all $z \in \partial\Omega \cap \mathscr{N}$, then $V \in H_0^1(\Omega)$. We denote by

$$\mathbb{V}(\mathscr{T}) := \mathbb{S}^{n,0}(\mathscr{T}) \cap H_0^1(\Omega)$$

the subspace of finite element functions which vanish on $\partial\Omega$. Note that we do not explicitly refer to the polynomial degree, which will be clear in each context.

For each simplex $T \in \mathscr{T}$, generated by vertices $\{z_i\}_{i=0}^d$, the *dual functions* $\{\lambda_i^*\}_{i=0}^d \subset \mathbb{P}_1(T)$ to the barycentric coordinates $\{\lambda_i\}_{i=0}^d$ satisfy the bi-orthogonality relation $\int_T \lambda_i^*\lambda_j = \delta_{ij}$, and are given by

$$\lambda_i^* = \frac{(1+d)^2}{|T|}\lambda_i - \frac{1+d}{|T|}\sum_{j\neq i}\lambda_j \qquad \forall\, 0 \le i \le d.$$

The *Courant dual basis* $\phi_z^* \in \mathbb{S}^{n,-1}(\mathscr{T})$ are the discontinuous piecewise linear functions over $\mathscr{T}$ given by

$$\phi_z^* = \frac{1}{\nu_z} \sum_{T \ni z} (\lambda_z^T)^* \chi_T \qquad \forall\, z \in \mathscr{N},$$

where $\nu_z \in \mathbb{N}$ is the valence of z (number of elements of $\mathscr{T}$ containing z) and χ_T is the characteristic function of T. These functions have the same support ω_z as the nodal basis ϕ_z and satisfy the global bi-orthogonality relation

$$\int_\Omega \phi_z^* \phi_y = \delta_{zy} \qquad \forall\, z,y \in \mathscr{N}.$$

1.5 Polynomial Interpolation in Sobolev Spaces

If $v \in C^0(\overline{\Omega})$ we define the *Lagrange interpolant* $I_{\mathscr{T}} v$ of v as follows:

$$I_{\mathscr{T}} v(x) = \sum_{z \in \mathscr{N}} v(z) \phi_z(x).$$

For functions without point values, such as functions in $H^1(\Omega)$ for $d > 1$, we need to determine nodal values by averaging. For any conforming refinement $\mathscr{T} \geq \mathscr{T}_0$ of $\mathscr{T}_0$, the averaging process extends beyond nodes and so gives rise to the discrete neigborhood

$$N_{\mathscr{T}}(T) := \{T' \in \mathscr{T} \mid T' \cap T \neq \emptyset\}$$

for each element $T \in \mathscr{T}$ along with the *local quasi-uniformity* properties

$$\max_{T \in \mathscr{T}} \# N_{\mathscr{T}}(T) \leq C(\mathscr{T}_0), \qquad \max_{T' \in N_{\mathscr{T}}(T)} \frac{|T|}{|T'|} \leq C(\mathscr{T}_0),$$

where $C(\mathscr{T}_0)$ depends only on the shape coefficient of $\mathscr{T}_0$ given by

$$\sigma(\mathscr{T}_0) := \max_{T \in \mathscr{T}_0} \frac{\overline{h}_T}{\underline{h}_T}.$$

We introduce now one such operator $I_{\mathscr{T}}$ due to Scott-Zhang [11, 48], from now on called *quasi-interpolation operator*. We focus on polynomial degree $n = 1$, but the construction is valid for any n; see [11, 48] for details. We recall that $\{\phi_z\}_{z \in \mathscr{N}}$ is the global Lagrange basis of $\mathbb{S}^{1,0}(\mathscr{T})$, $\{\phi_z^*\}_{z \in \mathscr{N}}$ is the global dual basis, and $\operatorname{supp} \phi_z^* = \operatorname{supp} \phi_z$ for all $z \in \mathscr{N}$. We thus define $I_{\mathscr{T}} : L^1(\Omega) \to \mathbb{S}^{1,0}(\mathscr{T})$ to be

$$I_{\mathscr{T}} v = \sum_{z \in \mathscr{N}} \langle v, \phi_z^* \rangle \phi_z,$$

If $0 \le s \le 2$ is a regularity index and $1 \le p \le \infty$ is an integrability index, then we would like to prove the *quasi-local error estimate*

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim h_T^{\mathrm{sob}(W_p^s)-\mathrm{sob}(W_q^t)} \|D^s v\|_{L^p(N_{\mathscr{T}}(T))} \tag{7}$$

for all $T \in \mathscr{T}$, provided $0 \le t \le s$, $1 \le q \le \infty$ are such that $\mathrm{sob}(W_p^s) > \mathrm{sob}(W_q^t)$.

We first observe that by construction $I_{\mathscr{T}}$ is invariant in $\mathbb{S}^{1,0}(\mathscr{T})$, namely,

$$I_{\mathscr{T}} P = P \qquad \text{for all } P \in \mathbb{S}^{1,0}(\mathscr{T}).$$

Since the averaging process giving rise to the values of $I_{\mathscr{T}} v$ for each element $T \in \mathscr{T}$ takes place in the neighborhood $N_{\mathscr{T}}(T)$, we also deduce the local invariance

$$I_{\mathscr{T}} P|_T = P \qquad \text{for all } P \in \mathbb{P}_1(N_{\mathscr{T}}(T))$$

as well as the local stability estimate for any $1 \le q \le \infty$

$$\|I_{\mathscr{T}} v\|_{L^q(T)} \lesssim \|v\|_{L^q(N_{\mathscr{T}}(T))}.$$

We thus may write

$$v - I_{\mathscr{T}} v|_T = (v-P) - I_{\mathscr{T}}(v-P)|_T \qquad \text{for all } T \in \mathscr{T},$$

where $P \in \mathbb{P}_{s-1}$ is arbitrary ($P = 0$ if $s = 0$). It suffices now to prove (7) in the reference element $\widehat{T}$ and scale back and forth to T; the definition (3) of Sobolev number accounts precisely for this scaling. We keep the notation T for $\widehat{T}$, apply the inverse estimate for linear polynomials $\|D^t(I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim \|I_{\mathscr{T}} v\|_{L^q(T)}$ to $v - P$ instead of v, and use the above local stability estimate, to infer that

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim \|v - P\|_{W_q^t(N_{\mathscr{T}}(T))} \lesssim \|v - P\|_{W_p^s(N_{\mathscr{T}}(T))}.$$

The last inequality is a consequence $W_p^s(N_{\mathscr{T}}(T)) \subset W_q^t(N_{\mathscr{T}}(T))$ because $\mathrm{sob}(W_p^s) > \mathrm{sob}(W_q^t)$ and $t \le s$. Estimate (7) now follows from the Bramble-Hilbert lemma [11, Lemma 4.3.8], [19, Theorem 3.1.1]

$$\inf_{P \in \mathbb{P}_{s-1}(N_{\mathscr{T}}(T))} \|v - P\|_{W_p^s(N_{\mathscr{T}}(T))} \lesssim \|D^s v\|_{L^p(N_{\mathscr{T}}(T))}. \tag{8}$$

This proves (7) for $n = 1$. The construction of $I_{\mathscr{T}}$ and ensuing estimate (7) are still valid for any $n > 1$ [11, 48].

Proposition 1 (Quasi-Interpolant without Boundary Values). *Let s, t be regularity indices with $0 \le t \le s \le n+1$, and $1 \le p, q \le \infty$ be integrability indices so that $\mathrm{sob}(W_p^s) > \mathrm{sob}(W_q^t)$.*

There exists a quasi-interpolation operator $I_{\mathscr{T}} : L^1(\Omega) \to \mathbb{S}^{n,0}(\mathscr{T})$, which is invariant in $S^{n,0}(\mathscr{T})$ and satisfies

$$\|D^t(v - I_{\mathscr{T}} v)\|_{L^q(T)} \lesssim h_T^{\mathrm{sob}(W_p^s)-\mathrm{sob}(W_q^t)} \|D^s v\|_{L^p(N_{\mathscr{T}}(T))} \quad \forall T \in \mathscr{T}. \tag{9}$$

The hidden constant in (7) *depends on the shape coefficient of* $\mathscr{T}_0$ *and* d.

To impose a vanishing trace on $I_{\mathscr{T}}v$ we may suitably modify the averaging process for boundary nodes. We thus define a set of dual functions with respect to an L^2-scalar product over $(d-1)$-subsimplices contained on $\partial\Omega$; see again [11, 48] for details. This retains the invariance property of $I_{\mathscr{T}}$ on $\mathbb{S}^{n,0}(\mathscr{T})$ and guarantees that $I_{\mathscr{T}}v$ has a zero trace if $v \in W_1^1(\Omega)$ does. Hence, the above argument applies and (9) follows provided $s \geq 1$.

Proposition 2 (Quasi-Interpolant with Boundary Values). *Let* s,t,p,q *be as in Proposition 1. There exists a quasi-interpolation operator* $I_{\mathscr{T}} : W_1^1(\Omega) \to \mathbb{S}^{n,0}(\mathscr{T})$ *invariant in* $\mathbb{S}^{n,0}(\mathscr{T})$ *which satisfies* (9) *for* $s \geq 1$ *and preserves the boundary values of* v *provided they are piecewise polynomial of degree* $\leq n$. *In particular, if* $v \in W_1^1(\Omega)$ *has a vanishing trace on* $\partial\Omega$, *then so does* $I_{\mathscr{T}}v$.

Remark 3 (Fractional Regularity). We observe that (7) does not require the regularity indices t and s to be integer. The proof follows the same lines but replaces the polynomial degree $s-1$ by the greatest integer smaller that s; the generalization of (8) can be taken from [26].

Remark 4 (Local Error Estimate for Lagrange Interpolant). Let the regularity index s and integrability index $1 \leq p \leq \infty$ satisfy $s - d/p > 0$. This implies that $\mathrm{sob}(W_p^s) > \mathrm{sob}(L^\infty)$, whence $W_p^s(\Omega) \subset C(\overline{\Omega})$ and the Lagrange interpolation operator $I_{\mathscr{T}} : W_p^s(\Omega) \to S^{n,0}(\mathscr{T})$ is well defined and satisfies the *local error estimate*

$$\|D^t(v - I_{\mathscr{T}}v)\|_{L^q(T)} \lesssim h_T^{\mathrm{sob}(W_p^s)-\mathrm{sob}(W_q^t)} \|D^s v\|_{L^p(T)}, \tag{10}$$

provided $0 \leq t \leq s$, $1 \leq q \leq \infty$ are such that $\mathrm{sob}(W_p^s) > \mathrm{sob}(W_q^t)$. We point out that $N_{\mathscr{T}}(T)$ in (7) is now replaced by T in (10). We also remark that if v vanishes on $\partial\Omega$ so does $I_{\mathscr{T}}v$. The proof of (10) proceeds along the same lines as that of Proposition 1 except that the nodal evaluation does not extend beyond the element $T \in \mathscr{T}$ and the inverse and stability estimates over the reference element are replaced by

$$\|D^t I_{\mathscr{T}} v\|_{L^q(\widehat{T})} \lesssim \|I_{\mathscr{T}} v\|_{L^q(\widehat{T})} \lesssim \|v\|_{L^\infty(\widehat{T})} \lesssim \|v\|_{W_p^s(\widehat{T})}.$$

We are now in a position to derive a global interpolation error estimate. To this end, it is convenient to introduce the mesh-size function $h \in L^\infty(\Omega)$ given by

$$h_{|T} = h_T \quad \text{for all } T \in \mathscr{T}. \tag{11}$$

Notice that the following estimate encompasses the linear as well as the nonlinear Sobolev scales.

Theorem 2 (Global Interpolation Error Estimate). *Let* $1 \leq s \leq n+1$ *and* $1 \leq p \leq 2$ *satisfy* $r := \mathrm{sob}(W_p^s) - \mathrm{sob}(H^1) > 0$. *If* $v \in W_p^s(\Omega)$, *then*

$$\|\nabla(v - I_{\mathscr{T}}v)\|_{L^2(\Omega)} \lesssim \|h^r D^s v\|_{L^p(\Omega)}. \tag{12}$$

Proof. Use Proposition 1 along with the elementary property of series $\sum_n a_n \leq (\sum_n a_n^q)^{1/q}$ for $0 < q := p/2 \leq 1$. $\square$

Quasi-Uniform Meshes. We now apply Theorem 2 to quasi-uniform meshes, namely meshes $\mathscr{T} \in \mathbb{T}$ for which all its elements are of comparable size h regardless of the refinement level. In this case, we have

$$h \approx (\#\mathscr{T})^{-1/d}.$$

Corollary 1 (Quasi-Uniform Meshes). *Let $1 \leq s \leq n+1$ and $u \in H^s(\Omega)$. If $\mathscr{T} \in \mathbb{T}$ is quasi-uniform, then*

$$\|\nabla(v - I_{\mathscr{T}} v)\|_{L^2(\Omega)} \lesssim |v|_{H^s(\Omega)} (\#\mathscr{T})^{-(s-1)/d}. \tag{13}$$

Remark 5 (Optimal Rate). If $s = n+1$, and so v has the maximal regularity $v \in H^{n+1}(\Omega)$, then we obtain the optimal convergence rate in a linear Sobolev scale

$$\|\nabla(v - I_{\mathscr{T}} v)\|_{L^2(\Omega)} \lesssim |v|_{H^{n+1}(\Omega)} (\#\mathscr{T})^{-n/d}. \tag{14}$$

The order $-n/d$ is just dictated by the polynomial degree n and cannot be improved upon assuming either higher regularity that $H^{n+1}(\Omega)$ or a graded mesh $\mathscr{T}$.

Example (Corner Singularity in $2d$). To explore the effect of a geometric singularity on (13), we let Ω be the L-shaped domain of Figure 1.5 and $v \in H^1(\Omega)$ be

$$u(r,\theta) = r^{\frac{2}{3}} \sin(2\theta/3) - r^2/4.$$

This function $v \in H^1(\Omega)$ exhibits the typical corner singularity of the solution of $-\Delta v = f$ with suitable Dirichlet boundary condition: $v \in H^s(\Omega)$ for $s < 5/3$. Table 1 displays the best approximation error for polynomial degree $n = 1,2,3$ and the sequence of *uniform* refinements depicted in Figure 1.5 in the seminorm $|\cdot|_{H^1(\Omega)}$. This gives a *lower* bound for the interpolation error in (13).

h	linear ($n=1$)	quadratic ($n=2$)	cubic ($n=3$)
1/4	1.14	9.64	9.89
1/8	0.74	0.67	0.67
1/16	0.68	0.67	0.67
1/32	0.66	0.67	0.67
1/64	0.66	0.67	0.67
1/128	0.66	0.67	0.67

Table 1 The asymptotic rate of convergence in term of mesh-size h is about $h^{2/3}$, or equivalently $(\#\mathscr{T})^{-1/3}$, irrespective of the polynomial degree n. This provides a lower bound for $\|v - I_{\mathscr{T}} v\|_{L^2(\Omega)}$ and thus shows that (13) is sharp.

Even though s is fractional, the error estimate (13) is still valid as stated in Remark 3. In fact, for uniform refinement, (13) can be derived by space interpolation between $H^1(\Omega)$ and $H^{n+1}(\Omega)$. The asymptotic rate $(\#\mathscr{T})^{-1/3}$ reported in Table 1 is

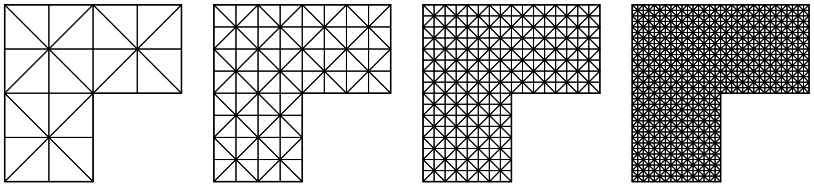


Fig. 8 Sequence of consecutive uniform meshes for L-shaped domain Ω created by 2 bisections.

consistent with (13) and independent of the polynomial degree n; this shows that (13) is sharp. It is also suboptimal as compared with the optimal rate $(\#\mathscr{T})^{-n/2}$ of Remark 5.

The question arises whether the rate $(\#\mathscr{T})^{-1/3}$ in Table 1 is just a consequence of uniform refinement or unavoidable. It is important to realize that $v \notin H^s(\Omega)$ for $s \geq 5/3$ and thus (13) is not applicable. However, the problem is not that second order derivatives of v do not exist but rather that they are not square-integrable. In particular, it is true that $v \in W^2_p(\Omega)$ if $1 \leq p < 3/2$. We therefore may apply Theorem 2 with, e.g., $n = 1$, $s = 2$, and $p \in [1, 3/2)$ and then ask whether the structure of (12) can be exploited, e.g., by compensating the local behavior of $D^s u$ with the local mesh-size h. This enterprise naturally leads to *graded* meshes adapted to v.

1.6 Adaptive Approximation

Principle of Error Equidistribution. We investigate the relation between local mesh-size and regularity for the design of graded meshes adapted to $v \in H^1(\Omega)$ for $d = 2$. We formulate this as an optimization problem:

> *Given a function $v \in C^2(\Omega) \cap W^2_p(\Omega)$ and an integer $N > 0$ find conditions for a shape regular mesh $\mathscr{T}$ to minimize the error $|v - I_{\mathscr{T}} v|_{H^1(\Omega)}$ subject to the constraint that the number of degrees of freedom $\#\mathscr{T} \leq N$.*

We first convert this *discrete* optimization problem into a *continuous model*, following Babuška and Rheinboldt [5]. Let

$$\#\mathscr{T} = \int_\Omega \frac{dx}{h(x)^2}$$

be the number of elements of $\mathscr{T}$ and let the Lagrange interpolation error

$$\|\nabla(v - I_{\mathscr{T}} v)\|^p_{L^2(\Omega)} = \int_\Omega h(x)^{2(p-1)} |D^2 v(x)|^p dx$$

be dictated by (12) with $s = 2$ and $1 < p \leq 2$; note that $r = \mathrm{sob}(W^2_p) - \mathrm{sob}(H^1) = 2 - 2/p$ whence $rp = 2(p-1)$ is the exponent of $h(x)$. We next propose the Lagrangian

$$\mathscr{L}[h,\lambda] = \int_\Omega \Big(h(x)^{2(p-1)} |D^2 v(x)|^p - \frac{\lambda}{h(x)^2} \Big) dx$$

with Lagrange multiplier $\lambda \in \mathbb{R}$. The optimality condition reads (Problem 4)

$$h(x)^{2(p-1)+2} |D^2 v(x)|^p = \Lambda \tag{15}$$

where $\Lambda > 0$ is a constant. In order to interpret this expression, we compute the interpolation error E_T incurred in element $T \in \mathscr{T}$. According to (10), E_T is given by

$$E_T^p \approx h_T^{2(p-1)} \int_T |D^2 v(x)|^p \approx \Lambda$$

provided $D^2 v(x)$ is about constant in T. Therefore we reach the heuristic, but insightful, conclusion that E_T is about constant, or equivalently

A graded mesh is quasi-optimal if the local error is equidistributed. (16)

Corner Singularities. Meshes satisfying (16) have been constructed by Babuška et al [3] for corner singularities and $d = 2$; see also [30]. If the function v possess the typical behavior

$$v(x) \approx r(x)^\gamma, \quad 0 < \gamma < 1,$$

where $r(x)$ is the distance from $x \in \Omega$ to a reentrant corner of Ω, then (15) implies the mesh grading

$$h(x) = \Lambda^{\frac{1}{2p}} r(x)^{-\frac{1}{2}(\gamma-2)}$$

whence

$$\#\mathscr{T} = \int_\Omega h(x)^{-2} dx = \Lambda^{-\frac{1}{p}} \int_0^{\operatorname{diam}(\Omega)} r^{\gamma-1} dr \approx \Lambda^{-\frac{1}{p}}.$$

This crucial relation is valid for any $\gamma > 0$ and $p > 1$; in fact the only condition on p is that $r = 2 - 2/p > 0$, or equivalently $\operatorname{sob}(W_p^2) > \operatorname{sob}(H^1)$. Therefore,

$$\|\nabla(v - I_{\mathscr{T}} v)\|_{L^2(\Omega)}^2 = \sum_{T \in \mathscr{T}} E_T^2 = \Lambda^{\frac{2}{p}} (\#\mathscr{T}) = (\#\mathscr{T})^{-1} \tag{17}$$

gives the optimal decay rate for $d = 2, n = 1$, according to Remark 5. We explore the case $d \geq 2$ and $n \geq 1$ in Problem 6. What this argument does not address is whether such meshes $\mathscr{T}$ exist in general and, more importantly, whether they can actually be constructed upon bisecting the initial mesh $\mathscr{T}_0$ so that $\mathscr{T} \in \mathbb{T}$.

Thresholding. We now construct graded bisection meshes $\mathscr{T}$ for $n = 1, d = 2$ that achieve the optimal decay rate $(\#\mathscr{T})^{-1/2}$ of (14) and (17) under the global regularity assumption

$$v \in W_p^2(\Omega), \quad p > 1. \tag{18}$$

Following the work of Binev et al. [8], we use a thresholding algorithm that is based on the knowledge of the element errors and on bisection. The algorithm hinges

on (16): if $\delta > 0$ is a given tolerance, the element error is equidistributed, that is $E_T \approx \delta^2$, and the global error decays with maximum rate $(\#\mathscr{T})^{-1/2}$, then

$$\delta^4 \#\mathscr{T} \approx \sum_{T\in\mathscr{T}} E_T^2 = |v - I_{\mathscr{T}} v|^2_{H^1(\Omega)} \lesssim (\#\mathscr{T})^{-1}$$

that is $\#\mathscr{T} \lesssim \delta^{-2}$. With this in mind, we impose $E_T \leq \delta^2$ as a common threshold to stop refining and expect $\#\mathscr{T} \lesssim \delta^{-2}$. The following algorithm implements this idea.

Thresholding Algorithm. Given a tolerance $\delta > 0$ and a conforming mesh $\mathscr{T}_0$, THRESHOLD finds a conforming refinement $\mathscr{T} \geq \mathscr{T}_0$ of $\mathscr{T}_0$ by bisection such that $E_T \leq \delta^2$ for all $T \in \mathscr{T}$: let $\mathscr{T} = \mathscr{T}_0$ and

THRESHOLD$(\mathscr{T}, \delta)$
while $\mathscr{M} := \{T \in \mathscr{T} \,|\, E_T > \delta^2\} \neq \emptyset$
  $\mathscr{T} :=$ REFINE$(\mathscr{T}, \mathscr{M})$
end while
return$(\mathscr{T})$

We get $W^2_p(\Omega) \subset C^0(\overline{\Omega})$, because $p > 1$, and can use the Lagrange interpolant and local estimate (10) with $r = \mathrm{sob}(W^2_p) - \mathrm{sob}(H^1) = 2 - 2/p > 0$. We deduce that

$$E_T \lesssim h_T^r \, \|D^2 v\|_{L^p(T)}, \tag{19}$$

and that THRESHOLD *terminates* because h_T decreases monotonically to 0 with bisection. The quality of the resulting mesh is assessed next.

Theorem 3 (Thresholding). *If $v \in H^1_0(\Omega)$ verifies* (18)*, then the output $\mathscr{T} \in \mathbb{T}$ of* THRESHOLD *satisfies*

$$|v - I_{\mathscr{T}} v|_{H^1(\Omega)} \leq \delta^2 (\#\mathscr{T})^{1/2}, \qquad \#\mathscr{T} - \#\mathscr{T}_0 \lesssim \delta^{-2} \, |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega)}.$$

Proof. Let $k \geq 1$ be the number of iterations of THRESHOLD before termination. Let $\mathscr{M} = \mathscr{M}_0 \cup \cdots \cup \mathscr{M}_{k-1}$ be the set of marked elements. We organize the elements in $\mathscr{M}$ by size in such a way that allows for a counting argument. Let $\mathscr{P}_j$ be the set of elements T of $\mathscr{M}$ with size

$$2^{-(j+1)} \leq |T| < 2^{-j} \quad \Rightarrow \quad 2^{-(j+1)/2} \leq h_T < 2^{-j/2}.$$

We proceed in several steps.

1 We first observe that all T's in $\mathscr{P}_j$ are *disjoint*. This is because if $T_1, T_2 \in \mathscr{P}_j$ and $\mathring{T}_1 \cap \mathring{T}_2 \neq \emptyset$, then one of them is contained in the other, say $T_1 \subset T_2$, due to the bisection procedure. Thus

$$|T_1| \leq \frac{1}{2} |T_2|$$

contradicting the definition of $\mathscr{P}_j$. This implies

$$2^{-(j+1)} \, \#\mathscr{P}_j \leq |\Omega| \quad \Rightarrow \quad \#\mathscr{P}_j \leq |\Omega| \, 2^{j+1}. \tag{20}$$

[2] In light of (19), we have for $T \in \mathscr{P}_j$

$$\delta^2 \le E_T \lesssim 2^{-(j/2)r} \|D^2 v\|_{L^p(T)}.$$

Therefore

$$\delta^{2p} \# \mathscr{P}_j \lesssim 2^{-(j/2)rp} \sum_{T \in \mathscr{P}_j} \|D^2 v\|^p_{L^p(T)} \le 2^{-(j/2)rp} \|D^2 v\|^p_{L^p(\Omega)}$$

whence

$$\# \mathscr{P}_j \lesssim \delta^{-2p} 2^{-(j/2)rp} \|D^2 v\|^p_{L^p(\Omega)}. \tag{21}$$

[3] The two bounds for $\# \mathscr{P}$ in (20) and (21) are complementary. The first is good for j small whereas the second is suitable for j large (think of $\delta \ll 1$). The crossover takes place for j_0 such that

$$2^{j_0+1} |\Omega| = \delta^{-2p} 2^{-j_0(rp/2)} \|D^2 v\|^p_{L^p(\Omega)} \quad \Rightarrow \quad 2^{j_0} \approx \delta^{-2} \frac{\|D^2 v\|_{L^p(\Omega)}}{|\Omega|^{1/p}}.$$

[4] We now compute

$$\# \mathscr{M} = \sum_j \# \mathscr{P}_j \lesssim \sum_{j \le j_0} 2^j |\Omega| + \delta^{-2p} \|D^2 v\|^p_{L^p(\Omega)} \sum_{j > j_0} (2^{-rp/2})^j.$$

Since

$$\sum_{j \le j_0} 2^j \approx 2^{j_0}, \qquad \sum_{j > j_0} (2^{-rp/2})^j \lesssim 2^{-(rp/2) j_0} = 2^{-(p-1) j_0}$$

we can write

$$\# \mathscr{M} \lesssim \left(\delta^{-2} + \delta^{-2p} \delta^{2(p-1)} \right) |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega)} \approx \delta^{-2} |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega)}.$$

We finally apply Theorem 1 to arrive at

$$\# \mathscr{T} - \# \mathscr{T}_0 \lesssim \# \mathscr{M} \lesssim \delta^{-2} |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega)}.$$

[5] It remains to estimate the energy error. We have, upon termination of THRESHOLD, that $E_T \le \delta^2$ for all $T \in \mathscr{T}$. Then

$$|v - I_{\mathscr{T}} v|^2_{H^1(\Omega)} = \sum_{T \in \mathscr{T}} E_T^2 \le \delta^4 \# \mathscr{T}.$$

This concludes the Theorem. □

By relating the threshold value δ and the number of refinements N, we obtain a result about the convergence rate.

Corollary 2 (Convergence Rate). *Let $v \in H^1_0(\Omega)$ satisfy* (18)*. Then for $N > \# \mathscr{T}_0$ integer there exists $\mathscr{T} \in \mathbb{T}$ such that*

$$|v-I_{\mathscr{T}}v|_{H^1(\Omega)} \lesssim |\Omega|^{1-1/p}\,\|D^2v\|_{L^p(\Omega)}N^{-1/2}, \qquad \#\mathscr{T}-\#\mathscr{T}_0 \lesssim N.$$

Proof. Choose $\delta^2 = |\Omega|^{1-1/p}\,\|D^2v\|_{L^p(\Omega)}N^{-1}$ in Theorem 3. Then, there exists $\mathscr{T}\in\mathbb{T}$ such that $\#\mathscr{T}-\#\mathscr{T}_0 \lesssim N$ and

$$\begin{aligned}|v-I_{\mathscr{T}}v|_{H^1(\Omega)} &\lesssim |\Omega|^{1-1/p}\,\|D^2v\|_{L^p(\Omega)}N^{-1}\big(N+\#\mathscr{T}_0\big)^{1/2}\\ &\lesssim |\Omega|^{1-1/p}\,\|D^2v\|_{L^p(\Omega)}N^{-1/2}\end{aligned}$$

because $N>\#\mathscr{T}_0$. This finishes the Corollary. □

Remark 6 (Piecewise smoothness). The global regularity (18) can be weakened to *piecewise W^2_p regularity* over the initial mesh $\mathscr{T}_0$, namely $W^2_p(\Omega;\mathscr{T}_0)$, and global $H^1_0(\Omega)$. This is because $W^2_p(T)\hookrightarrow C^0(\overline{T})$ for all $T\in\mathscr{T}_0$, whence $I_{\mathscr{T}}$ can be taken to be the Lagrange interpolation operator.

Remark 7 (Case $p<1$). We consider now polynomial degree $n\geq 1$. The integrability p corresponding to differentiability $n+1$ results from equating Sobolev numbers:

$$n+1-\frac{d}{p}=\mathrm{sob}(H^1)=1-\frac{d}{2} \quad\Rightarrow\quad p=\frac{2d}{2n+d}.$$

Depending on $d\geq 2$ and $n\geq 1$, this may lead to $0<p<1$, in which case $W^{n+1}_p(\Omega)$ is to be replaced by the Besov space $B^{n+1}_{p,p}(\Omega)$ [22]; see Problem 6. The argument of Theorem 3 works provided we replace (19) by a modulus of regularity; in fact, $D^{n+1}v$ would not be locally integrable and so would fail to be a distribution.

Remark 8 (Isotropic vs anisotropic elements). Theorem 3 and Problem 5 show that isotropic graded meshes can always deal with geometric singularities for $d=2$. This is no longer the case for $d>2$ and is explored in Problem 6.

1.7 Nonconforming Meshes

More general subdivisions of Ω than those in §1.3 are used in practice. If the elements of $\mathscr{T}_0$ are quadrilaterals for $d=2$, or their multidimensional variant for $d>2$, then it is natural to allow for improper or *hanging nodes* for the resulting refinements $\mathscr{T}$ to be graded; see Figure 9 (a). On the other hand, if $\mathscr{T}_0$ is made of triangles for $d=2$, or simplices for $d>2$, then red refinement without green completion also gives rise to graded meshes with hanging nodes; see Figure 9 (b). In both cases, the presence of hanging nodes is inevitable to enforce mesh grading. Finally, bisection may produce meshes with hanging nodes, as depicted in Figure 9 (c), if the completion process is incomplete. All three refinements maintain shape regularity, but for both practice and theory, they cannot be arbitrary: we need to restrict the level of incompatibility; see Problem 10. We discuss this next.

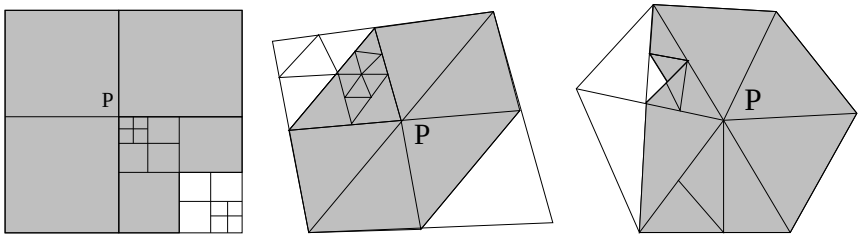


Fig. 9 Nonconforming meshes made of quadrilaterals (a), triangles with red refinement (b), and triangles with bisection (c). The shaded regions depict the domain of influence of a proper or conforming node P.

We start with the notion of domain of influence of a proper node, introduced by Babuška and Miller in the context of K-meshes [4]; see Figure 9. For simplicity, we restrict ourselves to polynomial degree $n = 1$. We say that a node P of $\mathscr{T}$ is a *proper* (or conforming) node if it is a vertex of all elements containing P; otherwise, we say that P is an *improper* (nonconforming or hanging) node. Since we only prescribe degrees of freedom at the proper nodes, it is natural to describe the canonical continuous piecewise linear basis functions ϕ_P associated with each proper node P.

We do this recursively. As in §1.3, the *generation* $g(T)$ of an element $T \in \mathscr{T}$ is the number of subdivisions needed to create T from its ancestor in the initial mesh $\mathscr{T}_0$, hereafter assumed to be conforming. We first rearrange the elements in $\mathscr{T} = \{T_i\}_{i=1}^{\#\mathscr{T}}$ by generation: $g(T_i) \le g(T_{i+1})$ for all $i \ge 0$. Suppose that ϕ_P has been already defined for each $T \in \mathscr{T}$ with $g(T) < i$. We proceed as follows to define ϕ_P at each vertex z of $T \in \mathscr{T}$ with $g(T) = i$:

- if z is a proper node, then we set $\phi_P(z) = 1$ if $z = P$ and $\phi_P(z) = 0$ otherwise;
- if z is a hanging node, then z belongs to an edge of another element $T' \in \mathscr{T}$ with $g(T') < i$ and set $\phi_P(z)|_T = \phi_P(z)|_{T'}$.

This definition is independent of the choice of T' since, by construction, ϕ_P is continuous across interelements of lower level. We also observe that $\{\phi_P\}_{P\in\mathscr{N}}$ is a basis of the finite element space $\mathbb{V}(\mathscr{T})$ of *continuous* piecewise linear functions, thus

$$V = \sum_{P\in\mathscr{N}} V(P)\phi_P \qquad \forall\, V \in \mathbb{V}(\mathscr{T}).$$

The *domain of influence* of a proper node P is the support of ϕ_P:

$$\omega_{\mathscr{T}}(T) = \operatorname{supp}(\phi_P).$$

We say that a sequence of nonconforming meshes $\{\mathscr{T}\}$ is *admissible* if there is a universal constant $\Lambda_* \le 1$, independent of the refinement level and $\mathscr{T}$, such that

$$\operatorname{diam}(\omega_{\mathscr{T}}(T)) \le \Lambda_* h_T \qquad \forall\, T \in \mathscr{T}. \tag{22}$$

An important example is quadrilaterals with *one* hanging node per edge. We observe, however, that (22) can neither be guaranteed with more than one hanging node per edge for quadrilaterals, nor for triangles with one hanging node per edge (see Problem 10).

Given an admissible grid $\mathscr{T}$, a subset $\mathscr{M}$ of elements marked for refinement, and a desired number $\rho \geq 1$ of subdivisions to be performed in each marked element, the procedure

$$\mathscr{T}_* = \mathsf{REFINE}(\mathscr{T}, \mathscr{M})$$

creates a minimal admissible mesh $\mathscr{T}_* \geq \mathscr{T}$ such that all the elements of $\mathscr{M}$ are subdivided at least ρ times. In order for $\mathscr{T}_*$ to be admissible, perhaps other elements not in $\mathscr{M}$ must be partitioned. Despite the fact that admissibility is a constraint on the refinement procedure weaker than conformity, it cannot avoid the propagation of refinements beyond $\mathscr{M}$. The complexity of REFINE is again an issue which we discuss in §6.4: we show that Theorem 1 extends to this case.

Lemma 3 (REFINE for Nonconforming Meshes). *Let $\mathscr{T}_0$ be an arbitrary conforming partition of Ω, except for bisection in which case $\mathscr{T}_0$ satisfies the labeling* (6) *for $d = 2$ or its higher dimensional counterpart [53]. Then the estimate*

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathscr{M}_j \quad \forall k \geq 1$$

holds with a constant Λ_0 depending on $\mathscr{T}_0, d$ and ρ.

We conclude by emphasizing that the polynomial interpolation and adaptive approximation theories of §§1.5 and 1.6 extend to nonconforming meshes with fixed level of incompatibility as well.

1.8 Notes

The use of Sobolev numbers is not so common in the finite element literature, but allows as to write compact error estimates and speak about nonlinear Sobolev scale. The latter concept is quite natural in nonlinear approximation theory [22].

The discussion of bisection for $d = 2$ follows Binev, Dahmen, and DeVore [7]. Stevenson extended the theory to $d > 2$ [53]. We refer to the survey by Nochetto, Siebert and Veeser [45] for a rather complete discussion for $d > 1$, and to §6 for a proof of Theorem 1 for $d = 2$, which easily extends to $d > 2$.

The discussion of finite element spaces [10, 11, 19] and polynomial interpolation [11, 26, 48] is rather classical. In contrast, the material of adaptive approximation is much less documented. The principle of equidistribution goes back to Babuška and Rheimboldt [5] and the a priori design of optimal meshes for corner singularities for $d = 2$ is due to Babuška, Kellogg, and Pitkäranta [3]. The construction of optimal meshes via bisection using thresholding is extracted from Binev, Dahmen, DeVore, and Petrushev [8].

Finally the discussion of nonconforming meshes follows Bonito and Nochetto [9], and continues in §6 with the proof of Lemma 3.

1.9 Problems

Problem 1 (Nonconforming element). Given a d-simplex T in R^d with vertices $z_0,\dots,z_d$, construct a basis $\lambda_0,\dots,\lambda_d$ of $\mathbb{P}_1(T)$ such that $\lambda_i(\overline{z}_j)=\delta_{ij}$ for all $i,j\in\{1,\dots,d\}$, where $\overline{z}_j$ denotes the barycenter of the face opposite to the vertex z_j. Does this local basis also lead to a global one in $\mathbb{S}^{1,0}(\mathscr{T})$?

Problem 2 (Quadratic basis functions). Express the nodal basis of $\mathbb{P}_2(T)$ in terms of barycentric coordinates of $T\in\mathscr{T}$.

Problem 3 (Quadratic dual functions). Derive expressions for the dual functions of the quadratic local Lagrange basis of $\mathbb{P}_2(T)$ for each element $T\in\mathscr{T}$. Construct a global discontinuous dual basis $\phi_z^*\in\mathbb{S}^{2,-1}(\mathscr{T})$ of the global Lagrange basis $\phi_z\in\mathbb{S}^{2,0}(\mathscr{T})$ for all $z\in\mathscr{N}_2(\mathscr{T})$.

Problem 4 (Lagrangian). Let $h(x)$ be a smooth function locally equivalent to the mesh-size and $v\in C^2(\Omega)\cap W_p^2(\Omega)$. Prove that a stationary point of the Lagrangian

$$\mathscr{L}[h,\lambda]=\int_\Omega\Big(h(x)^{2(p-1)}\,|D^2v(x)|^p-\frac{\lambda}{h(x)^2}\Big)\,dx$$

satisfies the optimality condition

$$h(x)^{2(p-1)+2}\,|D^2v(x)|^p=\text{constant}.$$

Problem 5 (W_p^2-regularity). Consider the function $v(r,\theta)=r^\gamma\phi(\theta)$ in polar coordinates (r,θ) for $d=2$ with $\phi(\theta)$ smooth. Show that $v\in W_p^2(\Omega)\setminus H^2(\Omega)$ for $1\le p<2/(2-\gamma)$.

Problem 6 (Edge singularities). This problem explores *formally* the effect of edge singularities for dimension $d>2$ and polynomial degree $n\ge1$. Since edge (or line) singularities are two dimensional locally, away from corners, we assume the behavior $v(x)\approx r(x)^\gamma$ where $r(x)$ is the distance of $x\in\Omega$ to an edge of Ω and $\gamma>0$.
(a) Use the Principle of Equidistribution with $p=2$ to determine the mesh grading

$$h(x)\approx\Lambda^{\frac{1}{2n+d}}\,r(x)^{2d\frac{\gamma-(n+1)}{2n+d}}.$$

(b) Show the following relation between Λ and number of elements $\#\mathscr{T}=\int_\Omega h(x)^{-d}$

$$\gamma>\frac{(d-2)n}{d}\quad\Rightarrow\quad\#\mathscr{T}\approx\Lambda^{-\frac{d}{2n+d}}.$$

(c) If $\gamma>\frac{(d-2)n}{d}$, then deduce the optimal interpolation error decay

$$\|\nabla(v - I_{\mathscr{T}} v)\|_{L^2(\Omega)} \lesssim (\#\mathscr{T})^{-\frac{n}{d}}.$$

(d) Prove that $\gamma > \frac{(d-2)n}{d}$ is equivalent to the regularity $\int_\Omega |D^{n+1} v|^p < \infty$ for $p > \frac{2d}{2n+d}$. If $\tau := \frac{2d}{2n+d} \geq 1$, then this would mean $v \in W_p^{n+1}(\Omega)$. However, it is easy to find examples $d > 2$ or $n > 1$ for which $\tau < 1$, in which case the Sobolev space $W_p^{n+1}(\Omega)$ must be replaced by the Besov space $B_{p,p}^{n+1}(\Omega)$ [22]. Note that $p > \tau$ is precisely what yields the crucial relation between Sobolev numbers

$$\mathrm{sob}(B_{p,p}^{n+1}) = n + 1 - \frac{d}{p} > \mathrm{sob}(H^1) = 1 - \frac{d}{2}.$$

We observe that for $d = 2$ all singular exponents $\gamma > 0$ lead to optimal meshes, but this is not true for $d = 3$: $n = 1$ requires $\gamma > \frac{1}{3}$ whereas $n = 2$ needs $\gamma > \frac{2}{3}$. The latter corresponds to a dihedral angle $\omega > \frac{3\pi}{2}$ and can be easily checked computationally. We thus conclude that *isotropic* graded meshes are sufficient to deal with geometric singularities for $d = 2$ but not for $d > 2$, for which *anisotropic* graded meshes are the only ones which exhibit optimal behavior. Their adaptive construction is open.

Problem 7 (Local H^2-regularity). Consider the function $v(x) \approx r(x)^\gamma$ where $r(x)$ is the distance to the origin and $d = 2$.
(a) Examine the construction of a graded mesh via the thresholding algorithm.
(b) Repeat the proof of Theorem 3 replacing the W_p^2 regularity by the corresponding local H^2-regularity of v depending on the distance to the origin.

Problem 8 (Thresholding for $d > 2$). Let $d > 2$, $n = 1$, and $v \in W_p^2(\Omega)$ with $p > \frac{2d}{2+d}$. This implies that $v \in H^1(\Omega)$ but not necessarily in $C^0(\overline{\Omega})$. Use the quasi-interpolant $I_{\mathscr{T}}$ of Proposition 1 to define the local H^1-error E_T for each element $T \in \mathscr{T}$ and use the thresholding algorithm to show Theorem 3 and Corollary 2.

Problem 9 (Reduced rate). Let $d \geq 2$, $n = 1$, and $v \in W_p^s(\Omega)$ with $1 < s < 2$ and $\mathrm{sob}(W_p^s) > \mathrm{sob}(H^1)$, namely $s - \frac{d}{p} > 1 - \frac{d}{2}$. Use the quasi-interpolant $I_{\mathscr{T}}$ of Proposition 1 to define the local H^1-error E_T for each element $T \in \mathscr{T}$ and use the thresholding algorithm to show Corollary 2: given $N > \#\mathscr{T}_0$ there exists $\mathscr{T} \in \mathbb{T}$ with $\#\mathscr{T} - \#\mathscr{T}_0 \lesssim N$ such that

$$\|v - I_{\mathscr{T}} v\|_{H^1(\Omega)} \lesssim \|D^s v\|_{L^p(\Omega)} N^{-\frac{s-1}{d}}.$$

Problem 10 (Level of incompatibility). This problem shows that keeping the number of hanging nodes per side bounded does not guarantee a bounded level of incompatibility for $d = 2$. The situation is similar for $d > 2$.
(a) *Square elements*: construct a selfsimilar quad-refinement of the unit square with only 2 hanging nodes per side and unbounded level of incompatibility.
(b) *Triangular elements*: construct selfsimilar red-refinements and bisection refinements of the unit reference triangle with 1 hanging node per side and unbounded level of incompatibility.

Problem 11 (Quasi-interpolation of discontinuous functions). Let $\mathscr{T}$ be an admissible nonconforming mesh. Let $\mathbb{V}(\mathscr{T})$ denote the space of discontinuous piecewise polynomials of degree $\leq n$ over $\mathscr{T}$, and $\mathbb{V}^0(\mathscr{T})$ be the subspace of continuous functions. Construct a local quasi-interpolation operator $I_{\mathscr{T}} : \mathbb{V}(\mathscr{T}) \to \mathbb{V}^0(\mathscr{T})$ with the following approximation property for all $V \in \mathbb{V}(\mathscr{T})$ and $|\alpha| = 0, 1$

$$\|D^\alpha (V - I_{\mathscr{T}} V)\|_{L^2(T)} \lesssim h_T^{\frac{1-|\alpha|}{2}} \| [\![V]\!] \|_{L^2(\Sigma_{\mathscr{T}}(T))} \quad \forall T \in \mathscr{T},$$

where $\Sigma_{\mathscr{T}}(T)$ stands for all sides within $N_{\mathscr{T}}(T)$ and $[\![V]\!]$ denotes the jump of V across sides.

2 Error Bounds for Finite Element Solutions

In §1 we have seen that approximating a given known function with meshes which are adapted to that function can impressively outperform the approximation with quasi-uniform meshes. In view of the fact that the solution of a boundary value problem is given only implicitly, it is not all clear if this is also true for its adaptive numerical solution. Considering a simple model problem and discretization, we now derive two upper bounds for the error of the finite element solution: an a priori one and an a posteriori one. The a priori bound reveals that an adaptive variant of the finite element method has the potential of a similar performance. The a posteriori bound is a first step to design such a variant, which has to face the complication that the target function is given only implicitly.

2.1 Model Boundary Value Problem

In order to minimize technicalities in the presentation, let us consider the following simple boundary value problem as a model problem: find a scalar function $u = u(x)$ such that

$$\begin{aligned} -\operatorname{div}(\boldsymbol{A}\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{23}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary $\partial\Omega$, $\boldsymbol{A} = \boldsymbol{A}(x)$ a map into the positive definite $d \times d$ matrices, and $f = f(x)$ a scalar load term. Introducing the Hilbert space

$$\mathbb{V} := H_0^1(\Omega) := \{v \in H^1(\Omega) \mid v_{|\partial\Omega} = 0\}, \qquad \|v\|_{\mathbb{V}} := \left(\int_\Omega |\nabla v|^2\right)^{1/2},$$

and the bilinear form

$$\mathscr{B}[v, w] := \int_\Omega \boldsymbol{A}\nabla v \cdot \nabla w, \qquad v, w \in \mathbb{V},$$

the weak solution of (23) is characterized by

$$u \in \mathbb{V}: \qquad \mathscr{B}[u, v] = \langle f, v\rangle \qquad \text{for all } v \in \mathbb{V}. \tag{24}$$

Hereafter $\langle\cdot,\cdot\rangle$ stands for the $L^2(\Omega)$-scalar product and also for a duality paring. We assume that $f \in \mathbb{V}^* = H^{-1}(\Omega) := H_0^1(\Omega)^*$ and that there exist constants $0 < \alpha_1 \le \alpha_2$ with

$$\forall x \in \Omega,\ \xi \in \mathbb{R}^d \quad \alpha_1|\xi|^2 \le \boldsymbol{A}(x)\xi \cdot \xi \text{ and } |\boldsymbol{A}(x)\xi| \le \alpha_2|\xi|. \tag{25}$$

The latter implies that the operator $-\operatorname{div}(\boldsymbol{A}\nabla\cdot)$ is uniformly elliptic. Moreover, the bilinear form $\mathscr{B}$ is coercive and continuous with constants α_1 and α_2, respectively.

Lax-Milgram Theorem and Poincaré-Friedrichs Inequality

$$\|v\|_\Omega \leq \operatorname{diam}(\Omega)\|\nabla v\|_\Omega \quad \text{for all } v \in \mathbb{V} = H^1_0(\Omega) \tag{26}$$

thus ensure existence and uniqueness of the weak solution (24).

Note that $\boldsymbol{A}$ is not assumed to be symmetric and so the bilinear form $\mathscr{B}$ may be nonsymmetric. For the a posteriori upper bound, we will require some additional regularity on the data f and $\boldsymbol{A}$ in §2.4.

2.2 Galerkin Solutions

Since $\mathbb{V}$ has infinite dimension, problem (24) cannot be implemented on a computer and solved numerically. Given a subspace $\mathbb{S} \subset \mathbb{V}$, the corresponding Galerkin solution or approximation of (24) is given by

$$U \in \mathbb{S}: \qquad \mathscr{B}[U,V] = \langle f, V\rangle \qquad \text{for all } V \in \mathbb{S}. \tag{27}$$

We simply replaced each occurence of $\mathbb{V}$ in (24) by $\mathbb{S}$. If $\mathbb{S}$ is finite-dimensional, we can choose a basis of $\mathbb{S}$ and the coefficients of the expansion of U can be determined by solving a square linear system.

Residual. Associate the functional $\mathscr{R} \in \mathbb{V}^*$ given by

$$\langle \mathscr{R}, v\rangle := \langle f, v\rangle - \mathscr{B}[U,v],$$

to $U \in \mathbb{S}$. The functional $\mathscr{R}$ is called the residual and depends only on the approximate solution U and data $\boldsymbol{A}$ and f. Moreover, it has the following properties:

- It relates to the typically unknown error function $u-U$ in the following manner:

$$\langle \mathscr{R}, v\rangle = \mathscr{B}[u-U,v] \qquad \text{for all } v \in \mathbb{V}. \tag{28}$$

  This is a direct consequence of the characterization (24) of the exact solution.
- It vanishes for discrete test functions, which in the case of symmetric $\boldsymbol{A}$ corresponds to the so-called Galerkin orthogonality:

$$\mathscr{B}[u-U,V] = \langle \mathscr{R}, V\rangle = 0 \quad \text{for all } V \in \mathbb{S}. \tag{29}$$

  This immediately follows from (28) and the definition (27) of the Galerkin solution.

Quasi-Best Approximation. Property (25) of $\boldsymbol{A}$, Galerkin orthogonality (29) and Cauchy-Schwarz Inequality in $L^2(\Omega)$ imply

$$\begin{aligned}\alpha_1\|u-U\|^2_{\mathbb{V}} &\leq \mathscr{B}[u-U,u-U] = \mathscr{B}[u-U,u-V]\\ &\leq \alpha_2\|u-U\|_{\mathbb{V}}\|u-V\|_{\mathbb{V}}\end{aligned}$$

for arbitrary $V \in \mathbb{S}$. This proves the famous

Theorem 4 (Céa Lemma). *The Galerkin solution is a quasi-best approximation from* $\mathbb{S}$ *with respect to the* $\mathbb{V}$*-norm:*

$$\|u-U\|_{\mathbb{V}} \leq \frac{\alpha_2}{\alpha_1} \inf_{V\in S} \|u-V\|_{\mathbb{V}}. \tag{30}$$

If the bilinear $\mathscr{B}$ is also symmetric and one considers the error with respect to the so-called energy norm $\mathscr{B}[\cdot,\cdot]^{1/2}$, the Galerkin solution is even the best approximation from $\mathbb{S}$; see Problem 12.

2.3 Finite Element Solutions and A Priori Bound

Problem (27) can be solved numerically on a computer, if we dispose of an implementable basis of $\mathbb{S}$. As an example of such space, let $\mathscr{T}$ be a conforming triangulation of Ω into d-simplices (this imposes further conditions on Ω) and consider

$$\mathbb{S} = \mathbb{V}(\mathscr{T}) := \{V \in \mathbb{S}^{n,0}(\mathscr{T}) \mid V_{|\partial\Omega} = 0\}, \tag{31}$$

where, as in §1.4, $\mathbb{S}^{n,0}(\mathscr{T})$ the space of continuous functions that are piecewise polynomial up to degree n. This is in fact a subspace of $\mathbb{V} = H^1_0(\Omega)$ thanks to the continuity requirement and boundary condition for the functions in $\mathbb{V}(\mathscr{T})$. Moreover, the basis $\{\phi_z\}_{z\in\mathscr{N}\cap\Omega}$ from §1.4 can be easily constructed in the computer; see for example Siebert [50].

The space $\mathbb{V}(\mathscr{T})$ is a popular choice for $\mathbb{S}$ in (27) and their combination may be viewed as a model finite element discretization.

In §1.5 we studied the approximation properties of $\mathbb{S}^{n,0}(\mathscr{T})$ with the help of (quasi-)interpolation operators $I_{\mathscr{T}}$. Since the right-hand side of (30) is bounded in terms of $\|u - I_{\mathscr{T}}u\|_{\mathbb{V}} = \|\nabla(u - I_{\mathscr{T}}u)\|_{L^2(\Omega)}$ with $I_{\mathscr{T}}$ as in Proposition 2, the discussion of §1.5 applies also to the error of the Galerkin solution $U_{\mathscr{T}}$ in $\mathbb{V}(\mathscr{T})$. In particular, the combination of the Céa Lemma and Theorem 2 yields the following upper bound. Since it does not involve the discrete solution, it is also comes with the adjective 'a priori'.

Theorem 5 (A priori upper bound). *Assume that the exact solution* u *of* (24) *satisfies* $u \in W^{s,p}(\Omega)$ *with* $1 \leq s \leq n+1$, $1 \leq p \leq 2$, *and set*

$$r := \operatorname{sob}(W^s_p(\Omega)) - \operatorname{sob}(H^1(\Omega)) > 0.$$

Then the error of the finite element solution $U_{\mathscr{T}} \in \mathbb{S} = \mathbb{V}(\mathscr{T})$ *of* (27) *satisfies the global a priori upper bound*

$$\|u - U_{\mathscr{T}}\|_{\mathbb{V}} \lesssim \|h^r D^s u\|_{L^p(\Omega)}. \tag{32}$$

The discussion in §1.6 about adaptively graded meshes only partially carries over to the error of the finite element solution $U_{\mathscr{T}}$, from now on denoted U. In view of the Céa Lemma, §1.6 shows that there are sequences of meshes such that the error of U decays as $\#\mathscr{T}^{-1/2}$ if, for example, $d = 2$ and $u \in W^{2,p}(\Omega)$ with $p > 1$. Notice however that the thresholding algorithm utilizes the local errors $E_T = \|\nabla(u - I_{\mathscr{T}} u)\|_{L^2(T)}$, which are typically not computable. The construction of appropriate meshes when the target function is given only implicitly by a boundary value problem is much more subtle. A first step towards this goal is developed in the next section.

2.4 A Posteriori Upper Bound

The *a priori* upper bound (32) is not computable and essentially provides only asymptotic information, namely the asymptotic convergence rate. The goal of this section is to derive an alternative bound, so-called *a posteriori* bound, that provides information beyond asymptotics and is computable in terms of data and the approximate solution. It is worth noting that such bounds are useful not only for adaptivity but also for the quality assessment of the approximate solution.

Since in this section the grid $\mathscr{T}$ is (arbitrary but) fixed, we simplify the notation by suppressing the subscript indicating the dependence on the grid in case of the approximate solution and similar quantities.

Error and Residual. Our starting point is the algebraic relationship (28) between the residual $\mathscr{R}$ and the error function $u - U$. It implies (Problem 13)

$$\|u - U\|_{\mathbb{V}} \leq \frac{1}{\alpha_1} \|\mathscr{R}\|_{\mathbb{V}^*} \leq \frac{\alpha_2}{\alpha_1} \|u - U\|_{\mathbb{V}}, \tag{33}$$

which means that the dual norm

$$\|\ell\|_{\mathbb{V}^*} := \sup \big\{ \langle \ell, v \rangle \mid v \in \mathbb{V}, \|v\|_{\mathbb{V}} \leq 1 \big\} \tag{34}$$

is a good measure for the residual $\mathscr{R}$ if we are interested in the error $\|u - U\|_{\mathbb{V}} = \|\nabla(u - U)\|_{L^2(\Omega)}$. However the evaluation of $\|\mathscr{R}\|_{\mathbb{V}^*} = \|\mathscr{R}\|_{H^{-1}(\Omega)}$ is impractical and, moreover, does not provide local information for guiding an adaptive mesh refinement. We therefore aim at a sharp upper bound of $\|\mathscr{R}\|_{H^{-1}(\Omega)}$ in terms of locally computable quantitites.

Assumptions and Structure of Residual. For the derivation of a computable upper bound of the dual norm of the residual, we require that

$$f \in L^2(\Omega) \quad \text{and} \quad \boldsymbol{A} \in W^{1,\infty}(\Omega; \mathscr{T}) \tag{35}$$

where the latter means that $\boldsymbol{A}$ is Lipschitz in each element of $\mathscr{T}$. Under these assumptions, we can write $\langle \mathscr{R}, v \rangle$ as integrals over elements $T \in \mathscr{T}$ and elementwise integration by parts yields the representation:

$$\begin{aligned}\langle \mathscr{R}, v\rangle &= \int_\Omega fv - \boldsymbol{A}\nabla U\cdot\nabla v = \sum_{T\in\mathscr{T}}\int_T fv - \boldsymbol{A}\nabla U\cdot\nabla v \\ &= \sum_{T\in\mathscr{T}}\int_T rv + \sum_{S\in\mathscr{S}}\int_S jv,\end{aligned} \tag{36}$$

where

$$\begin{aligned} &r = f + \operatorname{div}(\boldsymbol{A}\nabla U) \quad \text{in any simplex } T\in\mathscr{T},\\ &j = [\![\boldsymbol{A}\nabla U]\!]\cdot\boldsymbol{n} = \boldsymbol{n}^+\cdot\boldsymbol{A}\nabla U_{|T^+} + \boldsymbol{n}^-\cdot\boldsymbol{A}\nabla U_{|T^-} \quad \text{on any internal side } S\in\mathscr{S}\end{aligned} \tag{37}$$

and $\boldsymbol{n}^+$, $\boldsymbol{n}^-$ are unit normals pointing towards T^+, $T^-\in\mathscr{T}$. We see that the distribution $\mathscr{R}$ consists of a regular part r, called *interior or element residual*, and a singular part j, called *jump or interelement residual*. The regular part is absolutely continuous w.r.t. the d-dimensional Lebesgue measure and is related to the strong form of the PDE. The singular part is supported on the skeleton $\Gamma = \bigcup_{S\in\mathscr{S}} S$ of $\mathscr{T}$ and is absolutely continuous w.r.t. the $(d-1)$-dimensional Hausdorff measure.

We point out that this structure of the residual is not special to the model problem and its discretization but rather arises from the weak formulation of the PDE and the piecewise construction of finite element spaces.

Scaled Integral Norms. In view of the structure of the residual $\mathscr{R}$, we make our goal precise as follows: we aim at a sharp upper bound for $\|\mathscr{R}\|_{H^{-1}(\Omega)}$ in terms of local Lebesgue norms of the element and interelement residuals r and j, which are considered to be computable because they can be easily approximated with numerical integration. This approach is usually called standard a posteriori error estimation.

The sharpness of these bounds crucially hinges on appropriate local scaling constants for the aforementioned Lebesgue norms, which depend on the local geometry of the mesh. For simplicity, we will explicitly trace only the dependence on the local mesh-size and write '$\lesssim$' instead of '$\le C$', where the constant C is bounded in terms of the shape coefficient $\sigma(\mathscr{T}) = \max_{T\in\mathscr{T}} \bar{h}_T/\underline{h}_T$ of the triangulation $\mathscr{T}$ and the dimension d.

Localization. As a first step, we decompose the residual $\mathscr{R}$ into local contributions with the help of the Courant basis $\{\phi_z\}_{z\in\mathscr{V}}$ from §1.4. Hereafter $\mathscr{V}$ stands for the set of vertices of $\mathscr{T}$, which coincide with the nodes of $\mathbb{S}^{1,0}(\mathscr{T})$. The Courant basis has the following properties:

- It provides a partition of unity:

$$\sum_{z\in\mathscr{V}}\phi_z = 1 \quad \text{in } \Omega. \tag{38}$$

- For each interior vertex z, the corresponding basis function ϕ_z is contained in $\mathbb{V}(\mathscr{T})$ and so the residual is orthogonal to the interior contributions of the partition of unity:

$$\langle \mathscr{R}, \phi_z\rangle = 0 \quad \text{for all } z\in\mathring{\mathscr{V}} := \mathscr{V}\cap\Omega. \tag{39}$$

The second property corresponds to the Galerkin orthogonality. Notice that the first property involves all vertices, while in the second one the boundary vertices are excluded.

Given any $v \in H^1_0(\Omega)$, we apply (38) and then (39) to write

$$\langle \mathscr{R}, v\rangle = \sum_{z\in\mathscr{V}} \langle \mathscr{R}, v\phi_z\rangle = \sum_{z\in\mathscr{V}} \langle \mathscr{R}, (v-c_z)\phi_z\rangle, \tag{40}$$

where $c_z \in \mathbb{R}$ and $c_z = 0$ whenever $z \in \partial\Omega$. Exploiting representation (36), $0 \le \phi_z \le 1$, and the fact that the ϕ_z are locally supported, we can bound each local contribution $\langle \mathscr{R}, (v-c_z)\phi_z\rangle$ in the following manner:

$$|\langle \mathscr{R}, (v-c_z)\phi_z\rangle| \le \left|\int_{\omega_z} r(v-c_z)\phi_z\right| + \left|\int_{\gamma_z} j(v-c_z)\phi_z\right|, \tag{41}$$

where $\omega_z := \cup_{T\ni z} T$ is the star (or patch) around a vertex $z \in \mathscr{V}$ in $\mathscr{T}$ and γ_z is the skeleton of ω_z, i.e. the union of all sides emanating from z; note that r in (41) is computed elementwise. We examine the two terms on the right-hand side separately.

Bounding the Element Residual. We first consider the terms associated with the element residual r. The key tool for a sharp bound is the following local Poincaré-type inequality. Let

$$h_z := |\omega_z|^{1/d}$$

and notice that this quantity is, up to the shape coefficient $\sigma(\mathscr{T})$, equivalent to the diameter of ω_z, to $h_T = |T|^{1/d}$ if T is a d-simplex of ω_z and to $h_S := |S|^{1/(d-1)}$ if S is a side of γ_z.

Lemma 4 (Local Poincaré-type inequality). *For any $v \in H^1_0(\Omega)$ and $z \in \mathscr{V}$ there exists $c_z \in \mathbb{R}$ such that*

$$\|v-c_z\|_{L^2(\omega_z)} \lesssim h_z \|\nabla v\|_{L^2(\omega_z)}. \tag{42}$$

If $z \in \partial\Omega$ is a boundary vertex, then we can take $c_z = 0$.

We postpone the proof of Lemma 4. Combining the Cauchy-Schwarz inequality in $L^2(\omega_z)$ and Lemma 4 readily yields

$$\left|\int_{\omega_z} r(v-c_z)\phi_z\right| \le \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} \|v-c_z\|_{L^2(\omega_z)} \lesssim h_z \|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} \|\nabla v\|_{L^2(\omega_z)}. \tag{43}$$

Notice that the right-hand side consists of two factors: a computable one in the desired form and one that involves the test function in a local variant of the norm of the test space.

Bounding the Jump Residual. Next, we consider the terms associated to the jump residual j. Recall that j is supported on sides and so proceeding similarly as for the element residual will bring up traces of the test function. The following trace inequality exactly meets our needs.

Lemma 5 (Scaled trace inequality). *For any side S of a d-simplex T the following inequality holds:*

$$\|w\|_{L^2(S)} \lesssim h_S^{-1/2}\|w\|_{L^2(T)} + h_S^{1/2}\|\nabla w\|_{L^2(T)} \quad \text{for all } w \in H^1(T). \tag{44}$$

We again postpone the proof, now of Lemma 5. We apply first the Cauchy-Schwarz inequality in $L^2(\gamma_z)$, then Lemma 5 and finally Lemma 4 to obtain

$$\left|\int_{\gamma_z} j(v-c_z)\phi_z\right| \le \|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}\|v-c_z\|_{L^2(\gamma_z)} \lesssim h_z^{1/2}\|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}\|\nabla v\|_{L^2(\omega_z)}, \tag{45}$$

where the right-hand side has the same structure as that of the element residual.

Upper Bound for Residual Norm. We collect the local estimates and sum them up in order to arrive at the desired bound for the dual norm of the residual. Inserting the estimates (43) and (45) for element and jump residuals into (41) gives

$$|\langle \mathscr{R}, v\,\phi_z\rangle| \lesssim \left(h_z\|r\,\phi_z^{1/2}\|_{L^2(\omega_z)} + h_z^{1/2}\|j\,\phi_z^{1/2}\|_{L^2(\gamma_z)}\right)\|\nabla v\|_{L^2(\omega_z)}.$$

Recalling the decomposition (40), we sum over $z \in \mathscr{V}$ and use Cauchy-Schwarz in $\mathbb{R}^{\#\mathscr{T}}$ to arrive at

$$|\langle \mathscr{R}, v\rangle| \lesssim \left(\sum_{z\in\mathscr{V}} h_z^2\|r\,\phi_z^{1/2}\|^2_{L^2(\omega_z)} + h_z\|j\,\phi_z^{1/2}\|^2_{L^2(\gamma_z)}\right)^{1/2}\left(\sum_{z\in\mathscr{V}}\|\nabla v\|^2_{L^2(\omega_z)}\right)^{1/2}.$$

For bounding the second factor, we resort to the finite overlapping property of stars, namely

$$\sum_{z\in\mathscr{V}} \chi_{\omega_z}(x) \le d+1,$$

and infer that

$$\sum_{z\in\mathscr{V}}\|\nabla v\|^2_{L^2(\omega_z)} \lesssim \|\nabla v\|^2_{L^2(\Omega)}.$$

Since mesh refinement is typically based upon element subdivision, we regroup the terms within the first factor. To this end, denote by $h\colon \Omega \to \mathbb{R}^+$ the mesh-size function given by $h(x) := |S|^{1/k}$ if x belongs to the interior of the k-subsimplex S of $\mathscr{T}$ with $k \in \{1,\dots,d\}$. Then for all $x \in \omega_z$ we have $h_z \lesssim h(x)$. Therefore employing (38) once more and recalling that Γ is the union of all interior sides of $\mathscr{T}$, we deduce

$$\begin{aligned}\sum_{z\in\mathscr{V}} h_z^2\|r\,\phi_z^{1/2}\|^2_{L^2(\omega_z)} + h_z\|j\,\phi_z^{1/2}\|^2_{L^2(\gamma_z)} &\lesssim \sum_{z\in\mathscr{V}}\|h\,r\,\phi_z^{1/2}\|^2_{L^2(\Omega)} + \|h^{1/2}\,j\,\phi_z^{1/2}\|^2_{L^2(\Gamma)}\\ &= \|h\,r\|^2_{L^2(\Omega)} + \|h^{1/2}\,j\|^2_{L^2(\Gamma)}.\end{aligned}$$

Thus, introducing the *element indicators*

$$\mathscr{E}^2_{\mathscr{T}}(U,T) := h_T^2\|r\|^2_{L^2(T)} + h_T\|j\|^2_{L^2(\partial T\setminus\partial\Omega)} \tag{46}$$

and the *error estimator*

$$\mathscr{E}_{\mathscr{T}}^2(U) = \sum_{T\in\mathscr{T}} \mathscr{E}_{\mathscr{T}}^2(U,T) \tag{47}$$

we arrive at the following upper bound for the dual norm of the residual:

$$\|\mathscr{R}\|_{H^{-1}(\Omega)} \lesssim \mathscr{E}_{\mathscr{T}}(U). \tag{48}$$

Hereafter, we write $\mathscr{E}_{\mathscr{T}}(U,\mathscr{M})$ to indicate that the estimator is computed over $\mathscr{M}\subset\mathscr{T}$, whereas $\mathscr{E}_{\mathscr{T}}(U,\mathscr{T}) = \mathscr{E}_{\mathscr{T}}(U)$ if no confusion arises.

Proofs of Poincaré-Type and Trace Inequalities. We now prove Lemmas 4 and 5. We start with a formula for the mean value of a trace, which follows from the Divergence Theorem.

Lemma 6 (Trace identity). *Let T be a d-simplex, S a side of T, and z the vertex opposite to S. Defining the vector field $\boldsymbol{q}_S$ by*

$$\boldsymbol{q}_S(x) := x - z$$

the following equality holds

$$\frac{1}{|S|}\int_S w = \frac{1}{|T|}\int_T w + \frac{1}{d|T|}\int_T \boldsymbol{q}_S\cdot\nabla w \quad \textit{for all } w\in W_1^1(T).$$

Proof. We start with properties of the vector field $\boldsymbol{q}_S$. Let S' be an arbitrary side of T and fix some $y\in S'$. We then see $\boldsymbol{q}_S(x)\cdot\boldsymbol{n}_T = \boldsymbol{q}_S(y)\cdot\boldsymbol{n}_T + (x-y)\cdot\boldsymbol{n}_T = \boldsymbol{q}_S(y)\cdot\boldsymbol{n}_T$ for any $x\in S'$ since $x-y$ is a tangent vector to S'. Therefore, on each side of T, the associated normal flux $\boldsymbol{q}_S\cdot\boldsymbol{n}_T$ is constant. In particular, we see $\boldsymbol{q}_S\cdot\boldsymbol{n}_T$ vanishes on $\partial T\setminus S$ by choosing $y=z$ for sides emanating from z. Moreover, $\operatorname{div}\boldsymbol{q}_S = d$. Thus, if $w\in C^1(\overline{T})$, the Divergence Theorem yields

$$\int_T \boldsymbol{q}_S\cdot\nabla w = -d\int_T w + (\boldsymbol{q}_S\cdot\boldsymbol{n}_T)_{|S}\int_S w.$$

Take $w=1$ to show $(\boldsymbol{q}_S\cdot\boldsymbol{n}_T)_{|S} = d|T|/|S|$ and extend the result to $w\in W_1^1(T)$ by density. □

Proof of Lemma 5. Apply Lemma 6 to $|w|^2$; for the details see Problem 17. □

Proof of Lemma 4. 1 For any $z\in\mathscr{V}$ the value

$$\bar{c}_z = \frac{1}{|\omega_z|}\int_{\omega_z} v$$

is an optimal choice and (42) follows from (8) with $c_z = \bar{c}_z$.
2 If $z\in\partial\Omega$, then we observe that there exists a side $S\subset\partial\omega_z\cap\partial\Omega$ such that $v=0$ on S. We therefore can write

$$v = v - \frac{1}{|S|}\int_S v = (v-\bar{c}_z) - \frac{1}{|S|}\int_S (v-\bar{c}_z)$$

whence, using Lemma 5 and Step 1 for the second term,

$$\|v\|_{L^2(\omega_z)} \lesssim \|v-\overline{c}_z\|_{L^2(\omega_z)} + h_z\|\nabla v\|_{L^2(\omega_z)} \lesssim h_z\|\nabla v\|_{L^2(\omega_z)},$$

which establishes the supplement for boundary vertices. □

Upper Bound for Error. Inserting the bound (48) for the dual norm of the residual in the first bound of (33), we obtain the main result of this section.

Theorem 6 (A posteriori upper bound). *Let u be the exact solution of the model problem* (24) *satisfying* (25) *and* (35)*. The error of the finite element solution* $U \in \mathbb{S} = \mathbb{V}(\mathscr{T})$ *of* (27) *is bounded in terms of the estimator* (47) *as follows:*

$$\|u-U\|_{\mathbb{V}} \lesssim \frac{1}{\alpha_1}\mathscr{E}_{\mathscr{T}}(U), \tag{49}$$

where the hidden constant depends only on the shape coefficient $\sigma(\mathscr{T})$ *of the triangulation* $\mathscr{T}$ *and on the dimension d.*

Notice that the a posteriori bound in Theorem 6 does not require additional regularity on the exact solution *u* as the a priori one in Theorem 5. On the other hand, the dependence of the estimator on the approximate solution prevents us from directly extracting information such as asymptotic decay rate of the error. The question thus arises how sharp the a posteriori bound in Theorem 6 is.

In this context it is worth noticing that if we did not exploit orthogonality and used a global Poincaré-type inequality instead of the local ones, the resulting scalings of the element and jump residuals would be, respectively, 1 and $h_T^{-1/2}$ and the corresponding upper bound would have a lower asymptotic decay rate. We will show in the next §3 that the upper bound in Theorem 6 is sharp in an asymptotic sense.

2.5 Notes

The discussion of the quasi-best approximation and the a priori upper bound of the error of the finite element solution are classical; see Braess [10], Brenner-Scott [11], and Ciarlet [19]. The core of the a posteriori upper bound is a bound of the dual norm of the residual in terms of scaled Lebesgue norms. This approach is usually called *standard a posteriori error estimation* and has been successfully used for a variety of problems and discretizations. For alternative approaches we refer to the monographs of Ainsworth and Oden [2] and Verfürth [58] on a posteriori errror estimation.

Typically standard a posteriori error estimation is carried out with the help of error estimates for quasi-interpolation as in §1.5, which in turn rely on local Bramble-Hilbert lemmas. The above presentation invokes only the special case of Poincaré-type inequalities. It is a simplified version of derivation in Veeser and Verfürth [56], which has been influenced by Babuška and Rheinboldt [5], Carstensen and Funken

[12], and Morin, Nochetto and Siebert [42], and provides in particular constants that are explicit in terms of local Poincaré constants. It is worth mentioning that the ensuing constants are found in [56] for sample meshes and have values close to 1.

The setting and assumption of the model problem and discretization in this section avoids the following complications: numerical integration, approximation of boundary values, approximation of the domain, and inexact solution of the discrete system. While all these issues have been analyzed in an a priori context, only some of them have been considered in a posteriori error estimation; see Ainsworth and Kelly [1], Dörfler and Rumpf [25], Morin, Nochetto, and Siebert [42], Nochetto, Siebert, Schmidt and Veeser [46], and Sacchi and Veeser [47].

2.6 Problems

Problem 12 (Best approximation for symmetric problems). Consider the model problem (24), assume in addition to (25) that A is symmetric and denote the energy norm associated with the differential operator $-\operatorname{div}(A\nabla\cdot)$ by

$$|||v|||_\Omega := \left(\int_\Omega A\nabla v\cdot\nabla v\right)^{1/2}.$$

Prove that the Galerkin solution is the best approximation from $\mathbb{S}=\mathbb{V}(\mathscr{T})$ with respect to the energy norm:

$$|||u-U|||_\Omega = \min_{V\in\mathbb{S}} |||u-V|||_\Omega. \tag{50}$$

Derive from this that in this case (30) improves to

$$\|u-U\|_{\mathbb{V}} \le \sqrt{\frac{\alpha_2}{\alpha_1}} \inf_{V\in\mathbb{S}} \|u-V\|_{\mathbb{V}}.$$

Problem 13 (Equivalence of error and residual norm). Prove the equivalence (33) between error and dual norm of the residual. Consider the model problem also with a symmetric $\boldsymbol{A}$ and derive a similar relationship for the energy norm error.

Problem 14 (Dominance of jump residual). Considering the model problem (24) and its discretization (27) with (31) and $n=1$, show that, up to higher order terms, the jump residual

$$\eta_{\mathscr{T}}(U) = \Big(\sum_{S\in\mathscr{S}} \|h^{1/2} j\|^2_{L^2(S)}\Big)^{1/2}$$

bounds $\|\mathscr{R}\|_{H^{-1}(\Omega)}$, which entails that the estimator $\mathscr{E}_{\mathscr{T}}(U)$ is dominated by $\eta_{\mathscr{T}}(U)$. To this end, revise the proof of the upper bound for $\|\mathscr{R}\|_{H^{-1}(\Omega)}$, use

$$c_z = \frac{1}{\int_{\omega_z}\phi_z}\int_{\omega_z} f\,\phi_z.$$

and rewrite $\int_{\omega_z} f\,(v-c_z)\,\phi_z$ by exploiting this weighted L^2-orthogonality.

Problem 15 (A posteriori upper bound with quasi-interpolation). Consider the model problem (24) and its discretization (27) with space $\mathbb{S}=\mathbb{V}(\mathscr{T})$, and derive the upper a posteriori error bound without using the discrete partition of unity. To this end, use (36) and combine the scaled trace inequality (44) with the local interpolation error estimate (7). Show as an intermediate step the upper bound

$$|\langle\mathscr{R},v\rangle|\lesssim\sum_{T\in\mathscr{T}}\mathscr{E}_{\mathscr{T}}(U,T)\|\nabla v\|_{L^2(N_{\mathscr{T}}(T))} \tag{51}$$

with $N_{\mathscr{T}}(T)$ from §1.5. This bound will be useful in §4.

Problem 16 (Upper bound for certain singular loads). Revising the proof of Theorem 6, derive an a posteriori upper bound in the case of right-hand sides of the form

$$\langle f,v\rangle=\int_\Omega g_0v+\int_\Gamma g_1v,\qquad v\in\mathbb{V}=H^1_0(\Omega),$$

where $g_0\in L^2(\Omega)$, $g_1\in L^2(\Gamma)$, and Γ stands for the skeleton of the mesh $\mathscr{T}$.

Problem 17 (Scaled trace inequality). Work out the details of the proof of Lemma 5, taking into account that $h_T\approx|T|/|S|\approx h_S$.

Problem 18 (A posteriori upper bound for L^2-error). Assuming that Ω is convex and applying a duality argument, establish a variant of (33) between the L^2-error $\|u-U\|_{L^2(\Omega)}$ and a suitable dual norm of the residual. Use this to derive the a posteriori upper bound

$$\|u-U\|_{L^2(\Omega)}\lesssim\left(\sum_{T\in\mathscr{T}}h_T^2\mathscr{E}_{\mathscr{T}}(U,T)^2\right)^{1/2},$$

where the hidden constant depends in addition on the domain Ω.

3 Lower A Posteriori Bounds

The goal of this section is to assess the sharpness of the a posteriori upper bound for the model problem and discretization. We show not only that it is sharp in an asymptotic sense like the a priori bound but also in a local sense and, for certain data, in a non-asymptotic sense. Moreover, we verify that the latter cannot be true for all data and argue that this is the price to pay for the upper bound to be computable.

As in §2.4, '$\lesssim$' stands for '$\leq C$', where the constant C is bounded in terms of the shape coefficent $\sigma(\mathscr{T})$ of the triangulation $\mathscr{T}$ and the dimension d and, often, we do not indicate the dependence on the arbitrary but fixed triangulation.

3.1 Local Lower Bounds

The first step in the derivation of the upper bound (49) is that the error is bounded in terms of an appropriate dual norm of the residual. In the case of the model problem (24) this relies on the continuity of $[-\operatorname{div}(\boldsymbol{A}\nabla\cdot)]^{-1} : H^{-1}(\Omega) \to H^1_0(\Omega)$. Notice that the inverse is a global operator, while $-\operatorname{div}(\boldsymbol{A}\nabla\cdot)$ in the classical sense is a local one. One thus may suspect that an appropriate local dual norm of the residual is bounded in terms of the local error. Let us verify this for the model problem (24).

Local Dual Norms. Let ω be a subdomain of Ω and notice that $H^{-1}(\omega)$ is a good candidate for the local counterpart of $H^{-1}(\Omega)$. Given $v \in H^1_0(\omega)$ (and extending it by zero on $\Omega \setminus \omega$), the algebraic relationship (28), the Cauchy-Schwarz inequality in $L^2(\omega)$, and (25) readily yield

$$\langle \mathscr{R}, v\rangle = \mathscr{B}[u-U, v] = \int_\omega \boldsymbol{A}\nabla(u-U)\cdot\nabla v \leq \alpha_2 \|\nabla(u-U)\|_{L^2(\omega)} \|\nabla v\|_{L^2(\omega)}.$$

Consequently,

$$\|\mathscr{R}\|_{H^{-1}(\omega)} \leq \alpha_2 \|\nabla(u-U)\|_{L^2(\omega)}, \tag{52}$$

entailing that lower bounds for the local error $\|\nabla(u-U)\|_{L^2(\omega)}$ may be shown by bounding the local dual norm $\|\mathscr{R}\|_{H^{-1}(\omega)}$ from below.

Local Dual and Scaled Integral Norms. As for the a posteriori upper bound, we assume (35). If we take $\omega = T \in \mathscr{T}$ in the preceding paragraph, then there holds

$$\|\mathscr{R}\|_{H^{-1}(T)} = \sup_{\|\nabla v\|_{L^2(T)} \leq 1} \langle \mathscr{R}, v\rangle = \sup_{\|\nabla v\|_{L^2(T)} \leq 1} \int_T rv = \|r\|_{H^{-1}(T)} \tag{53}$$

thanks to the representation (36). Recall that the corresponding indicator $\mathscr{E}_{\mathscr{T}}(U,T)$ contains the term $h_T\|r\|_{L^2(T)}$ and therefore we wonder about the relationship of $\|r\|_{H^{-1}(T)}$ and $h_T\|r\|_{L^2(T)}$. Mimicking the local part in the derivation of the a posteriori upper bound in §2.4, we obtain

$$\int_T rv \le \|r\|_{L^2(T)}\|v\|_{L^2(T)} \lesssim h_T\|r\|_{L^2(T)}\|\nabla v\|_{L^2(T)}$$

with the help of the Poincaré-Friedrichs inequality (26). Hence there holds

$$\|r\|_{H^{-1}(T)} \lesssim h_T\|r\|_{L^2(T)}. \tag{54}$$

Since $L^2(\Omega)$ is a proper subspace of $H^{-1}(\Omega)$, the inverse inequality cannot hold for arbitrary r. Consequently, $h_T\|r\|_{L^2(T)}$ may overestimate $\|r\|_{H^{-1}(T)}$. On the other hand, if $r \in \mathbb{R}$ is *constant* and $\eta = \eta_T$ denotes a non-negative function with properties

$$|T| \lesssim \int_T \eta, \quad \operatorname{supp}\eta = T, \quad \|\nabla\eta\|_{L^\infty(T)} \lesssim h_T^{-1} \tag{55}$$

(postpone the question of existence until (59) below), we deduce

$$\begin{aligned}\|r\|^2_{L^2(T)} &\lesssim \int_T r(r\eta) \le \|r\|_{H^{-1}(T)}\|\nabla(r\eta)\|_{L^2(T)}\\ &\le \|r\|_{H^{-1}(T)}\|r\|_{L^2(T)}\|\nabla\eta\|_{L^\infty(T)} \lesssim h_T^{-1}\|r\|_{H^{-1}(T)}\|r\|_{L^2(T)},\end{aligned}$$

whence

$$h_T\|r\|_{L^2(T)} \lesssim \|r\|_{H^{-1}(T)}. \tag{56}$$

This shows that overestimation in (54) is caused by *oscillation* of r at a scale finer than the mesh-size. Notice that (56) is a so-called inverse estimate, where one norm is a dual norm. It is also valid for $r \in \mathbb{P}_n(T)$, but the constant deteriorates with the degree n; see Problem 22.

Local Lower Bound with Element Residual. Motivated by the observations of the preceding paragraph, we expect that $h_T\|r\|_{L^2(T)}$ bounds asymptotically $\|\mathscr{R}\|_{H^{-1}(T)}$ from below and introduce the *oscillation of the interior residual* in T defined by

$$h_T\|r-\bar r_T\|_{L^2(T)},$$

where $\bar r_T$ denotes the mean value of r in T. Replacing r by $\bar r_T$ in (56) and by $r-\bar r_T$ in (54), as well as recalling (53), we derive

$$\begin{aligned}h_T\|r\|_{L^2(T)} &\le h_T\|\bar r_T\|_{L^2(T)} + h_T\|r-\bar r_T\|_{L^2(T)}\\ &\lesssim \|\bar r_T\|_{H^{-1}(T)} + h_T\|r-\bar r_T\|_{L^2(T)}\\ &\lesssim \|r\|_{H^{-1}(T)} + \|r-\bar r_T\|_{H^{-1}(T)} + h_T\|r-\bar r_T\|_{L^2(T)}\\ &\lesssim \|\mathscr{R}\|_{H^{-1}(T)} + h_T\|r-\bar r_T\|_{L^2(T)}.\end{aligned} \tag{57}$$

This is the desired statement because the oscillation $h_T\|r-\bar r_T\|_{L^2(T)}$ is expected to convergence faster than $h_T\|r\|_{L^2(T)}$ under refinement. In particular, if $n=1$, then $r=f$ and the oscillation of the interior residual becomes *data oscillation*:

$$\operatorname{osc}_{\mathscr{T}}(f,T) := \|h(f-\overline{f_T})\|_{L^2(T)} \qquad \text{for all } T \in \mathscr{T}. \tag{58}$$

Note that in this case there is one additional order of convergence if $f \in H^1(\Omega)$.

The inequality (57) holds also with $\bar{r}_T$ chosen from $\mathbb{P}_{n_1}(T)$, with $n_1 \geq 1$, at the price of a larger constant hidden in $\lesssim$. We postpone the discussion of the higher order nature of the oscillation in this case after Theorem 7 below.

We conclude this paragraph by commenting on the choice of the cut-off function $\eta_T \in W^1_\infty(T)$ with (55). For example, we may take

$$\eta_T = (d+1)^{d+1} \prod_{z \in \mathscr{V} \cap T} \lambda_z, \tag{59}$$

where λ_z, $z \in \mathscr{V} \cap T$, are the barycentric coordinates of T from §1.4. This choice is due to Verfürth [57, 58]. Another choice, due to Dörfler [24], can be defined as follows: refine T such that there appears an interior node and take the corresponding Courant basis function on the virtual triangulation of T; see Fig. 10 for the 2-dimensional case.

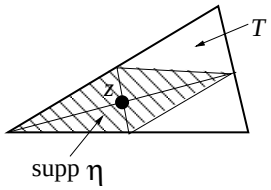


Fig. 10 Virtual refinement of a triangle for the Dörfler cut-off function.

The Dörfler cut-off function has the additional property that it is an element of a refined finite element space. This is not important here but useful when proving lower bounds for the differences of two discrete solutions; see e.g. Problem 23. Such estimates are therefore called *discrete lower bound* whereas the bound for the true error is called *continuous lower bound*.

Local Lower Bound with Jump Residual. We next strive for a local lower bound for the error in terms of the jump residual $h_S^{1/2}\|j\|_{L^2(S)}$, $S \in \mathscr{S}$, and use the local lower bound in terms of the element residual as guideline.

We first notice that $j = [\![\boldsymbol{A}\nabla U]\!]$ may not be constant on an interior side $S \in \mathscr{S}$ due to the presence of $\boldsymbol{A}$. We therefore introduce the *oscillation of the jump residual* in S,

$$h_S^{1/2}\|j - \bar{j}_S\|_{L^2(S)},$$

where $\bar{j}_S$ stands for the mean value of j on S, and write

$$h_S^{1/2}\|j\|_{L^2(S)} \leq h_S^{1/2}\|\bar{j}_S\|_{L^2(S)} + h_S^{1/2}\|j - \bar{j}_S\|_{L^2(S)}. \tag{60}$$

Notice that here the important question about the order of this oscillation is not obvious because, in contrast to the oscillation of the element residual in the case $n = 1$,

the approximate solution U is involved. We postpone the corresponding discussion until after Theorem 7 below.

To choose a counterpart of η_T, let ω_S denote the patch composed of the two elements of $\mathscr{T}$ sharing S; see Fig. 11 (left) for the 2-dimensional case. Obviously ω_S has a nonempty interior. Let $\eta_S \in W^1_\infty(\omega_S)$ be a cut-off function with the properties

$$|S| \lesssim \int_S \eta_S, \quad \operatorname{supp} \eta_S = \omega_S, \quad \|\eta_S\|_{L^\infty(\omega_S)} = 1, \quad \|\nabla \eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{-1}. \tag{61}$$

Following Verfürth [57, 58] we may take η_S given by

$$\eta_{S|T} = d^d \prod_{z \in \mathscr{V} \cap S} \lambda_z^T, \tag{62}$$

where $T \subset \omega_S$ and λ_z^T, $z \in \mathscr{V} \cap T$, are the barycentric coordinates of T. Also here

Fig. 11 Patch ω_S of triangles associated to interior side (left) and its refinement for Dörfler cut-off function (right).

Dörfler [24] proposed the following alternative: refine ω_S such that there appears an interior node of S and take the corresponding Courant basis function on the virtual triangulation of ω_S; see Fig. 11 (right) for the 2-dimensional case.

After these preparations we are ready to derive a counterpart of (57). In view of the properties of η_S, we have

$$\|\bar{j}_S\|^2_{L^2(S)} \lesssim \int_S \bar{j}_S(\bar{j}_S \eta_S) = \int_S j v_S + \int_S (\bar{j}_S - j) v_S \tag{63}$$

with $v_S = \bar{j}_S \eta_S$. We rewrite the first term on the right-hand side with the representation formula (36) as follows:

$$\int_S j v_S = -\int_{\omega_S} r v_S + \langle \mathscr{R}, v_S \rangle;$$

in contrast to (53), the jump residual couples with the element residual. Hence

$$\left| \int_{\omega_S} j v_S \right| \le \|r\|_{L^2(\omega_S)} \|v_S\|_{L^2(\omega_S)} + \|\mathscr{R}\|_{H^{-1}(\omega_S)} \|\nabla v_S\|_{L^2(\omega_S)}.$$

In view of the Poincaré-Friedrichs inequality (26), $|\omega_S| \lesssim h_S |S|$ and (61), we have

$$\|v_S\|_{L^2(\omega_S)} \lesssim h_S \|\nabla v_S\|_{L^2(\omega_S)} \le h_S \|\bar{j}_S\|_{L^2(\omega_S)} \|\nabla \eta_S\|_{L^\infty(\omega_S)} \lesssim h_S^{1/2} \|\bar{j}_S\|_{L^2(S)}.$$

We thus infer that

$$\left| \int_{\omega_S} j v_S \right| \lesssim \left(h_S^{1/2} \|r\|_{L^2(\omega_S)} + h_S^{-1/2} \|\mathscr{R}\|_{H^{-1}(\omega_S)} \right) \|\bar{j}_S\|_{L^2(S)}$$

and, using (44),

$$\left| \int_S (\bar{j}_S - j) v_S \right| \le \|\bar{j}_S - j\|_{L^2(S)} \|v_S\|_{L^2(S)} \lesssim \|\bar{j}_S - j\|_{L^2(S)} \|\bar{j}_S\|_{L^2(S)}.$$

Inserting these estimates into (63) yields

$$\|\bar{j}_S\|_{L^2(S)}^2 \lesssim \left(h_S^{1/2} \|r\|_{L^2(\omega_S)} + h_S^{-1/2} \|\mathscr{R}\|_{H^{-1}(\omega_S)} + \|\bar{j}_S - j\|_{L^2(S)} \right) \|\bar{j}_S\|_{L^2(S)}$$

whence, recalling (60),

$$h_S^{1/2} \|j\|_{L^2(S)} \lesssim \|\mathscr{R}\|_{H^{-1}(\omega_S)} + h_S \|r\|_{L^2(\omega_S)} + h_S^{1/2} \|\bar{j}_S - j\|_{L^2(S)}. \tag{64}$$

This estimate also holds if $\bar{j}_S \in \mathbb{P}_{n_2}(S)$ is a polynomial of degree $\le n_2$ (Problem 27).

Local Lower Bound with Indicator. We combine the two results on interior and jump residual and exploit also the local relationship between residual and error in order to obtain a local lower bound in terms of a single indicator.

To this end, we introduce the following notation for the oscillations. Recall the mesh-size function h from §1.5 and let

$$\bar{r} = P_{2n-2} r, \quad \bar{j} = P_{2n-1} j,$$

where $P_{2n-2} r_{|T}$ and $P_{2n-1} j_{|S}$ are the L^2-orthogonal projections of r and j onto $\mathbb{P}_{2n-2}(T)$ and $\mathbb{P}_{2n-1}(S)$, respectively. The choice of the polynomial degrees arise from the desire that the oscillations are of higher order. Details are discussed after Theorem 7. Moreover, we associate with each simplex $T \in \mathscr{T}$ the patch

$$\omega_T := \bigcup_{S \subset \partial T \setminus \partial\Omega} \omega_S$$

(see Fig. 12 for the 2-dimensional case), and define the oscillation in ω_T by

$$\operatorname{osc}_{\mathscr{T}}(U, \omega_T) = \|h(r - \bar{r})\|_{L^2(\omega_T)} + \|h^{1/2}(j - \bar{j})\|_{L^2(\partial T \setminus \partial\Omega)}. \tag{65}$$

In general, as indicated by the notation, the oscillation depends on the approximation U. However, in certain situations, it may be independent of the approximation U and then becomes *data* oscillation (58); see also Problem 19.

Theorem 7 (Local lower bound). *Let u be the exact solution of the model problem* (24) *satisfying* (25) *and* (35). *Each element indicator of* (46) *bounds, up to oscilla-*

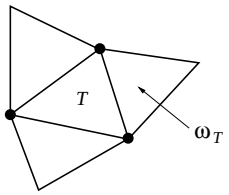


Fig. 12 Patch associated to a triangle in the local lower bound.

tion, the local error of an approximation $U \in \mathbb{V}(\mathscr{T})$ from below:

$$\mathscr{E}_{\mathscr{T}}(U,T) \lesssim \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_T)} + \operatorname{osc}_{\mathscr{T}}(U,\omega_T) \qquad \textit{for all } T \in \mathscr{T}, \tag{66}$$

where the hidden constant depends only on the shape coefficients of the simplices in ω_T, the dimension d and the polynomial degree n.

Proof. We combine (57) and (64), where $\bar{r}$ and $\bar{j}$ are piecewise polynomial of degree $2n-2$ and $2n-1$, respectively. Noting $h_S \approx h_T$ for all interior sides $S \in \mathscr{S}$ with $S \subset \partial T$ and and $\|\mathscr{R}\|_{H^{-1}(T')} \leq \|\mathscr{R}\|_{H^{-1}(\omega_T)}$ for $T' \subset T$, we thus derive

$$\mathscr{E}_{\mathscr{T}}(U,T) \lesssim \|\mathscr{R}\|_{H^{-1}(\omega_T)} + \|h(r-\bar{r})\|_{L^2(\omega_T)} + \|h^{1/2}(\bar{j}-j)\|_{L^2(\partial T \setminus \partial\Omega)}.$$

Thus, the special case

$$\|\mathscr{R}\|_{H^{-1}(\omega_T)} \leq \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_T)}$$

of (52) finishes the proof. □

A discussion of the significance of local lower bound in Theorem 7 is in order. To this end, we first consider the decay properties of the oscillation terms, which are crucial for the relevance of the aforementioned bound. Then we remark about the importance of the fact that the bound in Theorem 7 is local. Finally, in the next section, we provide a global lower bound as corollary and discuss its relationship with the upper bound in Theorem 6.

Higher Order Nature of Oscillation. In some sense the oscillation pollutes the local lower bound in Theorem 7. It is therefore important that the oscillaton is or gets small relative to the local error. We therefore compare the convergence order of the oscillation (65) with that of the local error.

To this end, let us first observe that the choices of the polynomial degrees in the oscillation allow us to derive the following upper bound of the oscillation (see Problem 29):

$$\begin{aligned}\operatorname{osc}_{\mathscr{T}}(U,\omega_T) &\lesssim \|h(f-P_{2n-2}f)\|_{L^2(\omega_T)} \\ &\quad + \Big(\|h(\operatorname{div}\boldsymbol{A} - P_{n-1}(\operatorname{div}\boldsymbol{A}))\|_{L^\infty(\omega_T)} + \|\boldsymbol{A}-P_n\boldsymbol{A}\|_{L^\infty(\omega_T)}\Big)\|\nabla U\|_{L^2(\omega_T)}.\end{aligned} \tag{67}$$

If f and $\boldsymbol{A}$ are smooth, one expects that the local error vanishes like

$$\|\nabla(u-U)\|_{L^2(T)} = \mathscr{O}(h_T^{d/2+n})$$

and, in view of (67), oscillation like

$$\operatorname{osc}_{\mathscr{T}}(U,\omega_T) = \mathscr{O}(h_T^{d/2+n+1}).$$

See also Problem 30 for a stronger result for the jump residual.

The oscillation $\operatorname{osc}_{\mathscr{T}}(U,\omega_T)$ is therefore expected to be of higher order as $h_T \downarrow 0$. However, as Problem 32 below illustrates, it may be relevant on relatively coarse triangulations $\mathscr{T}$.

Local Lower Bound and Marking. In contrast to the upper bound in Theorem 6, the lower bound in Theorem 7 is local. This is very welcome in a context of adaptivity. In fact, if $\operatorname{osc}_{\mathscr{T}}(U,\omega_T) \ll \|\nabla(u-U)\|_{L^2(\omega_T)}$, as we expect asymptotically, then (66) translates into

$$\mathscr{E}_{\mathscr{T}}(U,T) \lesssim \alpha_2 \|\nabla(u-U)\|_{L^2(\omega_T)}. \tag{68}$$

This means that an element T with relatively large error indicator contains a large portion of the error. To improve the solution U effectively, such T must be split giving rise to a procedure that tries to equidistribute errors. This is consistent with the discussion of adaptive approximation of §1.1 for $d=1$ and of §1.6 for $d>1$.

3.2 Global Lower Bound

We derive a global lower bound from the local one in Theorem 7 and discuss its relationship with the global upper bound in Theorem 6.

The global counterpart of $\operatorname{osc}_{\mathscr{T}}(U,\omega_T)$ from (65) is given by

$$\operatorname{osc}_{\mathscr{T}}(U) = \|h(r-\bar{r})\|_{L^2(\Omega)} + \|h^{1/2}(j-\bar{j})\|_{L^2(\Gamma)}, \tag{69}$$

where r is computed elementwise over $\mathscr{T}$ and Γ is the interior skeleton of $\mathscr{T}$.

Corollary 3 (Global lower bound). *Let u be the exact solution of the model problem* (24) *satisfying* (25) *and* (35). *The estimator* (47) *bounds, up to oscillation, the error of an approximation $U \in \mathbb{V}(\mathscr{T})$ from below:*

$$\mathscr{E}_{\mathscr{T}}(U) \lesssim \alpha_2 \|u-U\|_{\mathbb{V}} + \operatorname{osc}_{\mathscr{T}}(U) \tag{70}$$

where the hidden constant depends on the shape coefficient of $\mathscr{T}$, the dimension d, and the polynomial degree n.

Proof. Sum (66) over all $T \in \mathscr{T}$ and take into account that each element is contained in at most by $d+2$ patches ω_T. □

Supposing that the approximation U is the Galerkin solution (27) with (31), the upper and lower a posteriori bounds in Theorem 6 and Corollary 3 imply

$$\|u-U\|_{\mathbb{V}} \lesssim \frac{1}{\alpha_1}\mathscr{E}_{\mathscr{T}}(U) \lesssim \frac{\alpha_2}{\alpha_1}\|u-U\|_{\mathbb{V}} + \frac{1}{\alpha_1}\operatorname{osc}_{\mathscr{T}}(U). \tag{71}$$

In other words, the error and estimator are equivalent up to oscillation.

In Problem 32 we present an example for which the ratio $\|u-U\|_{\mathbb{V}}/\mathscr{E}_{\mathscr{T}}(U)$ can be made arbitrarily small. Consequently, a lower bound without pollution and a perfect equivalence of error and estimator cannot be true in general. Moreover, for that example there holds $\mathscr{E}_{\mathscr{T}}(U) = \operatorname{osc}_{\mathscr{T}}(U)$, indicating that $\operatorname{osc}_{\mathscr{T}}(U)$ is a good measure to account for the discrepancy.

We see that $\operatorname{osc}_{\mathscr{T}}(U)$ intervenes in the relationship of error and estimator and, therefore, cannot be ignored in an analysis of an adaptive algorithm using the estimator $\mathscr{E}_{\mathscr{T}}(U)$; we will come back to this in §7. The case of data oscillation will be simpler than the general case in which $\operatorname{osc}_{\mathscr{T}}(U)$ depends on the approximation U; the latter dependence creates a nonlinear interaction in the adaptive algorithm.

The presence of oscillation is also consistent with our previous comparison of local dual norms and scaled integral norms. Since we invoked scaled integral norm in order to have an (almost) computable upper bound, this suggests that, at least for standard a posteriori error estimation, oscillation is a price that we have to pay for computability.

Fortunately, as we have illustrated in §3.1, oscillation is typically of higher order and then the a posteriori upper bound in Theorem 6 is asymptotically sharp in that its decay rate coincides with the one of the error, as the a priori bound of Theorem 5. Notice however the lower bound in Corollary 3 provides information beyond asymptotics: for example, if we consider the linear finite element method, that is $n = 1$, then $\operatorname{osc}_{\mathscr{T}}(U)$ vanishes for all triangulations on which f and $\boldsymbol{A}$ are piecewise constant and in this class of meshes error and estimator are thus equivalent:

$$\|\nabla(u-U)\|_{L^2(\Omega)} \approx \mathscr{E}_{\mathscr{T}}(U).$$

In summary: the estimator $\mathscr{E}_{\mathscr{T}}(U)$ from (47) is computable, it may be used to quantify the error and, in view of the local properties in §3.1, its indicators may be employed to provide the problem-specific information for local refinement.

3.3 Notes

Local lower bounds first appear in the work of Babuška and Miller [4]. Their derivation with continuous bubble functions is due to Verfürth [57], while the discrete lower bounds are due to Dörfler [24].

The discussion of the relationship between local dual norms and scaled integral norms as the reason for oscillation is an elaborated version of Sacchi-Veeser's one [47, Remark 3.1]. It is worth mentioning that there the indicators associated with

the approximation of the Dirichlet boundary values do not need to invoke scaled Lebesgue norms and are overestimation-free. Binev, Dahmen and DeVore [7] and Stevenson [52] arrange the a posteriori analysis such that oscillation is measured in $H^{-1}(\Omega)$. This avoids overestimation but brings back the question how to (approximately) evaluate the $H^{-1}(\Omega)$-norm at acceptable cost. This question is open.

One may think that the issue of oscillation is specific to standard a posteriori error estimation. However all estimators we are aware of suffer from oscillations of the data that are finer than the mesh-size. For example, in the case of hierarchical estimators $\eta_{\mathscr{T}}(U)$ [2, 55, 58], as well as those based upon local discrete problems [2, 12, 42] or on gradient recovery [2, 27], the oscillation arises in the upper but not in the lower bounds and so creates a similar gap as that discussed here, namely

$$\eta_{\mathscr{T}}(U) \lesssim \|\nabla(u-U)\|_{L^2(\Omega)} \lesssim \eta_{\mathscr{T}}(U) + \operatorname{osc}_{\mathscr{T}}(U). \tag{72}$$

3.4 Problems

Problem 19 (Data oscillation). Check that $\operatorname{osc}_{\mathscr{T}}(U,\omega_T)$ in (65) does not depend on the approximation U if U is piecewise affine and $\boldsymbol{A}$ is piecewise constant, and is given by

$$\operatorname{osc}(f,\omega_T) = \|h(f-\overline{f})\|_{L^2(\omega_T)},$$

which corresponds to element data oscillation in (58).

Problem 20 (Energy norm case). Consider model problem (24) and discretization (27) with $\mathbb{S} = \mathbb{V}(\mathscr{T})$ and $\boldsymbol{A}$ symmetric. Derive the counterparts of (66) and (71) for the energy norm and discuss the difference to the case presented here.

Problem 21 (Cut-off functions for simplices). Verify that a suitable multiple of the Verfürth cut-off function (59) satisfies the properties (55). To this end, exploit affine equivalence of T to a fixed reference simplex and shape regularity. Repeat for the corresponding Dörfler cut-off function.

Problem 22 (Inverse estimate for general polynomials). (Try this problem after Problem 21.) Show that the choice (59) for η_T verifies, for all $p \in \mathbb{P}_n(T)$,

$$\int_T p^2 \lesssim \int_T p^2 \eta_T, \qquad \|\nabla(p\eta_T)\|_{L^2(T)} \lesssim h_T^{-1}\|p\|_{L^2(T)}$$

with constants depending on n and the shape coefficient of T. To this end, recall the equivalence of norms in finite-dimensional spaces. Derive the estimate

$$h_T \|r\|_{L^2(T)} \lesssim \|r\|_{H^{-1}(T)}$$

for $r \in \mathbb{P}_n(T)$.

Problem 23 (Lower bound for correction). Consider the model problem and its discretization for $d = 2$ and $n = 1$. Let U_1 be the solution over a triangulation $\mathscr{T}_1$

and U_2 the solution over $\mathscr{T}_2$, where $\mathscr{T}_2$ has been obtained by applying at least 3 bisections to every triangle of $\mathscr{T}_1$. Moreover, suppose that f is piecewise constant over $\mathscr{T}_1$. Show

$$\|\nabla(U_2-U_1)\|_{L^2(\Omega)} \geq \|h_1 f\|_{L^2(\Omega)},$$

where h_1 is the mesh-size function of $\mathscr{T}_1$.

Problem 24 (Cut-off functions for sides). Verify that a suitable multiple of the Verfürth cut-off function (62) satisfies the properties (61). Repeat for the corresponding Dörfler cut-off function.

Problem 25 (Polynomial extension). Let S be a side of a simplex T. Show that for each $q \in \mathbb{P}_n(S)$ there exists a $p \in \mathbb{P}_n(T)$ such that

$$p = q \text{ on } S \qquad \text{and} \qquad \|p\|_{L^2(T)} \lesssim h_T^{1/2}\|q\|_{L^2(S)}.$$

Problem 26 (Norm equivalences with cut-off functions of sides). Let S be a side of a simplex T. Show that the choice (62) for η_S verifies, for all $q \in \mathbb{P}_n(S)$ and all $p \in \mathbb{P}_m(T)$,

$$\int_S q^2 \lesssim \int_S q^2\eta_S, \qquad \|\nabla(p\eta_S)\|_{L^2(T)} \lesssim h_T^{-1}\|p\|_{L^2(T)}$$

with constants depending on m, n, and the shape coefficient of T.

Problem 27 (Lower bound with jump residual and general oscillation). Exploit the claims in Problems 25 and 26, to rederive the estimate (64) but this time with $\overline{r}$ and $\overline{j}$ piecewise polynomials of degree $\leq n_1$ and n_2.

Problem 28 (Best approximation of a product). Let K be either a d or a $(d-1)$-simplex. For $\ell \in \mathbb{N}$ denote by $P_m^p\colon L^p(K,\mathbb{R}^\ell) \to \mathbb{P}_m(K,\mathbb{R}^\ell)$ the operator of best L^p-approximation in K. Prove that, for all $v \in L^\infty(K,\mathbb{R}^\ell)$, $V \in \mathbb{P}_n(K,\mathbb{R}^\ell)$ and $m \geq n$,

$$\|vV - P_m^2(vV)\|_{L^2(K)} \leq \|v - P_{m-n}^\infty v\|_{L^\infty(K)}\|V\|_{L^2(K)}.$$

Problem 29 (Upper bound for oscillation). Verify the upper bound (67) for the oscillation by exploiting Problem 28.

Problem 30 (Superconvergence of jump residual oscillation). Show that if $\boldsymbol{A}$ is smooth across interelement boundaries, then the oscillation of the jump residual is superconvergent in that

$$\|j - \overline{j}_S\|_{L^2(S)} = \mathscr{O}(h_S^n)\|j\|_{L^2(S)} \qquad \text{as } h_S \searrow 0.$$

Problem 31 (Simplified bound of oscillation). Using (67), show that (35) implies

$$\operatorname{osc}_{\mathscr{T}}(U,\omega_T) \lesssim h_T\Big(\|f\|_{L^2(\omega_T)} + \|\nabla U\|_{L^2(\omega_T)}\Big), \tag{73}$$

where the hidden constant depends also on $\boldsymbol{A}$.

Problem 32 (Necessity of oscillation). Let $\varepsilon = 2^{-K}$ for K integer and extend the function $\frac{1}{2}x(\varepsilon - |x|)$ defined on $(-\varepsilon, \varepsilon)$ to a 2ε-periodic C^1 function u_ε on $\Omega = (-1,1)$. Moreover, let the forcing function be $f_\varepsilon = -u''$, which is 2ε-periodic and piecewise constant with values ± 1 that change at multiples of ε; see Fig. 13. Let $\mathscr{T}_\varepsilon$ be a uniform mesh with mesh-size $h = 2^{-k}$, with $k \ll K$. We consider piece-

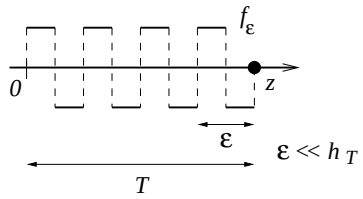


Fig. 13 An strongly oscillating forcing function.

wise linear finite elements $\mathbb{V}(\mathscr{T}_\varepsilon)$ and corresponding Galerkin solution $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$. Observing that f_ε is $L^2(\Omega)$-orthogonal to both the space of piecewise constants and linears over $\mathscr{T}_\varepsilon$, show that

$$\begin{aligned}\|u_\varepsilon' - U_\varepsilon'\|_{L^2(\Omega)} = \|u_\varepsilon'\|_{L^2(\Omega)} &= \frac{\varepsilon}{\sqrt{6}} = \frac{2^{-K}}{\sqrt{6}} \\ &\ll 2^{-k} = h = \|h f_\varepsilon\|_{L^2(\Omega)} = \operatorname{osc}_{\mathscr{T}_\varepsilon}(U_\varepsilon) = \mathscr{E}_{\mathscr{T}_\varepsilon}(U_\varepsilon).\end{aligned}$$

Extend this 1d example via a checkerboard pattern to any dimension.

4 Convergence of AFEM

The purpose of this section is to formulate an adaptive finite element method (AFEM) and to prove that it generates a sequence of approximate solutions converging to the exact one. The method consists in the following main steps:

$$\mathsf{SOLVE} \to \mathsf{ESTIMATE} \to \mathsf{MARK} \to \mathsf{REFINE}.$$

By their nature, adaptive algorithms define the sequence of approximate solutions as well as associated meshes and spaces only implicitly. This fact requires an approach that differs from 'classical' convergence proofs. In particular, a proof of convergence will hinge on results of an a posteriori analysis as in §§2.4 and 3.

The approach presented in this section covers wide classes of problems, discrete spaces, estimators and marking strategies. Here we do not strive for such generality but instead, in order to minimize technicalities, illustrate the main arguments only in a model case and then hint on possible generalizations.

It is worth noticing that, conceptually, the following convergence proof does not suppose any additional regularity of the exact solution. Consequently, it does not (and cannot) provide any information about the convergence speed. This important issue will be the concern of §7 for smaller classes of problems and algorithms.

4.1 A Model Adaptive Algorithm

We first present an AFEM for the model problem (24), which is an example of a *standard* iterative process that is often used in practice. In §4.2 we then prove its convergence and, finally, in §4.3 we comment on generalizations still covered by the given approach.

AFEM. The main structure of the adaptive finite element method is as follows: given an initial grid $\mathscr{T}_0$, set $k = 0$ and iterate

1. $U_k = \mathsf{SOLVE}(\mathscr{T}_k)$;
2. $\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k} = \mathsf{ESTIMATE}(U_k,\mathscr{T}_k)$;
3. $\mathscr{M}_k = \mathsf{MARK}\big(\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k},\mathscr{T}_k\big)$;
4. $\mathscr{T}_{k+1} = \mathsf{REFINE}(\mathscr{M}_k,\mathscr{T}_k)$; $k \leftarrow k+1$.

Thus, the algorithm produces sequences $(\mathscr{T}_k)_{k=0}^{\infty}$ of meshes, $(U_k)_{k=0}^{\infty}$ of approximate solutions, and, implicitly, $(\mathbb{V}_k)_{k=0}^{\infty}$ of discrete spaces.

We next state our main assumptions and define the aforementioned modules for the problem at hand in detail.

Assumptions on continuous problem. We assume that the model problem (24) satisfies (25) and (35) so that the a posteriori error bounds of §§2 and 3 are available.

Initial grid. Assume that $\mathscr{T}_0$ is some initial triangulation of Ω such that $\boldsymbol{A}$ is piecewise Lipschitz over $\mathscr{T}_0$.

Solve. Let

$$\mathbb{V}_k := \mathbb{V}(\mathscr{T}_k) := \left\{ V \in \mathbb{S}^{n,0}(\mathscr{T}_k) \mid V_{|\partial\Omega} = 0 \right\}$$

be the space of continuous functions that are piecewise polynomial of degree $\leq n$ over $\mathscr{T}_k$, and compute the Galerkin solution U_k in $\mathbb{V}_k$ given by

$$U_k \in \mathbb{V}_k : \qquad \int_\Omega \boldsymbol{A}\nabla U_k \cdot \nabla V = \int_\Omega fV \qquad \text{for all } V \in \mathbb{V}_k.$$

Estimate. Compute the error estimator $\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k}$ given by

$$\mathscr{E}_k(U_k,T) := \left(h_T^2 \|r\|^2_{L^2(T)} + h_T \|j\|^2_{L^2(\partial T\setminus\partial\Omega)} \right)^{1/2}$$

where $h_T = |T|^{1/d}$, r and j are the element and jump residuals from (37) associated to the approximate solution U_k.

Mark. Collect a subset $\mathscr{M}_k \subset \mathscr{T}_k$ of marked elements with the following property:

$$\forall T \in \mathscr{T}_k \quad \mathscr{E}_k(U_k,T) = \mathscr{E}_{k,\max} > 0 \quad \Longrightarrow \quad T \in \mathscr{M}_k \tag{74}$$

with $\mathscr{E}_{k,\max} := \max_{T\in\mathscr{T}_k} \mathscr{E}_k(U_k,T)$.

Refine. Refine $\mathscr{T}_k$ into $\mathscr{T}_{k+1}$ using bisection, as explained in §1.3, in such a way that each element in $\mathscr{M}_k$ is bisected at least once and, finally, increment k.

Classical convergence proofs consider the case of uniform, or 'non-adaptive', refinement, which is included in the above class of algorithms by choosing $\mathscr{M}_k = \mathscr{T}_k$, thereby ignoring the information provided by the estimator. These convergence proofs rely on the fact that the maximum mesh-size decreases to 0 and therefore $\overline{\cup_{k=0}^\infty \mathbb{V}_k} = H^1_0(\Omega)$. The above algorithm does not require this property, neither explicity nor implicitly in general. In fact this property is not desirable in an adaptive context, since (30) reveals that it is sufficient to approximate only one function of $H^1_0(\Omega)$, namely the exact solution u of (24). In the next section we see that the above alogrithm ensures just convergence to u by a subtle combination of properties of estimator and marking strategy.

4.2 Convergence

The goal of this section is to prove the convergence of the AFEM in §4.1. More precisely, we show that the sequence $(U_k)_{k=0}^\infty$ of approximate solutions converges to the exact solution u of the model problem (24).

Throughout this section '$\lesssim$' stands for '$\leq C$', where the constant is independent of the iteration number k in the adaptive algorithm.

Convergence to Some Function. We expect the Galerkin solutions $(U_k)_{k=0}^\infty$ to approximate the exact solution u in $\mathbb{V} = H_0^1(\Omega)$. In any event, we may regard them as approximations to the Galerkin solution U_∞ in the limit

$$\mathbb{V}_\infty := \overline{\bigcup_{k=0}^{\infty} \mathbb{V}_k}$$

of the discrete spaces. Notice that $\mathbb{V}_\infty$ is a subspace of $\mathbb{V}$, which may not coincide with $\mathbb{V}$ (see below). In the next lemma we adopt this viewpoint and show that $(U_k)_{k=0}^\infty$ converges to U_∞.

Lemma 7 (Limit of approximate solutions). *The finite element solutions $(U_k)_{k=0}^\infty$ converge in $\mathbb{V}$ to the Galerkin solution $U_\infty \in \mathbb{V}_\infty$ given by*

$$\int_\Omega \boldsymbol{A}\nabla U_\infty \cdot \nabla V = \int_\Omega fV \qquad \text{for all } V \in \mathbb{V}_\infty.$$

Proof. Since the sequence of $(\mathbb{V}_k)_{k=0}^\infty$ is nested (see Problem 33), the set $\mathbb{V}_\infty$ is a closed linear subspace of $\mathbb{V}$. Hence $\mathbb{V}_\infty$ is a Hilbert space and the bilinear form $\mathscr{B}$ is coercive and continuous on $\mathbb{V}_\infty$. The Lax-Milgram Theorem therefore ensures existence and uniqueness of U_∞.

Let $k \in \mathbb{N}_0$ and note that $\mathbb{V}_k \subset \mathbb{V}_\infty$. We can therefore replace $\mathbb{V}$ by $\mathbb{V}_\infty$ in Theorem 4 and obtain

$$\|U_\infty - U_k\|_{\mathbb{V}} \le \inf_{V \in \mathbb{V}_k} \|U_\infty - V\|_{\mathbb{V}}.$$

Sending $k \to \infty$ then finishes the proof, because the right-hand side decreases to 0 by the very definition of $\mathbb{V}_\infty$. □

Lemma 7 reduces our task to showing that $U_\infty = u$. Notice that this is equivalent to the condition $u \in \mathbb{V}_\infty$, illustrating that in general there is no need for $\mathbb{V}_\infty = \mathbb{V}$.

The identity $U_\infty = u$ hinges on the design of the adaptive algorithm. To illustrate this point, let us consider two extreme examples:

- It may happen that all indicators vanish in iteration k^*. Then $\mathscr{E}_{k^*,\max} = 0$ and (74) is compatible with $\mathscr{M}_k = \emptyset$ and $\mathbb{V}_\infty = \mathbb{V}_k$ for all $k \ge k^*$. In this case, $U_\infty = U_{k^*}$ and convergence is only ensured if a vanishing estimator implies a vanishing error. The latter is given in particular if the estimator bounds the error from above.
- It may happen that only the simplices containing a fixed point are bisected in each iteration, but the exact solution u has a more complex structure so that $u \notin \mathbb{V}_\infty$. Since $u \ne U_\infty$, and uniform refinement is not enforced, the adaptive procedure must depend on the unknown function u.

Convergence therefore will require that the module ESTIMATE extracts enough relevant information about the error, the module MARK uses this information correctly, and the module REFINE reduces the mesh-size where requested.

Mesh-Size before Bisection. The module REFINE bisects elements and so halves their volume. This implies the following useful property of elements to be bisected, which include the marked elements.

Lemma 8 (Sequences of elements to be bisected). *For any sequence $(T_k)_{k=0}^{\infty}$ of elements with $T_k \in \mathscr{T}_k \setminus \mathscr{T}_{k+1}$ there holds* $\lim_{k\to\infty} |T_k| = 0$.

Proof. Suppose that $\limsup_{k\to\infty} |T_k| \geq c > 0$, that is there exists a infinite subsequence $(T_{k_\ell})_\ell$ such that $\lim_{\ell\to\infty} |T_{k_\ell}| \geq c$. Recall that the children of a bisection have half the volume of the parent. Consequently, only a finite number of children of any generation of each T_{k_ℓ} can appear in the sequence $(T_{k_\ell})_\ell$. Eliminating inductively these children, we obtain an infinite sequence of simplices with disjoint interiors and volume greater than $c > 0$. This however contradicts the boundedness of Ω, whence $\limsup_{k\to\infty} |T_k| \leq 0$, which is equivalent to the assertion. □

It is instructive and convenient to reformulate Lemma 8 in terms of mesh-size functions.

Lemma 9 (Mesh-size of elements to be bisected). *If χ_k denotes the characteristic function of the union $\cup_{T\in\mathscr{T}_k\setminus\mathscr{T}_{k+1}} T$ of elements to be bisected and h_k is the mesh-size function of $\mathscr{T}_k$, then*

$$\lim_{k\to\infty} \|h_k \chi_k\|_{L^\infty(\Omega)} = 0$$

Proof. We may assume that $\mathscr{T}_k \setminus \mathscr{T}_{k+1} \neq \emptyset$ for all $k \in \mathbb{N}_0$ without loss of generality. Choose $(T_k)_{k=0}^{\infty}$ such that $T_k \in \mathscr{T}_k \setminus \mathscr{T}_{k+1}$ and $h_{T_k} = \max_{T\in\mathscr{T}_k\setminus\mathscr{T}_{k+1}} h_T$ and, recalling that $h_T = |T|^{1/d}$, use Lemma 8 to deduce the assertion. □

Lemma 9 may be viewed as a generalization of $\lim_{k\to\infty} \|h_k\|_{L^\infty(\Omega)} = 0$ in the case of uniform refinement. It may be proven also by invoking the limiting mesh-size h_∞; see Problems 34 and 35. The limiting mesh-size describes the local structure of $\mathbb{V}_\infty$ and may differ from the zero function.

Convergence to Exact Solution. In order to achieve $U_\infty = u$, we may investigate the residual of U_∞, which is related to the residuals of the finite element solutions U_k. The latter are in turn controlled by the element indicators $\mathscr{E}_k(U_k, T)$, $T \in \mathscr{T}_k$, which are employed in the step MARK. The fact that indicators with maximum value are marked yields the following property of the largest element indicator $\mathscr{E}_{k,\max}$.

Lemma 10 (Convergence of maximum indicator). *There holds*

$$\lim_{k\to\infty} \mathscr{E}_{k,max} = 0.$$

Proof. We may assume that $\mathscr{M}_k \neq \emptyset$ for all $k \in \mathbb{N}_0$ without loss of generality. Choose a sequence $(T_k)_{k=0}^{\infty}$ of elements such that $T_k \in \mathscr{T}_k$ and $\mathscr{E}_k(U_k, T_k) = \mathscr{E}_{k,\max}$. Thanks to (74), we have $T_k \in \mathscr{M}_k$ and so Lemma 8 and module REFINE yield $\lim_{k\to\infty} |T_k| = 0$. Exploiting the local lower bound in Theorem 7 and the simplified upper bound for the local oscillation (73), we derive the following estimate for any indicator for $T \in \mathscr{T}_k$:

$$\begin{aligned} \mathscr{E}_k(U_k, T) \lesssim \|\nabla(U_k - U_\infty)\|_{L^2(\omega_T)} + \|\nabla(U_\infty - u)\|_{L^2(\omega_T)} \\ + h_T \Big(\|f\|_{L^2(\omega_T)} + \|\nabla U_k\|_{L^2(\omega_T)} \Big). \end{aligned} \tag{75}$$

Taking $T = T_k$, we obtain

$$\mathscr{E}_{k,\max} = \mathscr{E}_k(U_k, T_k) \lesssim \|U_k - U_\infty\|_{\mathbb{V}} + \|\nabla(U_\infty - u)\|_{L^2(\omega_k)} + |T_k|^{1/d}\left(\|f\|_{L^2(\omega_k)} + \|U_k\|_{\mathbb{V}}\right)$$

with $\omega_k := \omega_{T_k}$. Consequently, Lemma 7 and $\lim_{k\to\infty}|T_k| = 0$, which also entails $\lim_{k\to\infty}|\omega_k| = 0$, prove the assertion. □

With these preparations we are ready for the first main result of this section.

Theorem 8 (Convergence of approximate solutions). *Let u be the exact solution of the model problem* (24) *satisfying* (25) *and* (35)*. The finite element solutions $(U_k)_{k=0}^\infty$ of the AFEM of §4.1 converge to the exact one in $\mathbb{V}$:*

$$U_k \to u \text{ in } \mathbb{V} \text{ as } k \to \infty.$$

Proof. 1 In view of Lemma 7, it remains to show that $U_\infty = u$. This is equivalent to

$$0 = \langle \mathscr{R}_\infty, v\rangle := \int_\Omega fv - \int_\Omega A\nabla U_\infty \cdot \nabla v \qquad \text{for all } v \in \mathbb{V} = H_0^1(\Omega). \tag{76}$$

Here we can take the test functions from $C_0^\infty(\Omega)$, because $C_0^\infty(\Omega)$ is a dense subset of the Hilbert space $H_0^1(\Omega)$. Lemma 7 therefore ensures that (76) follows from

$$0 = \lim_{k\to\infty}\langle \mathscr{R}_k, \varphi\rangle \quad \forall \varphi \in C_0^\infty(\Omega), \tag{77}$$

where $\mathscr{R}_k \in \mathbb{V}^*$ is the residual of U_k given by

$$\langle \mathscr{R}_k, v\rangle := \int_\Omega fv - \int_\Omega A\nabla U_k \cdot \nabla v.$$

2 In order to show (77), let $\varphi \in C_0^\infty(\Omega)$ and introduce the set

$$\mathscr{T}_\ell^* := \bigcap_{m\geq\ell} \mathscr{T}_m$$

of elements in $\mathscr{T}_\ell$ that will no longer be bisected; note that if $\mathscr{T}_\ell^* \neq \emptyset$, then $\mathbb{V} \neq \mathbb{V}_\infty$. Given $\ell \leq k$, $\mathbb{V}_\ell \subset \mathbb{V}_k$ and (51) imply

$$\langle \mathscr{R}_k, \varphi\rangle = \langle \mathscr{R}_k, \varphi - I_\ell\varphi\rangle \lesssim S_{\ell,k} + S_{\ell,k}^*, \tag{78}$$

where we expect that

$$S_{\ell,k} := \sum_{T\in\mathscr{T}_k\setminus\mathscr{T}_\ell^*} \mathscr{E}_k(U_k, T)\|\nabla(\varphi - I_\ell\varphi)\|_{L^2(N_k(T))}$$

gets small because of decreasing mesh-size whereas

$$S^*_{\ell,k} := \sum_{T\in\mathscr{T}^*_\ell} \mathscr{E}_k(U_k,T)\|\nabla(\varphi - I_\ell\varphi)\|_{L^2(N_k(T))}$$

gets small because of properties of the adaptive algorithm.
3 We first deal with $S_{\ell,k}$. The Cauchy-Schwarz inequality in some $\mathbb{R}^N$ yields

$$S_{\ell,k} \le \mathscr{E}_k(U_k,\mathscr{T}_k\setminus\mathscr{T}^*_\ell)\Big(\sum_{T\in\mathscr{T}_k\setminus\mathscr{T}^*_\ell}\|\nabla(\varphi - I_\ell\varphi)\|^2_{L^2(N_k(T))}\Big)^{1/2},$$

where the first factor

$$\begin{aligned}\mathscr{E}_k(U_k,\mathscr{T}_k\setminus\mathscr{T}^*_\ell) &\lesssim \|U_k - U_\infty\|_{\mathbb{V}} + \|U_\infty - u\|_{\mathbb{V}}\\ &\quad + \|h_k\chi_\ell\|_{L^\infty(\Omega)}\big(\|f\|_{L^2(\Omega)} + \|U_k\|_{\mathbb{V}}\big) \lesssim 1\end{aligned} \tag{79}$$

is uniformly bounded thanks to (75) and the second factor satisfies

$$\begin{aligned}\Big(\sum_{T\in\mathscr{T}_k\setminus\mathscr{T}^*_\ell}\|\nabla(\varphi - I_\ell\varphi)\|^2_{L^2(N_k(T))}\Big)^{1/2} &\lesssim \Big(\sum_{T\in\mathscr{T}_\ell\setminus\mathscr{T}^*_\ell}\|\nabla(\varphi - I_\ell\varphi)\|^2_{L^2(N_l(T))}\Big)^{1/2}\\ &\lesssim \|h_\ell\chi_\ell\|^n_{L^\infty(\Omega)}\|D^{n+1}\varphi\|_{L^2(\Omega)}\end{aligned}$$

because of $\mathscr{T}_k \ge \mathscr{T}_\ell$ and Proposition 2. Hence Lemma 9 implies

$$S_{\ell,k} \to 0 \text{ as } \ell\to\infty \text{ uniformly in } k. \tag{80}$$

4 Next, we deal with $S^*_{\ell,k}$. Here the Cauchy-Schwarz inequality yields

$$S^*_{\ell,k} \le \mathscr{E}_k(U_k,\mathscr{T}^*_\ell)\Big(\sum_{T\in\mathscr{T}^*_\ell}\|\nabla(\varphi - I_\ell\varphi)\|^2_{L^2(N_k(T))}\Big)^{1/2},$$

where the first factor satisfies

$$\mathscr{E}_k(U_k,\mathscr{T}^*_\ell) \le \#\mathscr{T}_\ell\ \mathscr{E}_{k,\max} \tag{81}$$

and the second factor

$$\Big(\sum_{T\in\mathscr{T}^*_\ell}\|\nabla(\varphi - I_\ell\varphi)\|^2_{L^2(N_k(T))}\Big)^{1/2} \lesssim \|h_\ell\|^n_{L^\infty(\Omega)}\|D^{n+1}\varphi\|_{L^2(\Omega)} \lesssim 1$$

is uniformly bounded. Lemma 10 therefore implies

$$S^*_{\ell,k} \to 0 \text{ as } k\to\infty \text{ for any fixed } \ell. \tag{82}$$

5 Given $\varepsilon > 0$, we exploit (80) and (82) by first choosing ℓ so that $S_{\ell,k} \le \varepsilon/2$ and next $k \ge \ell$ so that $S^*_{\ell,k} \le \varepsilon/2$. Inserting this into (78) yields the desired convergence (77) and finishes the proof. □

Convergence of Estimator. Theorem 8 ensures convergence of the finite element solutions U_k but says nothing about the behavior of the estimators

$$\mathscr{E}_k(U_k) = \Big(\sum_{T\in\mathscr{T}_k} \mathscr{E}_k(U_k,T)^2 \Big)^{1/2},$$

which enables one to monitor that convergence. The convergence of the estimators is ensured by the following theorem. Notice that this is not a simple consequence of Theorem 8 and Corollary 3 because of the presence of the oscillation $\operatorname{osc}_k(U_k)$ in the global lower bound; see also Problem 36.

Corollary 4 (Estimator convergence). *Assume again that the model problem* (24) *satisfies* (25) *and* (35)*. The estimators* $(\mathscr{E}_k(U_k))_{k=0}^\infty$ *of AFEM in* §*4.1 converge to* 0*:*

$$\lim_{k\to\infty} \mathscr{E}_k(U_k) = 0.$$

Proof. Theorem 8 implies $U_\infty = u$. Using this and $h_k \le h_\ell$ for $\ell \le k$, along with $\|U_k - U_\infty\|_{\mathbb{V}} \lesssim \|U_\ell - U_\infty\|_{\mathbb{V}}$, after the first inequality of (79) yields

$$\mathscr{E}_k(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^*) \to 0 \text{ as } \ell\to\infty \text{ uniformly in } k \tag{83}$$

with the help of Lemmas 7 and 9. In view of

$$\mathscr{E}_k^2(U_k) = \mathscr{E}_k^2(U_k, \mathscr{T}_k \setminus \mathscr{T}_\ell^*) + \mathscr{E}_k^2(U_k, \mathscr{T}_\ell^*),$$

we realize that (81), (83), and Lemma 10 complete the proof. □

We conclude this section with a few remarks about variants of Theorem 8 and Corollary 4 for general estimators. Theorem 8 holds for any estimator that provides an upper bound of the form

$$|\langle \mathscr{R}_k, v\rangle| \lesssim \sum_{T\in\mathscr{T}_k} \mathscr{E}_k(U_k,T) \|\nabla v\|_{L^2(N_k(T))} \quad \text{for all } v\in\mathbb{V}, \tag{84}$$

which is locally stable in the sense

$$\mathscr{E}_k(U_k,T) \lesssim h_T \|f\|_{L^2(\omega_T)} + \|\nabla U_k\|_{L^2(\omega_T)} \quad \text{for all } T\in\mathscr{T}_k; \tag{85}$$

see Problem 37. While the first assumption (84) appears natural, and is in fact crucial in view of the first example after Lemma 7, the second assumption (85) may appear artificial. However, Problem 38 reveals that is also crucial and, thus, the suggested variant of Theorem 8 is 'sharp'. Problem 39 proposes the construction of an estimator verifying the two assumptions (84) and (85) which, however, does not decrease to 0. On the other hand, Corollary 4 hinges on the local lower bound (75), which is a sort of minimal requirement of efficiency if the finite element solutions U_k converge. Roughly speaking, convergence of U_k relies on reliability and stability of the estimator, while the convergence of the estimator depends on the efficiency of the estimator. This shows that the assumptions on the estimator for Theorem 8 and

Corollary 4 are of different nature. In particular, we see that convergence of U_k can be achieved even with estimators that are too poor to quantify the error.

4.3 Notes

The convergence proof in §4.2 is a simplified version of Siebert [49], which unifies the work of Morin, Siebert, and Veeser [44] with the standard a priori convergence theory based on (global) density. In order to further discuss the underlying assumptions of the approach in 4.2, we now compare these two works in more detail.

Solve. Both works [44] and [49] consider well-posed linear problems and invoke a generalization of Lemma 7 that follows from a discrete inf-sup condition on the discretization. The latter assumption appears natural since it is necessary for convergence in the particular case of uniform refinement; see [10, Problem 3.9]. In the case of a problem with potential or 'energy', the explicit construction of U_∞ can be replaced by a convergent sequence of approximate energy minima. Examples are the convergence analyses for the p-Laplacian by Veeser [55] and for the obstacle problem by Siebert and Veeser [51], which are the first steps in the terrain of nonlinear and nonsmooth problems and are predecessors of [44] and [49].

Estimate and Mark. Paper [44] differs from [49] on the assumptions on estimators and marking strategy. More precisely, [44] assumes that the estimator provides a discrete local lower bound and that the marking strategy essentially ensures

$$\mathscr{E}_k(U_k,T) \leq \Big(\sum_{T' \in \mathscr{M}_k} \mathscr{E}_k(U_k,T')^2 \Big)^{1/2} \qquad \text{for all } T \in \mathscr{T}_k \setminus \mathscr{M}_k, \tag{86}$$

whereas [49] essentially assumes (84), (85), and (74). Thus, the assumptions on the estimator are weaker in [49], while those on the marking strategy are weaker in [44]; see also Problem 40. Since both works verify that their assumptions on the marking strategy are necessary, this shows that (minimal) assumptions on the estimator and marking strategy are coupled.

Refine. Both [44, 49] consider the same framework for REFINE. This does not only include bisection for conforming meshes (see §1.3), but also nonconforming meshes (see §1.7) and other manners of subdividing elements. Moreover, [44, 49] assume the minimal requirement of subdividing the marked elements, as in §4.1.

Further Variants and Generalizations. These approaches can be further developed in several directions:

- Morin, Siebert, and Veeser [43] give a proof of convergence of a variant of the AFEM in §4.1 when the estimator provides upper and local lower bounds for the error in 'weak' norms, e.g. similar to the L^2-norm in Problem 18.
- Demlow [20] proves convergence of a variant of the AFEM in §4.1 with estimators for local energy norm errors.

- Garau, Morin, and Zuppa [28] show convergence of a variant of the AFEM in §4.1 for symmetric eigenvalue problems.
- Holst, Tsogtgerel, and Zhu [31] extend [44] to nonlinear partial differential equations, the linearization of which are well-posed.

4.4 Problems

Problem 33 (Nesting of spaces). Let $\mathscr{T}_1$ and $\mathscr{T}_2$ be triangulations such that $\mathscr{T}_1 \leq \mathscr{T}_2$, that is $\mathscr{T}_2$ is a refinement by bisection of $\mathscr{T}_1$. Show that the corresponding Lagrange finite element spaces from (31) are nested, i. e., $\mathbb{V}(\mathscr{T}_1) \subset \mathbb{V}(\mathscr{T}_2)$.

Problem 34 (Limiting mesh-size function). Prove that there exists a limiting mesh-size function $h_\infty \in L^\infty(\Omega)$ such that

$$\|h_k - h_\infty\|_{L^\infty(\Omega)} \to 0 \text{ as } k \to \infty.$$

Can you construct an example with $h_\infty \neq 0$?

Problem 35 (Alternative proof of Lemma 9). For any iteration k, let χ_k be the characteristic function of the union $\cup_{T \in \mathscr{T}_k \setminus \mathscr{T}_{k+1}} T$ of elements to be bisected and h_k the mesh-size function of $\mathscr{T}_k$. Show

$$\lim_{k \to \infty} \|h_k \chi_k\|_{L^\infty(\Omega)} = 0$$

by means of Problem 34 and the fact that bisection reduces the mesh-size.

Problem 36 (Persistence of oscillation). Choosing appropriately the data of the model problem (24), provide an example where the exact solution is (locally) piecewise affine and the (local) oscillation does not vanish.

Problem 37 (Convergence for general estimators). Check that Lemma 10 and Theorem 8 hold for any estimator $\{\mathscr{E}_k(U_k, T)\}_{T \in \mathscr{T}_k}$ that is reliable in the sense of (84) and locally stable in the sense of (85).

Problem 38 ('Necessity' of local estimator stability). Construct an estimator that satisfies (84) and its indicators are always largest around a fixed point, entailing that (74) is compatible with refinement only around that fixed point, irrespective of the exact solution u.

Problem 39 (No estimator convergence). Assuming that the exact solution u of the model problem (24) does not vanish, construct an estimator satisfying (84) and (85) which does not decrease to 0.

Problem 40 (Assumptions for marking strategies). Check that (86) is weaker that (74) by considering the bulk-chasing strategy (90).

5 Contraction Property of AFEM

This section discusses the contraction property for a special AFEM for the *model problem* (23), which we rewrite for convenience:

$$\begin{aligned} -\operatorname{div}(\boldsymbol{A}(x)\nabla u) &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{87}$$

with piecewise smooth coefficient matrix $\boldsymbol{A}$ on $\mathscr{T}_0$. The matrix $\boldsymbol{A}$ is assumed to be (uniformly) SPD so that the problem is *coercive*, and *symmetric*. We consider a loop of the form

$$\mathsf{SOLVE} \to \mathsf{ESTIMATE} \to \mathsf{MARK} \to \mathsf{REFINE}$$

that produces a sequence $(\mathscr{T}_k, \mathbb{V}_k, U_k)_{k=0}^{\infty}$ of conforming meshes $\mathscr{T}_k$, spaces of conforming elements $\mathbb{V}_k$ (typically C^0 piecewise linears $n = 1$), and Galerkin solutions $U_k \in \mathbb{V}_k$.

The desired contraction property hinges on *error monotonicity*. Since this is closely related to a minimization principle, it is natural to consider the coercive problem (87). We cannot expect a similar theory for problems governed by an inf-sup condition; this is an important open problem.

We next follow Cascón, Kreuzer, Nochetto and Siebert [14]. We refer to [7, 9, 16, 17, 23, 24, 37, 40, 41, 42] for other approaches and to §5.6 for a discussion.

5.1 Modules of AFEM for the Model Problem

We present further properties of the four basic modules of AFEM for (87). The main additional restrictions with respect to §4 are symmetry and coercivity of the bilinear form and the marking strategy.

Module SOLVE. If $\mathscr{T} \in \mathbb{T}$ is a conforming refinement of $\mathscr{T}_0$ and $\mathbb{V} = \mathbb{V}(\mathscr{T})$ is the finite element space of C^0 piecewise polynomials of degree $\le n$, then

$$U = \mathsf{SOLVE}(\mathscr{T})$$

determines the Galerkin solution *exactly*, namely,

$$U \in \mathbb{V}: \quad \int_\Omega \boldsymbol{A}\nabla U \cdot \nabla V = \int_\Omega fV \quad \text{for all } V \in \mathbb{V}. \tag{88}$$

Module ESTIMATE. Given a conforming mesh $\mathscr{T} \in \mathbb{T}$ and the Galerkin solution $U \in \mathbb{V}(\mathscr{T})$, the output of

$$\{\mathscr{E}_{\mathscr{T}}(U,T)\}_{T\in\mathscr{T}} = \mathsf{ESTIMATE}(U, \mathscr{T}).$$

are the element indicators defined in (46). For convenience, we recall the definitions (37) of *interior* and *jump residuals*

$$\begin{aligned} r(V)|_T &= f + \operatorname{div}(\boldsymbol{A}\nabla V) && \text{for all } T \in \mathscr{T} \\ j(V)|_S &= [\![\boldsymbol{A}\nabla V]\!] \cdot \boldsymbol{n}\,|_S && \text{for all } S \in \mathscr{S} \quad \text{(internal sides of } \mathscr{T}\text{)}, \end{aligned}$$

and $j(V)|_S = 0$ for boundary sides $S \in \mathscr{S}$, as well as the element indicator

$$\mathscr{E}^2_{\mathscr{T}}(V,T) = h_T^2 \, \|r(V)\|^2_{L^2(T)} + h_T \, \|j(V)\|^2_{L^2(\partial T)} \quad \text{for all } T \in \mathscr{T}. \tag{89}$$

We observe that we now write explicitly the argument V in both r and j because this dependence is relevant for the present discussion.

Module MARK. Given $\mathscr{T} \in \mathbb{T}$, the Galerkin solution $U \in \mathbb{V}(\mathscr{T})$, and element indicators $\{\mathscr{E}_{\mathscr{T}}(U,T)\}_{T\in\mathscr{T}}$, the module MARK selects elements for refinement using *Dörfler Marking* (or bulk chasing), i. e., using a fixed parameter $\theta \in (0,1]$ the output

$$\mathscr{M} = \mathsf{MARK}\big(\{\mathscr{E}_{\mathscr{T}}(U,T)\}_{T\in\mathscr{T}}, \mathscr{T}\big)$$

satisfies

$$\mathscr{E}_{\mathscr{T}}(U,\mathscr{M}) \geq \theta \, \mathscr{E}_{\mathscr{T}}(U,\mathscr{T}). \tag{90}$$

This marking guarantees that $\mathscr{M}$ contains a substantial part of the total (or bulk), thus its name. This is a crucial property in our arguments. The choice of $\mathscr{M}$ does not have to be minimal at this stage, that is, the marked elements $T \in \mathscr{M}$ do not necessarily must be those with largest indicators. However, minimality of $\mathscr{M}$ will be crucial to derive rates of convergence in §7.

Module REFINE. Let $b \in \mathbb{N}$ be the number of desired bisections per marked element. Given $\mathscr{T} \in \mathbb{T}$ and a subset $\mathscr{M}$ of marked elements, the output $\mathscr{T}_* \in \mathbb{T}$ of

$$\mathscr{T}_* = \mathsf{REFINE}\big(\mathscr{T}, \mathscr{M}\big)$$

is the smallest refinement $\mathscr{T}_*$ of $\mathscr{T}$ so that all elements of $\mathscr{M}$ are at least bisected b times. Therefore, we have $h_{\mathscr{T}_*} \leq h_{\mathscr{T}}$ and the strict reduction property

$$h_{\mathscr{T}_*}|_T \leq 2^{-b/d} h_{\mathscr{T}}|_T \qquad \text{for all } T \in \mathscr{M}. \tag{91}$$

We finally let $\mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}$ be the subset of refined elements of $\mathscr{T}$ and note that

$$\mathscr{M} \subset \mathscr{R}_{\mathscr{T}\to\mathscr{T}_*}.$$

AFEM. The following procedure is identical to that of §4.1 except for the module MARK, which uses Dörfler marking with parameter $0 < \theta \leq 1$: given an initial grid $\mathscr{T}_0$, set $k = 0$ and iterate

1. $U_k = \mathsf{SOLVE}(\mathscr{T}_k)$;
2. $\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k} = \mathsf{ESTIMATE}(U_k,\mathscr{T}_k)$;
3. $\mathscr{M}_k = \mathsf{MARK}\big(\{\mathscr{E}_k(U_k,T)\}_{T\in\mathscr{T}_k},\mathscr{T}_k\big)$;
4. $\mathscr{T}_{k+1} = \mathsf{REFINE}(\mathscr{T}_k,\mathscr{M}_k)$; $k \leftarrow k+1$.

5.2 Basic Properties of AFEM

We next summarize some basic properties of AFEM that emanate from the symmetry of the differential operator (i.e. of $\boldsymbol{A}$) and features of the modules. In doing this, any explicit constant or hidden constant in $\lesssim$ will only depend on the uniform shape-regularity of $\mathbb{T}$, the dimension d, the polynomial degree n, and the (global) eigenvalues of $\boldsymbol{A}$, but not on a specific grid $\mathscr{T}\in\mathbb{T}$, except if explicitly stated. Furthermore, u will always be the weak solution of (24).

The following property relies on the fact that the bilinear form $\mathscr{B}$ is coercive and symmetric, and so induces a scalar product in $\mathbb{V}$ equivalent to the H^1_0-scalar product.

Lemma 11 (Pythagoras). *Let $\mathscr{T},\mathscr{T}_*\in\mathbb{T}$ such that $\mathscr{T}\le\mathscr{T}_*$. The corresponding Galerkin solutions $U\in\mathbb{V}(\mathscr{T})$ and $U_*\in\mathbb{V}(\mathscr{T}_*)$ satisfy the following orthogonality property in the energy norm $|||\cdot|||_\Omega$*

$$|||u-U|||_\Omega^2 = |||u-U_*|||_\Omega^2 + |||U_*-U|||_\Omega^2\,. \tag{92}$$

Proof. See Problem 41. □

Property (92) is valid for (87) for the energy norm exclusively. This restricts the subsequent analysis to the energy norm, or equivalent norms, but does not extend to other, perhaps more practical, norms such as the maximum norm. This is an important open problem and a serious limitation of this theory.

We now recall the concept of oscillation from §3.1. In view of (65), we denote by $\operatorname{osc}_{\mathscr{T}}(V,T)$ the *element oscillation* for any $V\in\mathbb{V}$

$$\operatorname{osc}_{\mathscr{T}}(V,T) = \|h(r(V)-\overline{r(V)})\|_{L^2(T)} + \|h^{1/2}(j(V)-\overline{j(V)})\|_{L^2(\partial T\cap\Omega)}, \tag{93}$$

where $\overline{r(V)} = P_{2n-2}r(V)$ and $\overline{j(V)} = P_{2n-1}j(V)$ stand for L^2-projections of the residuals $r(V)$ and $j(V)$ onto the polynomials $\mathbb{P}_{2n-2}(T)$ and $\mathbb{P}_{2n-1}(S)$ defined on the element T or side $S\subset\partial T$, respectively. For variable $\boldsymbol{A}$, $\operatorname{osc}_{\mathscr{T}}(V,T)$ depends on the discrete function $V\in\mathbb{V}$, and its study is more involved than for piecewise constant $\boldsymbol{A}$. In the latter case, $\operatorname{osc}_{\mathscr{T}}(V,T)$ is given by (58) and is called *data oscillation*; see also Problem 19.

We now rewrite the a posteriori error estimates of Theorems 6 and 7 in the energy norm.

Lemma 12 (A posteriori error estimates). *There exist constants $0<C_2\le C_1$, such that for any $\mathscr{T}\in\mathbb{T}$ and the corresponding Galerkin solution $U\in\mathbb{V}(\mathscr{T})$ there holds*

$$|||u-U|||_\Omega^2 \le C_1\,\mathscr{E}_{\mathscr{T}}^2(U) \tag{94a}$$

$$C_2\,\mathscr{E}_{\mathscr{T}}^2(U) \le |||u-U|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}}^2(U). \tag{94b}$$

The constants C_1 and C_2 depend on the smallest and largest global eigenvalues of $\boldsymbol{A}$. This dependence can be improved if the a posteriori analysis is carried out directly in the energy norm instead of the H_0^1-norm; see Problem 20. The definitions of $\overline{r(V)}$ and $\overline{j(V)}$, as well as the lower bound (94b), are immaterial for deriving a contraction property. However, they will be important for proving convergence rates in §7.

One serious difficulty in dealing with AFEM is that one has access to the energy error $|||u-U|||_\Omega$ only through the estimator $\mathscr{E}_{\mathscr{T}}(U)$. The latter, however, fails to be monotone because it depends on the discrete solution $U \in \mathbb{V}(\mathscr{T})$ that changes with the mesh. We first show that $\mathscr{E}_{\mathscr{T}}(V)$ decreases strictly provided V does not change (Lemma 13) and next we account for the effect of changing V but keeping the mesh (Lemma 14). Combining these two lemmas we get Proposition 3. In formulating these results we rely on the following notation: given $\mathscr{T} \in \mathbb{T}$ let $\mathscr{M} \subset \mathscr{T}$ denote a set of elements that are bisected $b \ge 1$ times at least, let $\mathscr{T}_* \ge \mathscr{T}$ be a conforming refinement of $\mathscr{T}$ that contains the bisected elements of $\mathscr{M}$, and let

$$\lambda = 1 - 2^{-b/d}.$$

Lemma 13 (Reduction of $\mathscr{E}_{\mathscr{T}}(V)$ wrt $\mathscr{T}$). *For any $V \in \mathbb{V}(\mathscr{T})$, we have*

$$\mathscr{E}_{\mathscr{T}_*}^2(V,\mathscr{T}_*) \le \mathscr{E}_{\mathscr{T}}^2(V,\mathscr{T}) - \lambda\,\mathscr{E}_{\mathscr{T}}^2(V,\mathscr{M}). \tag{95}$$

Proof. We decompose $\mathscr{E}_{\mathscr{T}_*}^2(V,\mathscr{T}_*)$ over elements $T \in \mathscr{T}$, and distinguish whether or not $T \in \mathscr{M}$. If $T \in \mathscr{M}$, then T is bisected at least b times and so T can be written as the union of elements $T' \in \mathscr{T}_*$ We denote this set of elements $\mathscr{T}_*(T)$ and observe that, according with (91), $h_{T'} \le 2^{-b/d}\,h_T$ for all $T' \in \mathscr{T}_*(T)$. Therefore

$$\sum_{T'\in\mathscr{T}_*(T)} h_{T'}^2 \|r(V)\|_{L^2(T')}^2 \le 2^{-2b/d}\,h_T^2\,\|r(V)\|_{L^2(T)}^2$$

and

$$\sum_{T'\in\mathscr{T}_*(T)} h_{T'}\,\|j(V)\|_{L^2(\partial T'\cap\Omega)}^2 \le 2^{-b/d}\,h_T\,\|j(V)\|_{L^2(\partial T\cap\Omega)}^2,$$

because $V \in \mathbb{V}(\mathscr{T})$ only jumps across the boundary of T. This implies

$$\mathscr{E}_{\mathscr{T}_*}^2(V,T) \le 2^{-b/d}\,\mathscr{E}_{\mathscr{T}}^2(V,T) \qquad \text{for all } T \in \mathscr{M}.$$

For the remaining elements $T \in \mathscr{T}\setminus\mathscr{M}$ we only know that mesh-size does not increased because $\mathscr{T} \le \mathscr{T}_*$, whence

$$\mathscr{E}_{\mathscr{T}_*}^2(V,T) \le \mathscr{E}_{\mathscr{T}}^2(V,T) \qquad \text{for all } T \in \mathscr{T}\setminus\mathscr{M}.$$

Combining the two estimates we see that

$$\begin{aligned}\mathscr{E}^2_{\mathscr{T}_*}(V,\mathscr{T}_*) &\le 2^{-b/d}\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M}) + \mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}\setminus\mathscr{M})\\ &= \mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}) - \big(1-2^{-b/d}\big)\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M}),\end{aligned}$$

which, in light of the definition of λ, is the desired estimate. □

Lemma 14 (Lipschitz property of $\mathscr{E}_{\mathscr{T}}(V,T)$ wrt V). *For all $T\in\mathscr{T}$, let ω_T denote the union of elements sharing a side with T, $\operatorname{div}\boldsymbol{A}\in L^\infty(\Omega;\mathbb{R}^d)$ be the divergence of $\boldsymbol{A}$ computed by rows, and*

$$\eta_{\mathscr{T}}(\boldsymbol{A},T) := h_T\|\operatorname{div}\boldsymbol{A}\|_{L^\infty(T)} + \|\boldsymbol{A}\|_{L^\infty(\omega_T)}.$$

Then the following estimate is valid

$$|\mathscr{E}_{\mathscr{T}}(V,T)-\mathscr{E}_{\mathscr{T}}(W,T)| \lesssim \eta_{\mathscr{T}}(\boldsymbol{A},T)\|\nabla(V-W)\|_{L^2(\omega_T)} \quad \text{for all } V,W\in\mathbb{V}(\mathscr{T}).$$

Proof. Recalling the definition of the indicators, the triangle inequality yields

$$|\mathscr{E}_{\mathscr{T}}(V,T)-\mathscr{E}_{\mathscr{T}}(W,T)| \le h_T\|r(V)-r(W)\|_{L^2(T)} + h_T^{1/2}\,\|j(V)-j(W)\|_{L^2(\partial T)}.$$

We set $E:=V-W\in\mathbb{V}(\mathscr{T})$, and observe that

$$r(V)-r(W) = \operatorname{div}(\boldsymbol{A}\nabla E) = \operatorname{div}\boldsymbol{A}\cdot\nabla E + \boldsymbol{A}:D^2E,$$

where D^2E is the Hessian of E. Since E is a polynomial of degree $\le n$ in T, applying the inverse estimate $\|D^2E\|_{L^2(T)}\lesssim h_T^{-1}\|\nabla E\|_{L^2(T)}$, we deduce

$$h_T\|r(V)-r(W)\|_{L^2(T)} \lesssim \eta_{\mathscr{T}}(\boldsymbol{A},T)\|\nabla E\|_{L^2(T)}.$$

On the other hand, for any $S\subset\partial T$ applying the inverse estimate of Problem 43 gives

$$\|j(V)-j(W)\|_{L^2(S)} = \|j(E)\|_{L^2(S)} = \|\,[\![A\nabla E]\!]\,\|_{L^2(S)} \lesssim h_T^{-1/2}\,\|\nabla E\|_{L^2(\omega_T)}$$

where the hidden constant is proportional to $\eta_{\mathscr{T}}(\boldsymbol{A},T)$. This finishes the proof. □

Proposition 3 (Estimator reduction). *Given $\mathscr{T}\in\mathbb{T}$ and a subset $\mathscr{M}\subset\mathscr{T}$ of marked elements, let $\mathscr{T}_*=\mathsf{REFINE}\big(\mathscr{T},\mathscr{M}\big)$. Then there exists a constant $\Lambda>0$, such that for all $V\in\mathbb{V}(\mathscr{T})$, $V_*\in\mathbb{V}_*(\mathscr{T}_*)$ and any $\delta>0$ we have*

$$\begin{aligned}\mathscr{E}^2_{\mathscr{T}_*}(V_*,\mathscr{T}_*) \le (1+\delta)\big(\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{T}) - \lambda\,\mathscr{E}^2_{\mathscr{T}}(V,\mathscr{M})\big)\\ + (1+\delta^{-1})\,\Lambda\,\eta^2_{\mathscr{T}}(\boldsymbol{A},\mathscr{T})\,|||V_*-V|||^2_\Omega\,.\end{aligned}$$

Proof. Apply Lemma 14 to $V,V_*\in\mathbb{V}(\mathscr{T}_*)$ in conjunction with Lemma 13 for V (see Problem 44). □

5.3 Contraction Property of AFEM

A key question to ask is what is (are) the quantity(ies) that AFEM may contract. In light of (92), an obvious candidate is the energy error $|||u-U_k|||_\Omega$. We first show, in the simplest scenario of piecewise constant data $\boldsymbol{A}$ and f, that this is in fact the case provided an interior node property holds; see Lemma 15. However, the energy error may not contract in general unless REFINE enforces several levels of refinement; see Example 1. We then present a more general approach that eliminates the interior node property at the expense of a more complicated contractive quantity, the quasi error; see Theorem 9.

Piecewise Constant Data. We now assume that both f and $\boldsymbol{A}$ are piecewise constant in the initial mesh $\mathscr{T}_0$, so that $\operatorname{osc}_k(U_k)=0$ for all $k\geq 0$. The following property was introduced by Morin, Nochetto, and Siebert [40].

Definition 1 (Interior node property). The refinement $\mathscr{T}_{k+1}\geq\mathscr{T}_k$ satisfies an interior node property with respect to $\mathscr{T}_k$ if each element $T\in\mathscr{M}_k$ contains at least one node of $\mathscr{T}_{k+1}$ in the interiors of T and of each side of T.

This property is valid upon enforcing a fixed number b_* of bisections ($b_*=3,6$ for $d=2,3$). An immediate consequence of this property, proved in [40, 41], is the following *discrete* lower a posteriori bound:

$$C_2\mathscr{E}_k^2(U_k,\mathscr{M}_k)\leq |||U_k-U_{k+1}|||_\Omega^2+\operatorname{osc}_k^2(U_k); \tag{96}$$

see also Problem 23 for a related result.

Lemma 15 (Contraction property for piecewise constant data). *Let $\boldsymbol{A}, f$ be piecewise constant in the initial mesh $\mathscr{T}_0$. If $\mathscr{T}_{k+1}$ satisfies an interior node property with respect to $\mathscr{T}_k$, then for $\alpha:=(1-\theta^2\frac{C_2}{C_1})^{1/2}<1$ there holds*

$$|||u-U_{k+1}|||_\Omega\leq\alpha\,|||u-U_k|||_\Omega\,, \tag{97}$$

where $0<\theta<1$ is the parameter in (90) *and $C_1\geq C_2$ are the constants in* (94).

Proof. For convenience, we use the notation

$$e_k=|||u-U_k|||_\Omega\,,\ E_k=|||U_{k+1}-U_k|||_\Omega\,,\ \mathscr{E}_k=\mathscr{E}_k(U_k,\mathscr{T}_k),\ \mathscr{E}_k(\mathscr{M}_k)=\mathscr{E}_k(U_k,\mathscr{M}_k).$$

The key idea is to use the Pythagoras equality (11)

$$e_{k+1}^2=e_k^2-E_k^2,$$

and show that E_k is a significant portion of e_k. Since (96) with $\operatorname{osc}_k(U_k)=0$ implies

$$C_2\mathscr{E}_k^2(\mathscr{M}_k)\leq E_k^2, \tag{98}$$

applying Dörfler marking (90) and the upper bound (94a), we deduce

$$E_k^2 \geq C_2 \theta^2 \mathscr{E}_k^2 \geq \frac{C_2}{C_1} \theta^2 e_k^2.$$

This is the desired property of E_k and leads to (97). □

Example 1 (Strict monotoniticity). Let $\Omega = (0,1)^2$, $\boldsymbol{A} = \boldsymbol{I}$, $f = 1$ (constant data), and consider the following sequences of meshes depicted in Figure 14. If ϕ_0 denotes the basis function associated with the only interior node of the initial mesh $\mathscr{T}_0$, then

$$U_0 = U_1 = \frac{1}{12}\,\phi_0, \qquad U_2 \neq U_1.$$

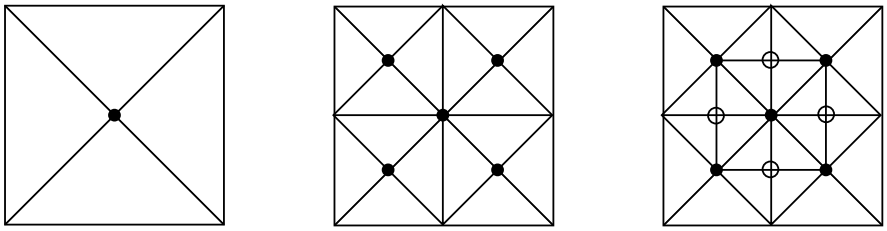


Fig. 14 Grids $\mathscr{T}_0$, $\mathscr{T}_1$, and $\mathscr{T}_2$ of Example 1. The mesh $\mathscr{T}_1$ has nodes in the middle of sides of $\mathscr{T}_0$, but only $\mathscr{T}_2$ has nodes in the interior of elements of $\mathscr{T}_0$. Hence, $\mathscr{T}_2$ satisfies the interior node property of Definition 1 with respect to $\mathscr{T}_0$.

The mesh $\mathscr{T}_1 \geq \mathscr{T}_0$ is produced by a standard 2-step bisection $(b = 2)$ in $2d$. Since $U_0 = U_1$ we conclude that the energy error may not change

$$|||u - U_0|||_\Omega = |||u - U_1|||_\Omega$$

between two consecutive steps of AFEM for $b = d = 2$. This is no longer true provided an interior node in each marked element is created, as in Definition 1, because then Lemma 15 holds. This example appeared first in [40, 41], and was used to justify the interior node property.

General Data. If $\operatorname{osc}_k(U_k) \neq 0$, then the contraction property of AFEM becomes trickier because the energy error and estimator are no longer equivalent regardless of the interior node property. The first question to ask is what quantity replaces the energy error in the analysis. We explore this next and remove the interior node property.

Heuristics. According to (92), the energy error is monotone

$$|||u - U_{k+1}|||_\Omega \leq |||u - U_k|||_\Omega\,,$$

but the previous example shows that strict inequality may fail. However, if $U_{k+1} = U_k$, estimate (95) reveals a strict estimator reduction $\mathscr{E}_{k+1}(U_k) < \mathscr{E}_k(U_k)$. We thus

expect that, for a suitable scaling factor $\gamma > 0$, the so-called *quasi error*

$$|||u - U_k|||_{\Omega}^2 + \gamma \mathscr{E}_k^2(U_k) \tag{99}$$

may be contractive. This heuristics illustrates a distinct aspect of AFEM theory, the interplay between continuous quantities such the energy error $|||u - U_k|||_{\Omega}$ and discrete ones such as the estimator $\mathscr{E}_k(U_k)$: no one alone has the requisite properties to yield a contraction between consecutive adaptive steps.

Theorem 9 (Contraction property). *Let $\theta \in (0,1]$ be the Dörfler Marking parameter, and $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ be a sequence of conforming meshes, finite element spaces and discrete solutions created by AFEM for the model problem* (87)*.*

Then there exist constants $\gamma > 0$ and $0 < \alpha < 1$, additionally depending on the number $b \geq 1$ of bisections and θ, such that for all $k \geq 0$

$$|||u - U_{k+1}|||_{\Omega}^2 + \gamma \mathscr{E}_{k+1}^2(U_{k+1}) \leq \alpha^2 \Big(|||u - U_k|||_{\Omega}^2 + \gamma \mathscr{E}_k^2(U_k) \Big). \tag{100}$$

Proof. We split the proof into four steps and use the notation in Lemma 15.

1 The error orthogonality (92) reads

$$e_{k+1}^2 = e_k^2 - E_k^2. \tag{101}$$

Employing Proposition 3 with $\mathscr{T} = \mathscr{T}_k$, $\mathscr{T}_* = \mathscr{T}_{k+1}$, $V = U_k$ and $V_* = U_{k+1}$ gives

$$\mathscr{E}_{k+1}^2 \leq (1+\delta)\big(\mathscr{E}_k^2 - \lambda\, \mathscr{E}_k^2(\mathscr{M}_k)\big) + (1+\delta^{-1})\,\Lambda_0\, E_k^2, \tag{102}$$

where $\Lambda_0 = \Lambda\, \eta_{\mathscr{T}_0}^2(\boldsymbol{A}, \mathscr{T}_0) \geq \Lambda\, \eta_{\mathscr{T}_k}^2(\boldsymbol{A}, \mathscr{T}_k)$. After multiplying (102) by $\gamma > 0$, to be determined later, we add (101) and (102) to obtain

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \big(\gamma(1+\delta^{-1})\,\Lambda_0 - 1\big)\, E_k^2 + \gamma(1+\delta)\,\big(\mathscr{E}_k^2 - \lambda\, \mathscr{E}_k^2(\mathscr{M}_k)\big).$$

2 We now choose the parameters δ, γ, the former so that

$$(1+\delta)\big(1 - \lambda\,\theta^2\big) = 1 - \frac{\lambda\,\theta^2}{2},$$

and the latter to verify

$$\gamma(1+\delta^{-1})\,\Lambda_0 = 1.$$

Note that this choice of γ yields

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \gamma(1+\delta)\,\big(\mathscr{E}_k^2 - \lambda\, \mathscr{E}_k^2(\mathscr{M}_k)\big).$$

3 We next employ Dörfler Marking, namely $\mathscr{E}_k(\mathscr{M}_k) \geq \theta \mathscr{E}_k$, to deduce

$$e_{k+1}^2 + \gamma \mathscr{E}_{k+1}^2 \leq e_k^2 + \gamma(1+\delta)(1 - \lambda\,\theta^2)\mathscr{E}_k^2$$

which, in conjunction with the choice of δ, gives

$$e_{k+1}^2+\gamma\mathscr{E}_{k+1}^2\le e_k^2+\gamma\left(1-\frac{\lambda\theta^2}{2}\right)\mathscr{E}_k^2=e_k^2-\frac{\gamma\lambda\theta^2}{4}\mathscr{E}_k^2+\gamma\left(1-\frac{\lambda\theta^2}{4}\right)\mathscr{E}_k^2.$$

[4] Finally, the upper bound (94a), namely $e_k^2\le C_1\mathscr{E}_k^2$, implies that

$$e_{k+1}^2+\gamma\mathscr{E}_{k+1}^2\le\left(1-\frac{\gamma\lambda\theta^2}{4C_1}\right)e_k^2+\gamma\left(1-\frac{\lambda\theta^2}{4}\right)\mathscr{E}_k^2.$$

This in turn leads to

$$e_{k+1}^2+\gamma\mathscr{E}_{k+1}^2\le\alpha^2\big(e_k^2+\gamma\mathscr{E}_k^2\big),$$

with

$$\alpha^2:=\max\left\{1-\frac{\gamma\lambda\theta^2}{4C_1},1-\frac{\lambda\theta^2}{4}\right\},$$

and proves the theorem because $\alpha^2<1$. □

Remark 9 (Ingredients). The basic ingredients of this proof are: Dörfler marking; coercivity and symmetry of $\mathscr{B}$ and nesting of spaces, which imply the Pythagoras identity (Lemma 11); the a posteriori upper bound (Lemma 12); and the estimator reduction property (Proposition 3). It does not use the lower bound (94b) and does not require marking by oscillation, as previous proofs do [17, 37, 40, 41, 42].

Remark 10 (Separate marking). MARK is driven by $\mathscr{E}_k$ exclusively, as it happens in all practical AFEM. Previous proofs in [17, 37, 40, 41, 42] require separate marking by estimator and oscillation. It is shown in [14] that separate marking may lead to suboptimal convergence rates. On the other hand, we will prove in §7 that the present AFEM yields quasi-optimal convergence rates.

5.4 Example: Discontinuous Coefficients

We invoke the formulas derived by Kellogg [34] to construct an exact solution of an elliptic problem with piecewise constant coefficients and vanishing right-hand side f; data oscillation is thus immaterial. We now write these formulas in the particular case $\Omega=(-1,1)^2$, $\boldsymbol{A}=a_1\boldsymbol{I}$ in the first and third quadrants, and $\boldsymbol{A}=a_2\boldsymbol{I}$ in the second and fourth quadrants. An exact weak solution u of the model problem (87) for $f\equiv 0$ is given in polar coordinates by $u(r,\theta)=r^\gamma\mu(\theta)$ (see Figure 15), where

$$\mu(\theta)=\begin{cases}\cos((\pi/2-\sigma)\gamma)\cdot\cos((\theta-\pi/2+\rho)\gamma) & \text{if } 0\le\theta\le\pi/2,\\ \cos(\rho\gamma)\cdot\cos((\theta-\pi+\sigma)\gamma) & \text{if } \pi/2\le\theta\le\pi,\\ \cos(\sigma\gamma)\cdot\cos((\theta-\pi-\rho)\gamma) & \text{if } \pi\le\theta<3\pi/2,\\ \cos((\pi/2-\rho)\gamma)\cdot\cos((\theta-3\pi/2-\sigma)\gamma) & \text{if } 3\pi/2\le\theta\le 2\pi,\end{cases}$$

and the numbers γ, ρ, σ satisfy the nonlinear relations

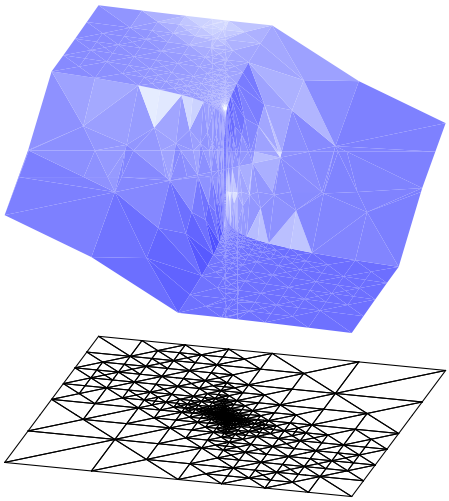


Fig. 15 Discontinuous coefficients in checkerboard pattern: Graph of the discrete solution, which is $u \approx r^{0.1}$, and underlying strongly graded grid. Notice the singularity of u at the origin.

$$\begin{cases} R := a_1/a_2 = -\tan((\pi/2-\sigma)\gamma)\cdot\cot(\rho\gamma), \\ 1/R = -\tan(\rho\gamma)\cdot\cot(\sigma\gamma), \\ R = -\tan(\sigma\gamma)\cdot\cot((\pi/2-\rho)\gamma), \\ 0<\gamma<2, \\ \max\{0,\pi\gamma-\pi\} < 2\gamma\rho < \min\{\pi\gamma,\pi\}, \\ \max\{0,\pi-\pi\gamma\} < -2\gamma\sigma < \min\{\pi,2\pi-\pi\gamma\}. \end{cases} \tag{103}$$

Since we want to test the algorithm AFEM in a worst case scenario, we choose $\gamma = 0.1$, which produces a very singular solution u that is barely in H^1; in fact $u \in H^s(\Omega)$ for $s < 1.1$ and piecewise in $W^2_p(\Omega)$ for $p > 1$. We then solve (103) for R, ρ, and σ using Newton's method to obtain

$$R = a_1/a_2 \cong 161.4476387975881, \quad \rho = \pi/4, \quad \sigma \cong -14.92256510455152,$$

and finally choose $a_1 = R$ and $a_2 = 1$. A smaller γ would lead to a larger ratio R, but in principle γ may be as close to 0 as desired.

We realize from Figure 16 that AFEM attains optimal decay rate for the energy norm. This is consistent with adaptive approximation for functions piecewise in $W^2_p(\Omega)$ (see §1.6), but nonobvious for AFEM which does not have direct access to u; this is the topic of §7. We also notice from Figure 17 that a graded mesh with mesh-size of order 10^{-10} at the origin is achieved with about 2×10^3 elements. To reach a similar resolution with a uniform mesh we would need $N \approx 10^{20}$ elements! This example clearly reveals the advantages and potentials of adaptivity for the FEM even with modest computational resources.

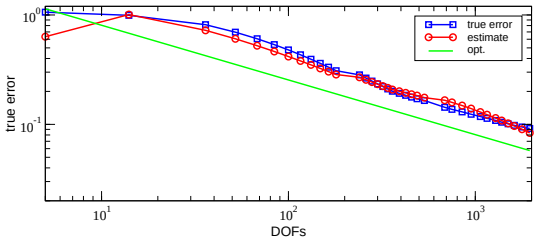


Fig. 16 Quasi-optimality of AFEM for discontinuous coefficients: estimate and true error. The optimal decay for piecewise linear elements in 2d is indicated by the line with slope $-1/2$.

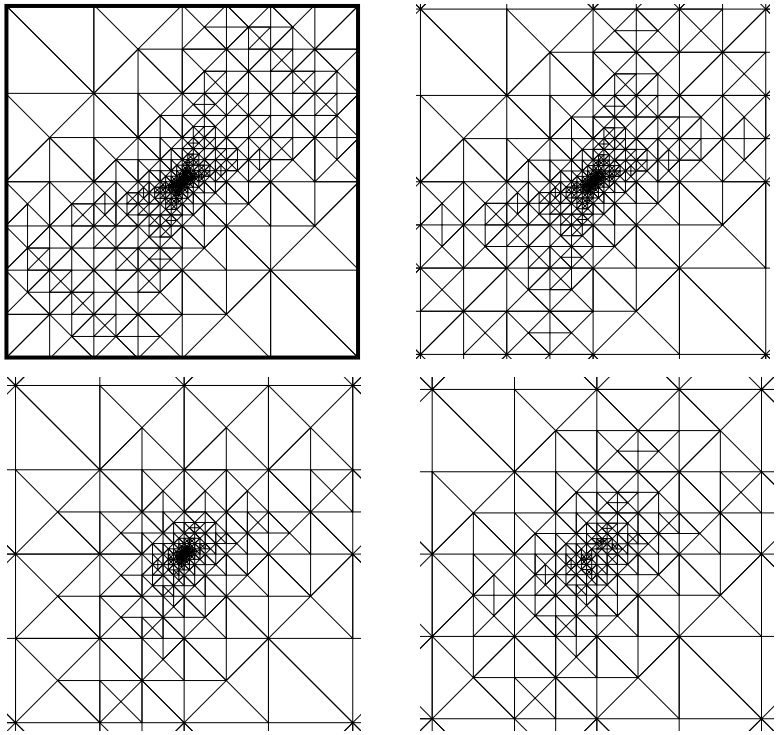


Fig. 17 Discontinuous coefficients in checkerboard pattern: Final grid (full grid with < 2000 nodes) (top left), zooms to $(-10^{-3}, 10^{-3})^2$ (top right), $(-10^{-6}, 10^{-6})^2$ (bottom left), and $(-10^{-9}, 10^{-9})^2$ (bottom right). The grid is highly graded towards the origin. For a similar resolution, a uniform grid would require $N \approx 10^{20}$ elements.

What is missing is an explanation of the recovery of optimal error decay $N^{-1/2}$ through mesh grading. This is the subject of §7, where we have to deal with the interplay between continuous and discrete quantities as already alluded to in the heuristics.

5.5 Extensions and Restrictions

It is important to take a critical look at the theory just developed and wonder about its applicability. Below we list a few extensions of the theory and acknowledge some restrictions.

Nonconforming Meshes. Theorem 9 easily extends to non-conforming meshes since conformity plays no role. This is reported in Bonito and Nochetto [9].

Non-Residual Estimators. The contraction property (100) has been derived for residual estimators $\mathscr{E}_k(U_k)$. This is because the estimator reduction property (95) is not known to hold for other estimators, such as hierarchical, Zienkiewicz-Zhu, and Braess-Schoerbel estimators, as well as those based on the solution of local problems. A common feature of these estimators $\eta_{\mathscr{T}}(U)$ is the lack of reliability in the preasymptotic regime, in which oscillation $\mathrm{osc}_{\mathscr{T}}(U)$ may dominate. In fact, we recall the upper a posteriori bound from (72)

$$|||u-U|||_{\Omega}^2 \le C_1\Big(\eta_{\mathscr{T}}^2(U)+\mathrm{osc}_{\mathscr{T}}^2(U)\Big) =: \mathscr{E}_{\mathscr{T}}^2(U),$$

which gives rise to Dörfler marking for the total estimator $\mathscr{E}_{\mathscr{T}}(U)$. Cascón and Nochetto [15] have recently extended Theorem 9 for $n=1$ upon allowing an interior node property after a fixed number of adaptive loops and combining Lemma 15 with Theorem 9; this is easy to implement within ALBERTA [50]. At the same time, using the local equivalence of the above estimators with the residual one, Kreuzer and Siebert have proved an error reduction property after several adaptive loops [35].

Elliptic PDE on Manifolds. Meckhay, Morin and Nochetto extended this theory to the Laplace-Beltrami operator [38]. In this case, an additional geometric error due to piecewise polynomial approximation of the surface must be accounted for.

Discontinuous Galerkin Methods (dG). The convergence results available in the literature are for the *interior penalty* method [9, 32, 33]. The simplest contraction property (97) for a right-hand side f in the finite element space and the Laplace operator was first derived by Karakashian and Pascal [33], and later improved by Hoppe, Kanschat, and Warburton [32] for $f \in L^2$ and just one bisection per marked element. In both cases, the theory is developed for $d=2$. The most general result, valid for $d \ge 2$, operators with discontinuous variable coefficients, and L^2 data, has been developed by Bonito and Nochetto [9]. The theory in [9] deals with nonconforming meshes made of quadrilaterals or triangles, or their multidimensional generalizations, which are natural in the dG context. A key theoretical issue is the control of the jump term, which is not monotone with refinement [33, 9].

Saddle Point Problems. The contraction properties (97) and (100) rely crucially on the Pythagoras orthogonality property (92) and does not extend to saddle point problems. However, a modified AFEM based on an inexact Uzawa iteration and separate marking was shown to converge by Bänsch, Morin, and Nochetto for the Stokes equation [6]. The situation is somewhat simpler for mixed FEM for scalar

second order elliptic PDE, and has been tackled directly for $d = 2$ by Carstensen and Hoppe for the lowest order Raviart-Thomas element [13], and by Chen, Holst, and Xu for any order [18]. They exploit the underlying special structure: the flux error is L^2-orthogonal to the discrete divergence free subspace, whereas the nonvanishing divergence component of the flux error can be bounded by data oscillation. This is not valid for the Stokes system, which remains open.

Beyond the Energy Framework. The contraction properties (97) and (100) may fail also for other norms of practical interest. An example is the maximum norm, for which there is no convergence result known yet of AFEM. Demlow proved a contraction property for local energy errors [20], and Demlow and Stevenson [21] showed a contraction property for the L^2 norm provided that the mesh grading is sufficiently mild.

5.6 Notes

The theory for conforming meshes in dimension $d > 1$ started with Dörfler [24], who introduced the crucial marking (90), the so-called *Dörfler marking*, and proved strict energy error reduction for the Laplacian provided the initial mesh $\mathscr{T}_0$ satisfies a fineness assumption. This marking plays an essential role in the present discussion, which does not seem to extend to other marking strategies such as those in §4. Morin, Nochetto, and Siebert [40, 41] showed that such strict energy error reduction does not hold in general even for the Laplacian. They introduced the concept of data oscillation and the interior node property, and proved convergence of the AFEM without restrictions on $\mathscr{T}_0$. The latter result, however, is valid only for $\boldsymbol{A}$ in (23) piecewise constant on $\mathscr{T}_0$. Inspired by the work of Chen and Feng [17], Mekchay and Nochetto [37] proved a contraction property for the *total error*, namely the sum of the energy error plus oscillation for $\boldsymbol{A}$ piecewise smooth. The total error will reappear in the study of convergence rates in §7.

Diening and Kreuzer proved a similar contraction property for the p-Laplacian replacing the energy norm by a so-called quasi-norm [23]. They were able to avoid marking for oscillation by using the fact that oscillation is dominated by the estimator. Most results for nonlinear problems utilize the equivalence of the energy error and error in the associated (nonlinear) energy; compare with Problem 42. This equivalence was first used by Veeser in a convergence analysis for the p-Laplacian [55] and later on by Siebert and Veeser for the obstacle problem [51].

The result of Diening and Kreuzer inspired the work by Cascón, Kreuzer, Nochetto, and Siebert [14]. This approach hinges solely on a strict reduction of the mesh-size within refined elements, the upper a posteriori error bound, an orthogonality property natural for (87) in nested approximation spaces, and Dörfler marking. This appears to be the simplest approach currently available.

5.7 Problems

Problem 41 (Pythagoras). Let $\mathbb{V}_1 \subset \mathbb{V}_2 \subset \mathbb{V} = H^1_0(\Omega)$ be nested, conforming and closed subspaces. Let $u \in \mathbb{V}$ be the weak solution to (87), $U_1 \in \mathbb{V}_1$ and $U_2 \in \mathbb{V}_2$ the respective Ritz-Galerkin approximations to u. Prove the orthogonality property

$$|||u-U_1|||^2_\Omega = |||u-U_2|||^2_\Omega + |||U_2-U_1|||^2_\Omega\,. \tag{104}$$

Problem 42 (Error in energy). Let $\mathbb{V}_1 \subset \mathbb{V}_2 \subset \mathbb{V}$ and U_1, U_2, u be as in Problem 41. Recall that u, U_1, U_2 are the unique minimizers of the quadratic energy

$$I[v] := \tfrac{1}{2}\mathscr{B}[v,v] - \langle f, v\rangle$$

in $\mathbb{V}, \mathbb{V}_1, \mathbb{V}_2$ respectively. Show that (104) is equivalent to the identity

$$I[U_1] - I[u] = (I[U_2] - I[u]) + (I[U_1] - I[U_2]).$$

To this end prove

$$I[U_i] - I[u] = \tfrac{1}{2}\,|||U_i - u|||^2_\Omega \qquad \text{and} \qquad I[U_1] - I[U_2] = \tfrac{1}{2}\,|||U_1 - U_2|||^2_\Omega\,.$$

Problem 43 (Inverse estimate). Let $S \in \mathscr{S}$ be an interior side of $T \in \mathscr{T}$, and let $\boldsymbol{A} \in L^\infty(S)$. Make use of a scaling argument to the reference element to show

$$\|\boldsymbol{A}\nabla V\|_S \lesssim h_S^{-1/2}\|\nabla V\|_T \quad \text{for all } V \in \mathbb{V}(\mathscr{T}),$$

where the hidden constant depends on the shape coefficient of $\mathscr{T}$, the dimension d, and $\|\boldsymbol{A}\|_{L^\infty(S)}$.

Problem 44 (Proposition 3). Complete the proof of Proposition 3 upon using Young inequality

$$(a+b)^2 \le (1+\delta)a^2 + (1+\delta^{-1})b^2 \qquad \text{for all } a,b \in \mathbb{R}.$$

Problem 45 (Quasi-local Lipschitz property). Let $\boldsymbol{A} \in W^1_\infty(T)$ for all $T \in \mathscr{T}$. Prove

$$|\operatorname{osc}_{\mathscr{T}}(V,T) - \operatorname{osc}_{\mathscr{T}}(W,T)| \lesssim \operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T)\,\|V-W\|_{H^1(\omega_T)} \qquad \text{for all } V,W \in \mathbb{V},$$

where $\operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T) = h_T\,\|\operatorname{div}\boldsymbol{A} - P^\infty_{n-1}(\operatorname{div}\boldsymbol{A})\|_{L^\infty(T)} + \|\boldsymbol{A} - P^\infty_n\boldsymbol{A}\|_{L^\infty(\omega_T)}$. Proceed as in the proof of Lemma 14 and use Problem 28.

Problem 46 (Perturbation). Let $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$, with $\mathscr{T} \le \mathscr{T}_*$. Use Problem 45 to prove that, for all $V \in \mathbb{V}(\mathscr{T})$ and $V_* \in \mathbb{V}(\mathscr{T}_*)$, there is a constant $\Lambda_1 > 0$ such that

$$\operatorname{osc}^2_{\mathscr{T}}(V, \mathscr{T}\cap\mathscr{T}_*) \le 2\operatorname{osc}^2_{\mathscr{T}_*}(V_*, \mathscr{T}\cap\mathscr{T}_*) + \Lambda_1 \operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A},\mathscr{T}_0)^2\,|||V-V_*|||^2_\Omega\,.$$

6 Complexity of Refinement

This section is devoted to proving Theorem 1 for conforming meshes and Lemma 3 for nonconforming meshes. The results of Sections 6.1 and 6.2 are valid for $d = 2$ but the proofs of Theorem 1 in Section 6.3 and Lemma 3 in Section 6.4 extend easily to $d > 2$. We refer to the survey [45] for a full discussion for $d \geq 2$.

6.1 Chains and Labeling for $d = 2$

In order to study nonlocal effects of bisection for $d = 2$ we introduce now the concept of chain [7]; this concept is not adequate for $d > 2$ [45, 53]. Recall that $E(T)$ denotes the edge of T assigned for refinement. To each $T \in \mathscr{T}$ we associate the element $F(T) \in \mathscr{T}$ sharing the edge $E(T)$ if $E(T)$ is interior and $F(T) = \emptyset$ if $E(T)$ is on $\partial\Omega$. A *chain* $\mathscr{C}(T,\mathscr{T})$, with starting element $T \in \mathscr{T}$, is a sequence $\{T, F(T), \ldots, F^m(T)\}$ with no repetitions of elements and with

$$F^{m+1}(T) = F^k(T) \text{ for } k \in \{0,\ldots,m-1\} \text{ or } F^{m+1}(T) = \emptyset;$$

see Figure 18. We observe that if an element T belongs to two different grids, then the corresponding chains may be different as well. Two adjacent elements $T, T' =$

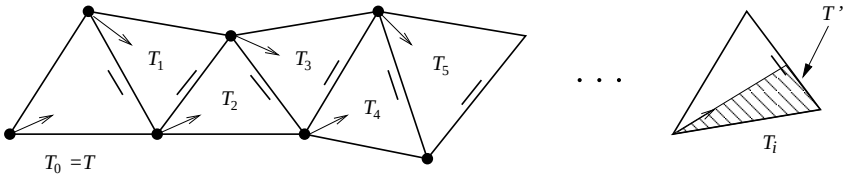


Fig. 18 Typical chain $\mathscr{C}(T,\mathscr{T}) = \{T_j\}_{j=0}^{i}$ emanating from $T = T_0 \in \mathscr{T}$ with $T_j = F(T_{j-1}), j \geq 1$.

$F(T)$ are *compatibly divisible* (or equivalently T, T' form a *compatible bisection patch*) if $F(T') = T$. Hence, $\mathscr{C}(T,\mathscr{T}) = \{T, T'\}$ and a bisection of either T or T' does not propagate outside the patch.

Example (Chains): Let $\mathscr{F} = \{T_i\}_{i=1}^{12}$ be the forest of Figure 3. Then $\mathscr{C}(T_6,\mathscr{T}) = \{T_6, T_7\}, \mathscr{C}(T_9,\mathscr{T}) = \{T_9\}$, and $\mathscr{C}(T_{10},\mathscr{T}) = \{T_{10}, T_8, T_2\}$ are chains, but only $\mathscr{C}(T_6,\mathscr{T})$ is a compatible bisection patch.

To study the structure of chains we rely on the initial labeling (6) and the bisection rule of Section 1.3 (see Figure 5):

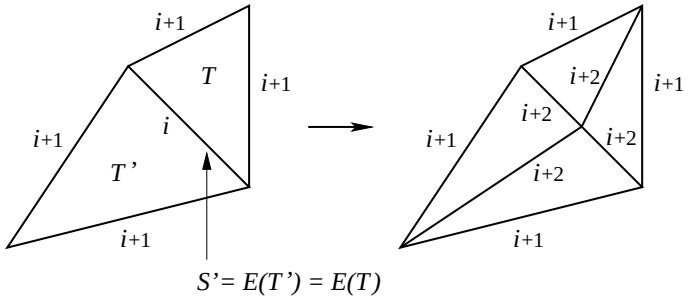
> *every triangle $T \in \mathscr{T}$ with generation $g(T) = i$ receives the label $(i+1, i+1, i)$ with i corresponding to the refinement edge $E(T)$, its side i is bisected and both new sides as well as the bisector are labeled $i+2$ whereas the remaining labels do not change.* (105)

We first show that once the initial labeling and bisection rule are set, the resulting master forest $\mathbb{F}$ is uniquely determined: the label of an edge is independent of the elements sharing this edge and no ambiguity arises in the recursion process.

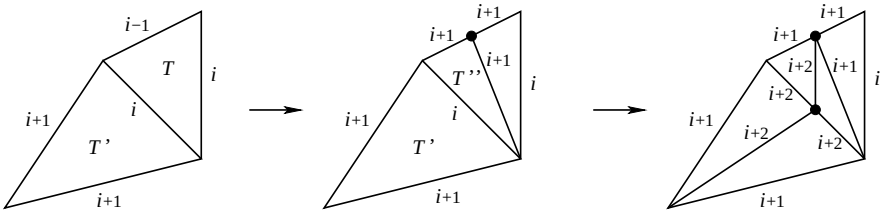
Lemma 16 (Labeling). *Let the initial labeling* (6) *for $\mathscr{T}_0$ and above bisection rule be enforced. If $\mathscr{T}_0 \leq \mathscr{T}_1 \leq \cdots \leq \mathscr{T}_n$ are generated according to* (105)*, then each side in $\mathscr{T}_k$ has a unique label independent of the two triangles sharing this edge.*

Proof. We argue by induction over $\mathscr{T}_k$. For $k = 0$ the assertion is valid due to the initial labeling. Suppose the statement is true for $\mathscr{T}_k$. An edge S in $\mathscr{T}_{k+1}$ can be obtained in two ways. The first is that S is a bisector, and so a new edge, in which case there is nothing to prove about its label being unique. The second possibility is that S was obtained by bisecting an edge $S' \in \mathscr{S}_k$. Let $T, T' \in \mathscr{T}_k$ be the elements sharing S', and let us assume that $E(T') = S'$. Let $(i+1, i+1, i)$ be the label of T', which means that S is assigned the label $i+2$. By induction assumption over $\mathscr{T}_k$, the label of S' as an edge of T is also i. There are two possible cases for the label of T:

- Label $(i+1, i+1, i)$: this situation is symmetric, $E(T) = S'$, and S' is bisected with both halves getting label $i+2$. This is depicted in the figure below.



- Label $(i, i, i-1)$: a bisection of side $E(T)$ with label $i-1$ creates a children T'' with label $(i+1, i+1, i)$ that is compatibly divisible with T'. Joining the new node of T with the midpoint of S' creates a conforming partition with level $i+2$ assigned to S. This is depicted in the figure below.



Therefore, in both cases the label $i+2$ assigned to S is the same from both sides, as asserted. □

The two possible configurations displayed in the two figures above lead readily to the following statement about generations.

Corollary 5 (Generation of Consecutive Elements). *For any $\mathscr{T} \in \mathbb{T}$ and $T, T' = F(T) \in \mathscr{T}$ we either have:*

(a) $g(T) = g(T')$ and T, T' are compatibly divisible, or

(b) $g(T') = g(T) - 1$ and T is compatibly divisible with a child of T'.

Corollary 6 (Generations within a Chain). *For all $\mathscr{T} \in \mathbb{T}$ and $T \in \mathscr{T}$, its chain $\mathscr{C}(T, \mathscr{T}) = \{T_k\}_{k=0}^m$ with $T_k = F^k(T)$ have the property*

$$g(T_k) = g(T) - k \quad 0 \le k \le m-1$$

and $T_m = F^m(T)$ has generation $g(T_m) = g(T_{m-1})$ or it is a boundary element with lowest labeled edge on $\partial\Omega$. In the first case, T_{m-1} and T_m are compatibly divisible.

Proof. Apply Corollary 5 repeatly to consecutive elements of $\mathscr{C}(T, \mathscr{T})$. □

6.2 Recursive Bisection

Given an element $T \in \mathscr{M}$ to be refined, the routine REFINE_RECURSIVE $(\mathscr{T}, T)$ recursively refines the chain $\mathscr{C}(T, \mathscr{T})$ of T, from the end back to T, and creates a minimal conforming partition $\mathscr{T}_* \ge \mathscr{T}$ such that T is bisected once. This procedure reads as follows:

```
REFINE_RECURSIVE (𝒯, T)
if g(F(T)) < g(T)
    𝒯 := REFINE_RECURSIVE (𝒯, F(T));
else
    bisect the compatible bisection patch 𝒞(T, 𝒯);
    update 𝒯;
end if
return (𝒯)
```

We denote by $\mathscr{C}_*(T, \mathscr{T}) \subset \mathscr{T}_*$ the recursive refinement of $\mathscr{C}(T, \mathscr{T})$ (or completion of $\mathscr{C}(T, \mathscr{T})$) caused by bisection of T. Since REFINE_RECURSIVE refines solely compatible bisection patches, intermediate meshes are always conforming.

We refer to Figure 19 for an example of recursive bisection $\mathscr{C}_*(T_{10}, \mathscr{T})$ of $\mathscr{C}(T_{10}, \mathscr{T}) = \{T_{10}, T_8, T_2\}$ in Figure 2: REFINE_RECURSIVE starts bisecting from the end of $\mathscr{C}(T_{10}, \mathscr{T})$, namely T_2, which is a boundary element, and goes back the chain bisecting elements twice until it gets to T_{10}.

We now establish a fundamental property of REFINE_RECURSIVE $(\mathscr{T}, T)$ relating the generation of elements within $\mathscr{C}_*(T, \mathscr{T})$.

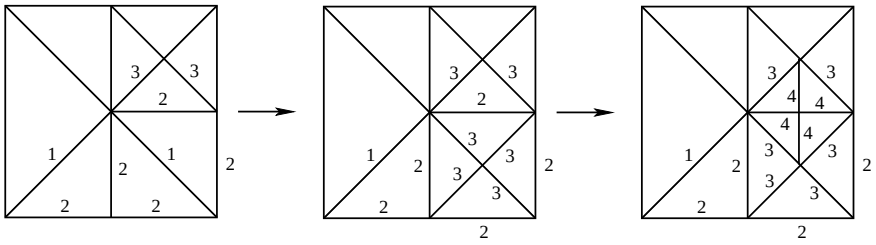


Fig. 19 Recursive refinement of $T_{10} \in \mathscr{T}$ in Figure 2 by REFINE_RECURSIVE. This entails refining the chain $\mathscr{C}(T_{10}, \mathscr{T}) = \{T_{10}, T_8, T_2\}$, starting from the last element $T_2 \in \mathscr{T}$, which form alone a compatible bisection patch because its refinement edge is on the boundary, and continuing with $T_8 \in \mathscr{T}$ and finally $T_{10} \in \mathscr{T}$. Note that the successive meshes are always conforming and that REFINE_RECURSIVE bisects elements in $\mathscr{C}(T_{10}, \mathscr{T})$ twice before getting back to T_{10}.

Lemma 17 (Recursive Refinement). *Let $\mathscr{T}_0$ satisfy the labeling* (6)*, and let $\mathscr{T} \in \mathbb{T}$ be a conforming refinement of $\mathscr{T}_0$. A call to REFINE_RECURSIVE $(\mathscr{T}, T)$ terminates, for all $T \in \mathscr{M}$, and outputs the smallest conforming refinement $\mathscr{T}_*$ of $\mathscr{T}$ such that T is bisected. In addition, all newly created $T' \in \mathscr{C}_*(T, \mathscr{T})$ satisfy*

$$g(T') \leq g(T) + 1. \tag{106}$$

Proof. We first observe that T has maximal generation within $\mathscr{C}(T, \mathscr{T})$. So recursion is applied to elements with generation $\leq g(T)$, whence the recursion terminates. We also note that this procedure creates children of T and either children or grandchildren of triangles $T_k \in \mathscr{C}(T, \mathscr{T}) = \{T_i\}_{i=0}^m$ with $k \geq 1$. If T' is a child of T there is nothing to prove. If not, we consider first $m = 1$, in which case T' is a child of T_1 because T_0 and T_1 are compatibly divisible and so have the same generation; thus $g(T') = g(T_1) + 1 = g(T_0) + 1$. Finally, if $m > 1$, then $g(T_k) < g(T)$ and we apply Corollary 6 to deduce

$$g(T') \leq g(T_k) + 2 \leq g(T) + 1,$$

as asserted. □

The following crucial lemma links generation and distance between T and $T' \in \mathscr{C}_*(T, \mathscr{T})$, the latter being defined as

$$\operatorname{dist}(T', T) := \inf_{x' \in T', x \in T} |x' - x|.$$

Lemma 18 (Distance and Generation). *Let $T \in \mathscr{M}$. Any newly created $T' \in \mathscr{C}_*(T, \mathscr{T})$ by REFINE_RECURSIVE $(\mathscr{T}, T)$ satisfies*

$$\operatorname{dist}(T', T) \leq D_2 \frac{2}{\sqrt{2} - 1} 2^{-g(T')/2}, \tag{107}$$

where $D_2 > 0$ is the constant in (5)*.*

Proof. Suppose $T' \subset T_i \in \mathscr{C}(T,\mathscr{T})$ have been created by subdividing T_i (see Figure 18). If $i \leq 1$ then $\operatorname{dist}(T',T) = 0$ and there is nothing to prove. If $i > 1$, then we observe that $\operatorname{dist}(T',T_{i-1}) = 0$, whence

$$\begin{aligned}\operatorname{dist}(T',T) &\leq \operatorname{dist}(T_{i-1},T) + \operatorname{diam}(T_{i-1}) \leq \sum_{k=1}^{i-1} \operatorname{diam}(T_k) \\ &\leq D_2 \sum_{k=1}^{i-1} 2^{-g(T_k)/2} < D_2 \frac{1}{1-2^{-1/2}}\, 2^{-g(T_{i-1})/2},\end{aligned}$$

because the generations decrease exactly by 1 along the chain $\mathscr{C}(T)$ according to Corollary 5(b). Since T' is a child or grandchild of T_i, we deduce

$$g(T') \leq g(T_i) + 2 = g(T_{i-1}) + 1,$$

whence

$$\operatorname{dist}(T',T) < D_2 \frac{2^{1/2}}{1-2^{-1/2}}\, 2^{-g(T')/2}.$$

This is the desired estimate. □

The recursive procedure REFINE_RECURSIVE is the core of the routine REFINE of Section 1.3: given a conforming mesh $\mathscr{T} \in \mathbb{T}$ and a subset $\mathscr{M} \subset \mathscr{T}$ of marked elements, REFINE creates a conforming refinement $\mathscr{T}_* \geq \mathscr{T}$ of $\mathscr{T}$ such that all elements of $\mathscr{M}$ are bisected at least once:

REFINE $(\mathscr{T},\mathscr{M})$
for all $T \in \mathscr{M} \cap \mathscr{T}$ do
  $\mathscr{T}$:= REFINE_RECURSIVE $(\mathscr{T},T)$;
end
return $(\mathscr{T})$

It may happen that an element $T' \in \mathscr{M}$ is scheduled prior to T for refinement and $T \in \mathscr{C}(T',\mathscr{T})$. Since the call REFINE_RECURSIVE $(\mathscr{T},T')$ bisects T, its two children replace T in $\mathscr{T}$. This implies that $T \notin \mathscr{M} \cap \mathscr{T}$, which prevents further refinement of T.

In practice, one often likes to bisect selected elements several times, for instance each marked element is scheduled for $b \geq 1$ bisections. This can be done by assigning the number $b(T) = b$ of bisections that have to be executed for each marked element T. If T is bisected then we assign $b(T)-1$ as the number of pending bisections to its children and the set of marked elements is $\mathscr{M} := \{T \in \mathscr{T} \mid b(T) > 0\}$.

6.3 Conforming Meshes: Proof of Theorem 1

Figure 19 reveals that the issue of propagation of mesh refinement to keep conformity is rather delicate. In particular, an estimate of the form

$$\#\mathscr{T}_k - \#\mathscr{T}_{k-1} \le \Lambda \#\mathscr{M}_{k-1}$$

is not valid with a constant Λ independent of k; in fact the constant can be proportional to k according to Figure 19.

Binev, Dahmen, and DeVore [7] for $d = 2$ and Stevenson [53] for $d > 2$ show that control of the propagation of refinement by bisection is possible when considering the collective effect:

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \le \Lambda_0 \sum_{j=0}^{k-1} \#\mathscr{M}_j.$$

This can be heuristically motivated as follows. Consider the set $\mathscr{M} := \bigcup_{j=0}^{k-1} \mathscr{M}_j$ used to generate the sequence $\mathscr{T}_0 \le \mathscr{T}_1 \le \cdots \le \mathscr{T}_k =: \mathscr{T}$. Suppose that each element $T_* \in \mathscr{M}$ is assigned a fixed amount C_1 of money to spend on refined elements in $\mathscr{T}$, i. e., on $T \in \mathscr{T} \setminus \mathscr{T}_0$. Assume further that $\lambda(T, T_*)$ is the portion of money spent by T_* on T. Then it must hold

$$\sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \lambda(T, T_*) \le C_1 \qquad \text{for all } T_* \in \mathscr{M}. \tag{108a}$$

In addition, we suppose that the investment of all elements in $\mathscr{M}$ is fair in the sense that each $T \in \mathscr{T} \setminus \mathscr{T}_0$ gets at least a fixed amount C_2, whence

$$\sum_{T_* \in \mathscr{M}} \lambda(T, T_*) \ge C_2 \qquad \text{for all } T \in \mathscr{T} \setminus \mathscr{T}_0. \tag{108b}$$

Therefore, summing up (108b) and using the upper bound (108a) we readily obtain

$$C_2(\#\mathscr{T} - \#\mathscr{T}_0) \le \sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \sum_{T_* \in \mathscr{M}} \lambda(T, T_*) = \sum_{T_* \in \mathscr{M}} \sum_{T \in \mathscr{T} \setminus \mathscr{T}_0} \lambda(T, T_*) \le C_1 \#\mathscr{M},$$

which proves Theorem 1 for $\mathscr{T}$ and $\mathscr{M}$. In the remainder of this section we design such an allocation function $\lambda : \mathscr{T} \times \mathscr{M} \to \mathbb{R}^+$ in several steps and prove that recurrent refinement by bisection yields (108) provided $\mathscr{T}_0$ satisfies (6).

Construction of the Allocation Function. The function $\lambda(T, T_*)$ is defined with the help of two sequences $\big(a(\ell)\big)_{\ell=-1}^{\infty}$, $\big(b(\ell)\big)_{\ell=0}^{\infty} \subset \mathbb{R}^+$ of positive numbers satisfying

$$\sum_{\ell \ge -1} a(\ell) = A < \infty, \qquad \sum_{\ell \ge 0} 2^{-\ell/2} b(\ell) = B < \infty, \qquad \inf_{\ell \ge 1} b(\ell)\, a(\ell) = c_* > 0,$$

and $b(0) \ge 1$. Valid instances are $a(\ell) = (\ell + 2)^{-2}$ and $b(\ell) = 2^{\ell/3}$.

With these settings we are prepared to define $\lambda : \mathscr{T} \times \mathscr{M} \to \mathbb{R}^+$ by

$$\lambda(T, T_*) := \begin{cases} a(g(T_*) - g(T)), & \operatorname{dist}(T, T_*) < D_3 B 2^{-g(T)/d} \text{ and } g(T) \le g(T_*) + 1 \\ 0, & \text{else,} \end{cases}$$

where $D_3 := D_2\big(1+2(\sqrt{2}-1)^{-1}\big)$. Therefore, the investment of money by $T_* \in \mathscr{M}$ is restricted to cells T that are sufficiently close and are of generation $g(T) \leq g(T_*)+1$. Only elements of such generation can be created during refinement of T_* according to Lemma 17. We stress that except for the definition of B, this construction is mutidimensional and we refer to [45, 53] for details.

The following lemma shows that the total amount of money spend by the allocation function $\lambda(T,T_*)$ per marked element T_* is bounded.

Lemma 19 (Upper Bound). *There exists a constant $C_1 > 0$ only depending on $\mathscr{T}_0$ such that λ satisfies* (108a)*, i. e.,*

$$\sum_{T\in\mathscr{T}\setminus\mathscr{T}_0} \lambda(T,T_*) \leq C_1 \qquad \text{for all } T_* \in \mathscr{M}.$$

Proof. [1] Given $T_* \in \mathscr{M}$ we set $g_* = g(T_*)$ and we let $0 \leq g \leq g_*+1$ be a generation of interest in the definition of λ. We claim that for such g the cardinality of the set

$$\mathscr{T}(T_*,g) = \{T \in \mathscr{T} \mid \operatorname{dist}(T,T_*) < D_3 B 2^{-g/2} \text{ and } g(T) = g\}$$

is uniformly bounded, i. e., $\#\mathscr{T}(T_*,g) \leq C$ with C solely depending on D_1, D_2, D_3, B.

From (5) we learn that $\operatorname{diam}(T_*) \leq D_2 2^{-g_*/2} \leq 2D_2 2^{-(g_*+1)/2} \leq 2D_2 2^{-g/2}$ as well as $\operatorname{diam}(T) \leq D_2 2^{-g/2}$ for any $T \in \mathscr{T}(T_*,g)$. Hence, all elements of the set $\mathscr{T}(T_*,g)$ lie inside a ball centered at the barycenter of T_* with radius $(D_3B + 3D_2)2^{-g/2}$. Again relying on (5) we thus conclude

$$\#\mathscr{T}(T_*,g) D_1 2^{-g} \leq \sum_{T\in\mathscr{T}(T_*,g)} |T| \leq c(D_3B+3D_2)^2 2^{-g},$$

whence $\#\mathscr{T}(T_*,g) \leq c D_1^{-1}(D_3B+3D_2)^2 =: C$.

[2] Accounting only for non-zero contributions $\lambda(T,T_*)$ we deduce

$$\sum_{T\in\mathscr{T}\setminus\mathscr{T}_0} \lambda(T,T_*) = \sum_{g=0}^{g_*+1} \sum_{T\in\mathscr{T}(T_*,g)} a(g_*-g) \leq C \sum_{\ell=-1}^{\infty} a(\ell) = CA =: C_1,$$

which is the desired upper bound. □

The definition of λ also implies that each refined element receives a fixed amount of money. We show this next.

Lemma 20 (Lower Bound). *There exists a constant $C_2 > 0$ only depending on $\mathscr{T}_0$ such that λ satisfies* (108b)*, i. e.,*

$$\sum_{T_*\in\mathscr{M}} \lambda(T,T_*) \geq C_2 \qquad \text{for all } T \in \mathscr{T}\setminus\mathscr{T}_0.$$

Proof. [1] Fix an arbitrary $T_0 \in \mathscr{T}\setminus\mathscr{T}_0$. Then there is an iteration count $1 \leq k_0 \leq k$ such that $T_0 \in \mathscr{T}_{k_0}$ and $T_0 \notin \mathscr{T}_{k_0-1}$. Therefore there exists an $T_1 \in \mathscr{M}_{k_0-1} \subset \mathscr{M}$

such that T_0 is generated during REFINE_RECURSIVE $(\mathscr{T}_{k_0-1}, T_1)$. Iterating this process we construct a sequence $\{T_j\}_{j=1}^J \subset \mathscr{M}$ with corresponding iteration counts $\{k_j\}_{j=1}^J$ such that T_j is created by REFINE_RECURSIVE $(\mathscr{T}_{k_j-1}, T_{j+1})$. The sequence is finite since the iteration counts are strictly decreasing and thus $k_J = 0$ for some $J > 0$, or equivalently $T_J \in \mathscr{T}_0$.

Since T_j is created during refinement of T_{j+1} we infer from (106) that

$$g(T_{j+1}) \geq g(T_j) - 1.$$

Accordingly, $g(T_{j+1})$ can decrease the previous value of $g(T_j)$ at most by 1. Since $g(T_J) = 0$ there exists a smallest value s such that $g(T_s) = g(T_0) - 1$. Note that for $j = 1, \ldots, s$ we have $\lambda(T_0, T_j) > 0$ if $\operatorname{dist}(T_0, T_j) \leq D_3 B g^{-g(T_0)/d}$.

2 We next estimate the distance $\operatorname{dist}(T_0, T_j)$. For $1 \leq j \leq s$ and $\ell \geq 0$ we define the set

$$\mathscr{T}(T_0, \ell, j) := \{T \in \{T_0, \ldots, T_{j-1}\} \mid g(T) = g(T_0) + \ell\}$$

and denote by $m(\ell, j)$ its cardinality. The triangle inequality combined with an induction argument yields

$$\begin{aligned}\operatorname{dist}(T_0, T_j) &\leq \operatorname{dist}(T_0, T_1) + \operatorname{diam}(T_1) + \operatorname{dist}(T_1, T_j) \\ &\leq \sum_{i=1}^{j} \operatorname{dist}(T_{i-1}, T_i) + \sum_{i=1}^{j-1} \operatorname{diam}(T_i).\end{aligned}$$

We apply (107) for the terms of the first sum and (5) for the terms of the second sum to obtain

$$\begin{aligned}\operatorname{dist}(T_0, T_j) &< D_2 \frac{2}{\sqrt{2}-1} \sum_{i=1}^{j} 2^{-g(T_{i-1})/2} + D_2 \sum_{i=1}^{j-1} 2^{-g(T_i)/2} \\ &\leq D_2 \left(1 + \frac{2}{\sqrt{2}-1}\right) \sum_{i=0}^{j-1} 2^{-g(T_i)/2} \\ &= D_3 \sum_{\ell=0}^{\infty} m(\ell, j)\, 2^{-(g(T_0)+\ell)/2} \\ &= D_3 2^{-g(T_0)/2} \sum_{\ell=0}^{\infty} m(\ell, j)\, 2^{-\ell/2}.\end{aligned}$$

For establishing the lower bound we distinguish two cases depending on the size of $m(\ell, s)$. This is done next.

3 *Case 1:* $m(\ell, s) \leq b(\ell)$ for all $\ell \geq 0$. From this we conclude

$$\operatorname{dist}(T_0, T_s) < D_3 2^{-g(T_0)/2} \sum_{\ell=0}^{\infty} b(\ell)\, 2^{-\ell/2} = D_3 B 2^{-g(T_0)/2}$$

and the definition of λ then readily implies

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \lambda(T_0,T_s) = a(g(T_s)-g(T_0)) = a(-1) > 0.$$

4 *Case 2:* There exists $\ell \geq 0$ such that $m(\ell,s) > b(\ell)$. For each of these ℓ's there exists a smallest $j = j(\ell)$ such that $m(\ell, j(\ell)) > b(\ell)$. We let ℓ^* be the index ℓ that gives rise to the smallest $j(\ell)$, and set $j^* = j(\ell^*)$. Consequently

$$m(\ell, j^*-1) \leq b(\ell) \quad \text{for all } \ell \geq 0 \qquad \text{and} \qquad m(\ell^*, j^*) > b(\ell^*).$$

As in Case 1 we see $\operatorname{dist}(T_0,T_i) < D_3 B 2^{-g(T_0)/2}$ for all $i \leq j^*-1$, or equivalently

$$\operatorname{dist}(T_0,T_i) < D_3 B 2^{-g(T_0)/2} \qquad \text{for all } T_i \in \mathscr{T}(T_0,\ell,j^*).$$

We next show that the elements in $\mathscr{T}(T_0,\ell^*,j^*)$ spend enough money on T_0. We first consider $\ell^* = 0$ and note that $T_0 \in \mathscr{T}(T_0,0,j^*)$. Since $m(0,j^*) > b(0) \geq 1$ we discover $j^* \geq 2$. Hence, there is an $T_i \in \mathscr{T}(T_0,0,j^*) \cap \mathscr{M}$, which yields the estimate

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \lambda(T_0,T_i) = a(g(T_i)-g(T_0)) = a(0) > 0.$$

For $\ell^* > 0$ we see that $T_0 \notin \mathscr{T}(T_0,\ell^*,j^*)$, whence $\mathscr{T}(T_0,\ell^*,j^*) \subset \mathscr{M}$. In addition, $\lambda(T_0,T_i) = a(\ell^*)$ for all $T_i \in \mathscr{T}(T_0,\ell^*,j^*)$. From this we conclude

$$\begin{aligned}\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) &\geq \sum_{T_*\in\mathscr{T}(T_0,\ell^*,j^*)} \lambda(T_0,T_*) = m(\ell^*,j^*)\, a(\ell^*) \\ &> b(\ell^*)\, a(\ell^*) \geq \inf_{\ell\geq 1} b(\ell)\, a(\ell) = c_* > 0.\end{aligned}$$

5 In summary we have proved the assertion since for any $T_0 \in \mathscr{T} \setminus \mathscr{T}_0$

$$\sum_{T_*\in\mathscr{M}} \lambda(T_0,T_*) \geq \min\{a(-1),a(0),c_*\} =: C_2 > 0. \qquad \square$$

Remark 11 (Complexity with $b > 1$ Bisections). To show the complexity estimate when REFINE performs $b > 1$ bisections, the set $\mathscr{M}_k$ is to be understood as a sequence of *single* bisections recorded in sets $\{\mathscr{M}_k(j)\}_{j=1}^b$, which belong to intermediate triangulations between $\mathscr{T}_k$ and $\mathscr{T}_{k+1}$ with $\#\mathscr{M}_k(j) \leq 2^{j-1}\#\mathscr{M}_k$, $j = 1,\dots,b$. Then we also obtain Theorem 1 because

$$\sum_{j=1}^{b} \#\mathscr{M}_k(j) \leq \sum_{j=1}^{b} 2^{j-1}\#\mathscr{M}_k = (2^b-1)\#\mathscr{M}_k.$$

In practice, it is customary to take $b = d$ [50].

6.4 Nonconforming Meshes: Proof of Lemma 3

We now examine briefly the refinement process for quadrilaterals with one hanging node per edge, which gives rise to the so-called *1-meshes*. The refinement of $T \in \mathscr{T}$ might affect four elements of $\mathscr{T}$ for $d = 2$ (or 2^d elements for any dimension $d \geq 2$), all contained in the *refinement patch* $R(T, \mathscr{T})$ of T in $\mathscr{T}$. The latter is defined as

$$R(T, \mathscr{T}) := \{T' \in \mathscr{T} |\ T' \text{ and } T \text{ share an edge and } g(T') \leq g(T)\},$$

and is called *compatible* provided $g(T') = g(T)$ for all $T' \in R(T, \mathscr{T})$. The generation gap between elements sharing an edge, in particular those in $R(T, \mathscr{T})$, is always ≤ 1 for 1-meshes, and is 0 if $R(T, \mathscr{T})$ is compatible. The element size satisfies

$$h_T = 2^{-g(T)} h_{T_0} \qquad \forall T \in \mathscr{T}$$

where $T_0 \in \mathscr{T}_0$ is the ancestor of T in the initial mesh $\mathscr{T}_0$. Lemma 2 is thus valid

$$h_T < \bar{h}_T \leq D_2 2^{-g(T)} \qquad \forall T \in \mathscr{T}. \tag{109}$$

Given an element $T \in \mathscr{M}$ to be refined, the routine REFINE_RECURSIVE $(\mathscr{T}, T)$ refines recursively $R(T, \mathscr{T})$ in such a way that the intermediate meshes are always 1-meshes, and reads as follows:

```
REFINE_RECURSIVE (𝒯, T)
if g = min{g(T'') : T'' ∈ R(T, 𝒯)} < g(T)
    let T' ∈ R(T, 𝒯) satisfy g(T') = g
    𝒯 := REFINE_RECURSIVE (𝒯, T');
else
    subdivide T;
    update 𝒯 upon replacing T by its children;
end if
return (𝒯)
```

The conditional prevents the generation gap within $R(T, \mathscr{T})$ from getting larger than 1. If it fails, then the refinement patch $R(T, \mathscr{T})$ is compatible and refinining T increases the generation gap from 0 to 1 without violating the 1-mesh structure. This implies Lemma 17: for all newly created elements $T' \in \mathscr{T}_*$

$$g(T') \leq g(T) + 1. \tag{110}$$

In addition, REFINE_RECURSIVE $(\mathscr{T}, T)$ creates a minimal 1-mesh $\mathscr{T}_* \geq \mathscr{T}$ refinement of $\mathscr{T}$ so that T is subdivided only *once*. This yields Lemma 18: there exist a geometric constant $D > 0$ such that for all newly created elements $T' \in \mathscr{T}_*$

$$\operatorname{dist}(T, T') \leq D 2^{g(T')}. \tag{111}$$

The procedure REFINE_RECURSIVE is the core of REFINE, which is conceptually identical to that in Section 6.2. Suppose that each marked element $T \in \mathscr{M}$ is to be subdivided $\rho \geq 1$ times. We assign a flag $q(T)$ to each element T which is initialized $q(T) = \rho$ if $T \in \mathscr{M}$ and $q(T) = 0$ otherwise. The marked set $\mathscr{M}$ is then the set of elements T with $q(T) > 0$, and every time T is subdivided it is removed from $\mathscr{T}$ and replaced by its children, which inherit the flag $q(T) - 1$. This avoids the conflict of subdividing again an element that has been previously refined by REFINE_RECURSIVE. The procedure REFINE $(\mathscr{T}, \mathscr{M})$ reads

```
REFINE (𝒯, ℳ)
for all T ∈ ℳ ∩ 𝒯 do
    𝒯 := REFINE_RECURSIVE (𝒯, T);
end
return (𝒯)
```

and its output is a minimal 1-mesh $\mathscr{T}_* \geq \mathscr{T}$, refinement of $\mathscr{T}$, so that all marked elements of $\mathscr{M}$ are refined at least ρ times. Since $\mathscr{T}_*$ has one hanging node per side it is thus admissible in the sense (22). However, the refinement may spread outside $\mathscr{M}$ and the issue of complexity of REFINE again becomes non-trivial.

With the above ingredients in place, the proof of Lemma 3 follows along the lines of Section 6.3; see Problem 50.

6.5 Notes

The complexity theory for bisection hinges on the initial labeling (6) for $d = 2$. That such a labeling exists is due to Mitchell [39, Theorem 2.9] and Binev, Dahmen, and DeVore [7, Lemma 2.1], but the proofs are not constructive. A couple of global bisections of $\mathscr{T}_0$, as depicted in Figure 6, guarantee (6) over the ensuing mesh. For $d > 2$ the corresponding initial labeling is due to Stevenson [53, Section 4 - Condition (b)], who in turn improves upon Maubach [36] and Traxler [54] and shows how to impose it upon further refining each element of $\mathscr{T}_0$. We refer to the survey [45] for a discussion of this condition: a key consequence is that every uniform refinement of $\mathscr{T}_0$ gives a conforming bisection mesh.

The fundamental properties of chains, especially Lemmas 17 and 18, along with the clever ideas of Section 6.3 are due to Binev, Dahmen, and DeVore [7] for $d = 2$, and Stevenson for $d > 2$; see [45]. Bonito and Nochetto [9] observed, in the context of dG methods, that such properties extend to admissible nonconforming meshes.

6.6 Problems

Problem 47 (Largest number of bisections). Show that REFINE_RECURSIVE $(\mathscr{T}, T)$ for $d = 2$ bisects T exactly once and all the elements in the chain $\mathscr{C}(T, \mathscr{T})$ at

most twice. This property extends to $d > 2$ provided the initial labeling of Stevenson [53, Section 4 - Condition (b)] is enforced.

Problem 48 (Properties of generation for quad-refinement). Prove (110) and (111).

Problem 49 (Largest number of subdivisions for quads). Show that the procedure REFINE_RECURSIVE $(\mathscr{T}, T)$ subdivides T exactly once and never subdivides any other quadrilateral of $\mathscr{T}$ more than once.

Problem 50 (Lemma 3). Combine (110) and (111) to prove Lemma 3 for any $\rho \geq 1$.

7 Convergence Rates

We have already realized in §1.6 that we can a priori accommodate the degrees of freedom in such a way that the finite element approximation retains optimal energy error decay for a class of singular functions. This presumes knowledge of the exact solution u. At the same time, we have seen numerical evidence in §5.4 that the standard AFEM of §5.1, achieves such a performance without direct access to the exact solution u. Practical experience strongly suggests that this is even true for a much larger class of problems and adaptive methods. The challenge ahead is to reconcile these two distinct aspects of AFEM.

A crucial insight in such a connection for the simplest scenario, the Laplacian and piecewise constant forcing f, is due to Stevenson [52]:

> *any marking strategy that reduces the energy error relative to the current value must contain a substantial portion of $\mathscr{E}_{\mathscr{T}}(U)$, and so it can be related to Dörfler Marking.* (112)

This allows one to compare meshes produced by AFEM with optimal ones and to conclude a quasi-optimal error decay. We discuss this issue in §7.3. However, this is not enough to handle the model problem (87) with variable $\boldsymbol{A}$ and f.

The objective of this section is to study (87) for general data $\boldsymbol{A}$ and f. This study hinges on the total error and its relation with the quasi error, which is contracted by AFEM. This approach allows us to improve upon and extend Stevenson [52] to variable data. In doing so, we follow closely Cascón, Kreuzer, Nochetto, and Siebert [14]. The present theory, however, does not extend to noncoercive problems and marking strategies other than Dörfler's. These remain important open questions.

As in §5, u will always be the weak solution of (87) and, except when stated otherwise, any explicit constant or hidden constant in $\lesssim$ may depend on the uniform shape-regularity of $\mathbb{T}$, the dimension d, the polynomial degree n, the (global) eigenvalues of $\boldsymbol{A}$, and the oscillation $\operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A})$ of $\boldsymbol{A}$ on the initial mesh $\mathscr{T}_0$, but not on a specific grid $\mathscr{T} \in \mathbb{T}$.

7.1 The Total Error

We first introduce the concept of *total error* for the Galerkin function $U \in \mathbb{V}(\mathscr{T})$

$$|||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(U), \tag{113}$$

(see Mekchay and Nochetto [37]), and next assert its equivalence to the quasi error (99). In fact, in view of the upper and lower a posteriori error bounds (94), and

$$\operatorname{osc}_{\mathscr{T}}^2(U) \leq \mathscr{E}_{\mathscr{T}}^2(U),$$

we have

$$C_2\,\mathscr{E}_{\mathscr{T}}^2(U) \le |||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(U)$$
$$\le |||u-U|||_{\Omega}^2 + \mathscr{E}_{\mathscr{T}}^2(U) \le (1+C_1)\,\mathscr{E}_{\mathscr{T}}^2(U),$$

whence

$$\mathscr{E}_{\mathscr{T}}^2(U) \approx |||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(U). \tag{114}$$

Since AFEM selects elements for refinement based on information extracted exclusively from the error indicators $\{\mathscr{E}_{\mathscr{T}}(U,T)\}_{T\in\mathscr{T}}$, we realize that the decay rate of AFEM must be characterized by the total error. Moreover, on invoking the upper bound (94a) again, we also see that the total error is equivalent to the quasi error

$$|||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(U) \approx |||u-U|||_{\Omega}^2 + \mathscr{E}_{\mathscr{T}}^2(U).$$

The latter is the quantity being strictly reduced by AFEM (Theorem 9). Finally, the total error satisfies the following Cea's type-lemma, or equivalently AFEM is quasi-optimal regarding the total error. In fact, if the oscillation vanishes, then this is Cea's Lemma stated in Theorem 4; see also Problem 12.

Lemma 21 (Quasi-optimality of total error). *There exists an explicit constant Λ_2, which depends on $\boldsymbol{A}$ $\mathscr{T}_0$, n and d, such that for any $\mathscr{T}\in\mathbb{T}$ and the corresponding Galerkin solution $U\in\mathbb{V}(\mathscr{T})$ there holds*

$$|||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(U) \le \Lambda_2 \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-V|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(V)\Big).$$

Proof. For $\varepsilon>0$ choose $V_\varepsilon\in\mathbb{V}(\mathscr{T})$, with

$$|||u-V_\varepsilon|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(V_\varepsilon) \le (1+\varepsilon) \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-V|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(V)\Big).$$

Applying Problem 46 with $\mathscr{T}_*=\mathscr{T}$, $V=U$, and $V_*=V_\varepsilon$ yields

$$\operatorname{osc}_{\mathscr{T}}^2(U) \le 2\,\operatorname{osc}_{\mathscr{T}}^2(V_\varepsilon) + C_3\,|||U-V_\varepsilon|||_{\Omega}^2,$$

with

$$C_3 := \Lambda_1 \operatorname{osc}_{\mathscr{T}_0}(\boldsymbol{A})^2.$$

Since $U\in\mathbb{V}(\mathscr{T})$ is the Galerkin solution, $U-V_\varepsilon\in\mathbb{V}(\mathscr{T})$ is orthogonal to $u-U$ in the energy norm, whence $|||u-U|||_{\Omega}^2 + |||U-V_\varepsilon|||_{\Omega}^2 = |||u-V_\varepsilon|||_{\Omega}^2$ and

$$|||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(U) \le \big(1+C_3\big)\,|||u-V_\varepsilon|||_{\Omega}^2 + 2\,\operatorname{osc}_{\mathscr{T}}^2(V_\varepsilon)$$
$$\le (1+\varepsilon)\,\Lambda_2 \inf_{V\in\mathbb{V}(\mathscr{T})} \Big(|||u-U|||_{\Omega}^2 + \operatorname{osc}_{\mathscr{T}}^2(V)\Big),$$

with $\Lambda_2 = \max\big\{2, 1+C_3\big\}$. The assertion follows upon taking $\varepsilon\to 0$. □

7.2 Approximation Classes

In view of (114) and Lemma 21, the definition of approximation class $\mathbb{A}_s$ hinges on the concept of best total error:

$$\inf_{V\in\mathbb{V}(\mathscr{T})}\Big(|||u-V|||_\Omega^2+\operatorname{osc}_{\mathscr{T}}^2(V)\Big).$$

We first let $\mathbb{T}_N\subset\mathbb{T}$ be the set of all possible conforming refinements of $\mathscr{T}_0$ with at most N elements more than $\mathscr{T}_0$, i. e.,

$$\mathbb{T}_N=\{\mathscr{T}\in\mathbb{T}\mid \#\mathscr{T}-\#\mathscr{T}_0\le N\}.$$

The quality of the best approximation in $\mathbb{T}_N$ with respect to the total error is characterized by

$$\sigma(N;u,f,\boldsymbol{A}):=\inf_{\mathscr{T}\in\mathbb{T}_N}\inf_{V\in\mathbb{V}(\mathscr{T})}\big(|||u-V|||_\Omega^2+\operatorname{osc}_{\mathscr{T}}^2(V)\big)^{1/2},$$

and the approximation class $\mathbb{A}_s$ for $s>0$ is defined by

$$\mathbb{A}_s:=\Big\{(v,f,\boldsymbol{A})\mid |v,f,\boldsymbol{A}|_s:=\sup_{N>0}\big(N^s\,\sigma(N;v,f,\boldsymbol{A})\big)<\infty\Big\}.$$

Therefore, if $(v,f,\boldsymbol{A})\in\mathbb{A}_s$, then $\sigma(N;v,f,\boldsymbol{A})\lesssim N^{-s}$ decays with rate N^{-s}. We point out the upper bound $s\le n/d$ for polynomial degree $n\ge 1$; this can be seen with full regularity and uniform refinement (see (14)). Note that if $(v,f,\boldsymbol{A})\in\mathbb{A}_s$ then for all $\varepsilon>0$ there exist $\mathscr{T}_\varepsilon\ge\mathscr{T}_0$ conforming and $V_\varepsilon\in\mathbb{V}(\mathscr{T}_\varepsilon)$ such that (see Problem 51)

$$|||v-V_\varepsilon|||_\Omega^2+\operatorname{osc}_{\mathscr{T}_\varepsilon}^2(V_\varepsilon)\le\varepsilon^2\qquad\text{and}\qquad \#\mathscr{T}_\varepsilon-\#\mathscr{T}_0\le|v,f,\boldsymbol{A}|_s^{1/s}\,\varepsilon^{-1/s}. \tag{115}$$

In addition, thanks to Lemma 21, the solution u with data $(f,\boldsymbol{A})$ satisfies

$$\sigma(N;u,f,\boldsymbol{A})\approx\inf_{\mathscr{T}\in\mathbb{T}_N}\big\{\mathscr{E}_{\mathscr{T}}(U,\mathscr{T})\mid U=\mathsf{SOLVE}(\mathbb{V}(\mathscr{T}))\big\}. \tag{116}$$

This establishes a direct connection between $\mathbb{A}_s$ and AFEM.

Mesh Overlay. For the subsequent discussion it will be convenient to merge two conforming meshes $\mathscr{T}_1,\mathscr{T}_2\in\mathbb{T}$. Given the corresponding forests $\mathscr{F}_1,\mathscr{F}_2\in\mathbb{F}$ we consider the set $\mathscr{F}_1\cup\mathscr{F}_2\in\mathbb{F}$, which satisfies $\mathscr{T}_0\subset\mathscr{F}_1\cup\mathscr{F}_2$. Then $\mathscr{F}_1\cup\mathscr{F}_2$ is a forest and its leaves are called the *overlay* of $\mathscr{F}_1$ and $\mathscr{F}_2$:

$$\mathscr{T}_1\oplus\mathscr{T}_2=\mathscr{T}(\mathscr{F}_1\cup\mathscr{F}_2).$$

We next bound the cardinality of $\mathscr{T}_1\oplus\mathscr{T}_2$ in terms of that of $\mathscr{T}_1$ and $\mathscr{T}_2$; see [14, 52].

Lemma 22 (Overlay). *The overlay $\mathscr{T}=\mathscr{T}_1\oplus\mathscr{T}_2$ is conforming and*

$$\#\mathscr{T}\le\#\mathscr{T}_1+\#\mathscr{T}_2-\#\mathscr{T}_0. \tag{117}$$

Proof. See Problem 52. □

Discussion of $\mathbb{A}_s$. We now would like to show a few examples of membership in $\mathbb{A}_s$ and highlight some important open questions. We first investigate the class $\mathbb{A}_s$ for piecewise polynomial coefficient matrix $\boldsymbol{A}$ of degree $\le n$ over $\mathscr{T}_0$. In this simplified scenario, the oscillation $\operatorname{osc}_{\mathscr{T}}(U)$ reduces to *data oscillation* (see (58) and (93)):

$$\operatorname{osc}_{\mathscr{T}}(U) = \operatorname{osc}_{\mathscr{T}}(f) := \|h(f - P_{2n-2}\, f)\|_{L^2(\Omega)}.$$

We then have the following characterization of $\mathbb{A}_s$ in terms of the standard approximation classes [7, 8, 52]:

$$\mathscr{A}_s := \Big\{ v \in \mathbb{V} \mid \ |v|_{\mathscr{A}_s} := \sup_{N>0} \big(N^s \inf_{\mathscr{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathscr{T})} |||v-V|||_\Omega \big) < \infty \Big\},$$
$$\overline{\mathscr{A}_s} := \Big\{ g \in L^2(\Omega) \mid \ |g|_{\overline{\mathscr{A}_s}} := \sup_{N>0} \big(N^s \inf_{\mathscr{T}\in\mathbb{T}_N} \operatorname{osc}_{\mathscr{T}}(g) \big) < \infty \Big\}.$$

Lemma 23 (Equivalence of classes). *Let $\boldsymbol{A}$ be piecewise polynomial of degree $\le n$ over $\mathscr{T}_0$. Then $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ if and only if $(u, f) \in \mathscr{A}_s \times \overline{\mathscr{A}_s}$ and*

$$|u, f, \boldsymbol{A}|_s \approx |u|_{\mathscr{A}_s} + |f|_{\overline{\mathscr{A}_s}}. \tag{118}$$

Proof. It is obvious that $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ implies $(u, f) \in \mathscr{A}_s \times \overline{\mathscr{A}_s}$ as well as the bound $|u|_{\mathscr{A}_s} + |f|_{\overline{\mathscr{A}_s}} \lesssim |u, f, \boldsymbol{A}|_s$.

In order to prove the reverse inequality, let $(u, f) \in \mathscr{A}_s \times \overline{\mathscr{A}_s}$. Then there exist $\mathscr{T}_1, \mathscr{T}_2 \in \mathbb{T}_N$ so that $\big|\!\big|\!\big| u - U_{\mathscr{T}_1} \big|\!\big|\!\big|_\Omega \le |u|_{\mathscr{A}_s} N^{-s}$ where $U_{\mathscr{T}_1} \in \mathbb{V}(\mathscr{T}_1)$ is the best approximation and $\operatorname{osc}_{\mathscr{T}_2}(f, \mathscr{T}_2) \le |f|_{\overline{\mathscr{A}_s}} N^{-s}$.

The overlay $\mathscr{T} = \mathscr{T}_1 \oplus \mathscr{T}_2 \in \mathbb{T}_{2N}$ according to (117), and

$$|||u - U_{\mathscr{T}}|||_\Omega + \operatorname{osc}_{\mathscr{T}}(f) \le \big|\!\big|\!\big| u - U_{\mathscr{T}_1} \big|\!\big|\!\big|_\Omega + \operatorname{osc}_{\mathscr{T}_2}(f) \le 2^s \big(|u|_{\mathscr{A}_s} + |f|_{\overline{\mathscr{A}_s}} \big) (2N)^{-s}.$$

This yields $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ together with the bound $|u, f, \boldsymbol{A}|_s \lesssim |u|_{\mathscr{A}_s} + |f|_{\overline{\mathscr{A}_s}}$. □

Corollary 7 (Membership in $\mathbb{A}_{1/2}$ with piecewise linear $\boldsymbol{A}$). *Let $d = 2$, $n = 1$, and $u \in H^1_0(\Omega)$ be the solution of the model problem with piecewise linear $\boldsymbol{A}$ and $f \in L^2(\Omega)$. If $u \in W^2_p(\Omega; \mathscr{T}_0)$ is piecewise W^2_p over the initial grid $\mathscr{T}_0$ and $p > 1$, then $(u, f, \boldsymbol{A}) \in \mathbb{A}_{1/2}$ and*

$$|u, f, \boldsymbol{A}|_{1/2} \lesssim \|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)}. \tag{119}$$

Proof. Since $f \in L^2(\Omega)$, we realize that for all uniform refinements $\mathscr{T} \in \mathbb{T}$ we have

$$\operatorname{osc}_{\mathscr{T}}(f) = \|h(f - P_0 f)\|_{L^2(\Omega)} \le h_{\max}(\mathscr{T}) \|f\|_{L^2(\Omega)} \lesssim (\#\mathscr{T})^{-1/2} \|f\|_{L^2(\Omega)},$$

This implies $f \in \overline{\mathscr{A}_{1/2}}$ with $|f|_{\overline{\mathscr{A}_{1/2}}} \lesssim \|f\|_{L^2(\Omega)}$. Moreover, for $u \in W^2_p(\Omega; \mathscr{T}_0)$ we learn from Corollary 2 and Remark 6 of §1.6 that $u \in \mathscr{A}_{1/2}$ and $|u|_{\mathscr{A}_{1/2}} \lesssim \|D^2 u\|_{L^2(\Omega;\mathscr{T}_0)}$. The assertion then follows from Lemma 23. □

Corollary 8 (Membership in $\mathbb{A}_{1/2}$ with variable A). *Let $d = 2$, $n = 1$, $p > 1$, $f \in L^2(\Omega)$. Let $\boldsymbol{A} \in W^1_\infty(\Omega, \mathscr{T}_0)$ be piecewise Lipschitz and $u \in W^2_p(\Omega; \mathscr{T}_0) \cap H^1_0(\Omega)$ be piecewise W^2_p over the initial mesh $\mathscr{T}_0$. Then $(u, f, \boldsymbol{A}) \in \mathbb{A}_{1/2}$ and*

$$|u, f, \boldsymbol{A}|_{1/2} \lesssim \|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)}. \tag{120}$$

Proof. Combine Problem 55 with Corollary 2. □

Corollary 9 (Membership in $\mathbb{A}_s$ with $s < 1/d$). *Let $d \geq 2$, $n = 1$, $1 < t < 2$, $p > 1$, and $f \in L^2(\Omega)$. Let $\boldsymbol{A} \in W^1_\infty(\Omega, \mathscr{T}_0)$ be piecewise Lipschitz and $u \in W^t_p(\Omega; \mathscr{T}_0) \cap H^1_0(\Omega)$ be piecewise W^t_p over the initial mesh $\mathscr{T}_0$ with $t - \frac{d}{p} > 1 - \frac{d}{2}$. Then $(u, f, \boldsymbol{A}) \in \mathbb{A}_{(t-1)/d}$ and*

$$|u, f, \boldsymbol{A}|_{(t-1)/d} \lesssim \|D^t u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)}. \tag{121}$$

Proof. Combine Problem 9 with Problem 55. □

Example 2 (Pre-asymptotics). Corollary 7 shows that oscillation decays with rate $1/2$ for $f \in L^2(\Omega)$. Since the decay rate of the total error is $s \leq 1/2$, oscillation can be ignored asymptotically; this is verified in Problems 56, 57, and 58. However, oscillation may dominate the total error, or equivalently the class $\mathbb{A}_s$ may fail to describe the behavior of $|||u - U_k|||_\Omega$, in the early stages of adaptivity. In fact, we recall from Problem 32 that the discrete solution $U_k = 0$, and $|||u - U_k|||_\Omega \approx 2^{-K}$ is constant for as many steps $k \leq K$ as desired. In contrast, $\mathscr{E}_k(U_k) = \mathrm{osc}_k(U_k) = \|h(f - \overline{f})\|_{L^2(\Omega)} = \|hf\|_{L^2(\Omega)}$ reduces strictly for $k \leq K$ but overestimates $|||u - U_k|||_\Omega$. The fact that the preasymptotic regime $k \leq K$ for the energy error could be made arbitrarily long would be problematic if we were to focus exclusively on $|||u - U_k|||_\Omega$.

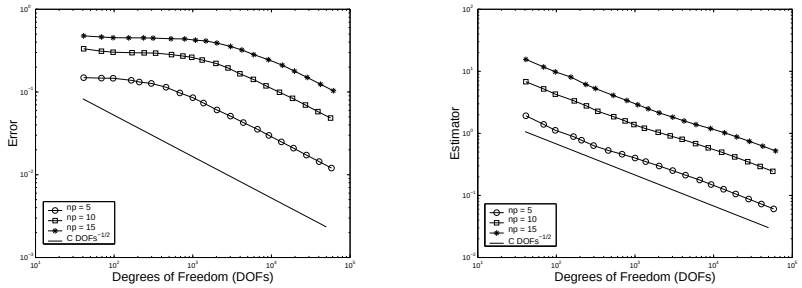


Fig. 20 Decay of the energy error (left) and the estimator (right) for the smooth solution u_S of (122) with frequencies $\kappa = 5, 10$, and 15. The energy error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. This behavior is allowed by $\mathbb{A}_s$.

In practice, this effect is typically less dramatic because f is not orthogonal to $\mathbb{V}(\mathscr{T}_k)$. Figure 20 displays the behavior of the AFEM for the smooth solution u_S

$$u_S(x,y) = 10^{-2} a_i^{-1}(x^2 + y^2)\sin^2(\kappa\pi x)\sin^2(\kappa\pi y), \qquad 1 \leq i \leq 4, \tag{122}$$

of the model problem (87) with discontinuous coefficients $\{a_i\}_{i=1}^4$ in checkerboard pattern as in §5.4 and frequencies $\kappa = 5, 10$, and 15. We can see that the error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. In contrast, the estimator decays always with the optimal rate. Since all decisions of the AFEM are based on the estimator, this behavior has to be expected and is consistent with our notion of approximation class $\mathbb{A}_s$, which can be characterized just by the estimator according to (116).

7.3 Quasi-Optimal Cardinality: Vanishing Oscillation

In this section we follow the ideas of Stevenson [52] for the simplest scenario with vanishing oscillation $\operatorname{osc}_{\mathscr{T}}(U) = 0$, and thereby explore the insight (112). We recall that in this case the a posteriori error estimates (94) become

$$C_2\,\mathscr{E}_{\mathscr{T}}^2(U) \leq |||u-U|||_{\Omega}^2 \leq C_1\,\mathscr{E}_{\mathscr{T}}^2(U). \tag{123}$$

It is then evident that the ratio $C_2/C_1 \leq 1$, between the *reliability* constant C_1 and the *efficiency* constant C_2, is a quality measure of the estimatior $\mathscr{E}_{\mathscr{T}}(U)$: the closer to 1 the better! This ratio is usually closer to 1 for non-residual estimators, such as those discussed in §5.5, but their theory is a bit more cumbersome.

Assumptions for Optimal Decay Rate. The following are further restrictions on AFEM to achieve optimal error decay, as predicted by the approximation class $\mathscr{A}_s$.

Assumption 1 (Marking parameter: vanishing oscillation). *The parameter θ of Dörfler marking satisfies $\theta \in (0, \theta_*)$ with*

$$\theta_* := \sqrt{\frac{C_2}{C_1}} \tag{124}$$

Assumption 2 (Cardinality of $\mathscr{M}$). *MARK selects a set $\mathscr{M}$ with* minimal *cardinality.*

Assumption 3 (Initial labeling). *The labeling of the initial mesh $\mathscr{T}_0$ satisfies* (6) *for $d = 2$ or its multimensional counterpart for $d > 2$ [52, 45].*

A few comments about these assumptions are now in order.

Remark 12 (Threshold $\theta_* < 1$). It is reasonable to be cautious in making marking decisions if the constants C_1 and C_2 are very disparate, and thus the ratio C_2/C_1 is far from 1. This justifies the upper bound $\theta_* \leq 1$ in Assumption 1.

Remark 13 (Minimal $\mathscr{M}$). According to the equidistribution principle (16) and the local lower bound (68) without oscillation, it is natural to mark elements with largest error indicators. This leads to a minimal set $\mathscr{M}$, as stated in Assumption 2, and turns out to be crucial to link AFEM with optimal meshes and approximation classes.

Remark 14 (Initial triangulation). Assumption 3 guarantees the complexity estimate of module REFINE stated in Theorem 1 and proved in §6.3:

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathscr{M}_j.$$

Assumption 3 is rather restrictive for dimension $d > 2$. Any other refinement giving the same complexity estimate can replace REFINE together with Assumption 3.

Even though we cannot expect local upper bounds between the continuous and discrete solution, the following crucial result shows that this is not the case between discrete solutions on nested meshes $\mathscr{T}_* \geq \mathscr{T}$: what matters is the set of elements of $\mathscr{T}$ which are no longer in $\mathscr{T}_*$.

Lemma 24 (Localized upper bound). *Let $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$ satisfy $\mathscr{T}_* \geq \mathscr{T}$ and let $\mathscr{R} := \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ be the refined set. If $U \in \mathbb{V}$, $U_* \in \mathbb{V}_*$ are the corresponding Galerkin solutions, then*

$$|||U_* - U|||_\Omega^2 \leq C_1 \, \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{R}). \tag{125}$$

Proof. See Problem 53. □

We are now ready to explore Stevenson's insight (112) for the simplest scenario.

Lemma 25 (Dörfler marking: vanishing oscillation). *Let θ satisfy Assumption 1 and set $\mu := 1 - \theta^2/\theta_*^2 > 0$. Let $\mathscr{T}_* \geq \mathscr{T}$ and the corresponding Galerkin solution $U_* \in \mathbb{V}(\mathscr{T}_*)$ satisfy*

$$|||u - U_*|||_\Omega^2 \leq \mu \, |||u - U|||_\Omega^2 . \tag{126}$$

Then the refined set $\mathscr{R} = \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ satisfies the Dörfler property

$$\mathscr{E}_{\mathscr{T}}(U, \mathscr{R}) \geq \theta \, \mathscr{E}_{\mathscr{T}}(U, \mathscr{T}). \tag{127}$$

Proof. Since $\mu < 1$ we use the lower bound in (123), in conjunction with (126) and Pythagoras equality (92), to derive

$$\begin{aligned}(1-\mu) C_2 \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{T}) &\leq (1-\mu) \, |||u - U|||_\Omega^2 \\ &\leq |||u - U|||_\Omega^2 - |||u - U_*|||_\Omega^2 = |||U - U_*|||_\Omega^2 .\end{aligned}$$

In view of Lemma 24, we thus deduce

$$(1-\mu) C_2 \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{T}) \leq C_1 \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{R}),$$

which is the assertion in disguise. □

To examine the cardinality of $\mathscr{M}_k$ in terms of $|||u - U_k|||_\Omega$ we must relate AFEM with the approximation class $\mathscr{A}_s$. Even though this might appear like an undoable task, the key to unravel this connection is given by Lemma 25. We show this now.

Lemma 26 (Cardinality of $\mathscr{M}_k$). *Let Assumptions 1 and 2 hold. If $u \in \mathscr{A}_s$ then*

$$\#\mathscr{M}_k \lesssim |u|_s^{1/s} \, |||u-U_k|||_\Omega^{-1/s} \quad \forall k \geq 0. \tag{128}$$

Proof. We invoke that $u \in \mathscr{A}_s$, together with Problem 51 with $\varepsilon^2 = \mu \, |||u-U_k|||_\Omega^2$, to find a mesh $\mathscr{T}_\varepsilon \in \mathbb{T}$ and the Galerkin solution $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ so that

$$|||u-U_\varepsilon|||_\Omega^2 \leq \varepsilon^2, \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u|_s^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}}.$$

Since $\mathscr{T}_\varepsilon$ may be totally unrelated to $\mathscr{T}_k$, we introduce the overlay $\mathscr{T}_* = \mathscr{T}_\varepsilon \oplus \mathscr{T}_k$. We exploit the property $\mathscr{T}_* \geq \mathscr{T}_\varepsilon$ to conclude that the Galerkin solution $U_* \in \mathbb{V}(\mathscr{T}_*)$ satisfies (127):

$$|||u-U_*|||_\Omega^2 \leq |||u-U_\varepsilon|||_\Omega^2 \leq \varepsilon^2 = \mu \, |||u-U|||_\Omega^2 .$$

Therefore, Lemma 25 implies that the refined set $\mathscr{R} = \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ satisfies a Dörfler marking with parameter $\theta < \theta_*$. But MARK delivers a minimal set $\mathscr{M}_k$ with this property, according to Assumption 2, whence

$$\#\mathscr{M}_k \leq \#\mathscr{R} \leq \#\mathscr{T}_* - \#\mathscr{T}_k \leq \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u|_s^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}},$$

where we use Lemma 22 to account for the overlay. The proof is complete. □

Proposition 4 (Quasi-optimality: vanishing oscillation). *Let Assumptions 1, 2, and 3 hold. If $u \in \mathscr{A}_s$, then AFEM gives rise to a sequence $(\mathscr{T}_k, \mathbb{V}_k, U_k)_{k=0}^\infty$ such that*

$$|||u-U_k|||_\Omega \lesssim |u|_s \, (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s} \quad \forall k \geq 1.$$

Proof. We make use of Assumption 3, along with Theorem 1, to infer that

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathscr{M}_j \lesssim |u|_s^{\frac{1}{s}} \sum_{j=0}^{k-1} \left|||u-U_j|||_\Omega^{-\frac{1}{s}} .$$

We now use the contraction property (97) of Lemma 15

$$|||u-U_k|||_\Omega \leq \alpha^{k-j} \left|||u-U_j|||_\Omega$$

to replace the sum above by

$$\sum_{j=0}^{k-1} \left|||u-U_j|||_\Omega^{-\frac{1}{s}} \leq |||u-U_k|||_\Omega^{-\frac{1}{s}} \sum_{j=0}^{k-1} \alpha^{\frac{k-j}{s}} < \frac{\alpha^{\frac{1}{s}}}{1-\alpha^{\frac{1}{s}}} \, |||u-U_k|||_\Omega^{-\frac{1}{s}} ,$$

because $\alpha < 1$ and the series is summable. This completes the proof. □

7.4 Quasi-Optimal Cardinality: General Data

In this section we remove the restriction $\operatorname{osc}_{\mathscr{T}}(U) = 0$, and thereby make use of the basic ingredients developed in §7.1 and §7.2. Therefore, we replace the energy error by the total error and the linear approximation class $\mathscr{A}_s$ for u by the nonlinear class $\mathbb{A}_s$ for the triple $(u, f, \boldsymbol{A})$. To account for the presence of general f and $\boldsymbol{A}$, we need to make an even more stringent assumption on the threshold θ_*.

Assumption 4 (Marking parameter: general data). *Let $C_3 = \Lambda_1 \operatorname{osc}^2_{\mathscr{T}_0}(\boldsymbol{A})$ be the constant in Problem 46 and Lemma 21. The marking parameter θ satisfies $\theta \in (0, \theta_*)$ with*

$$\theta_* = \sqrt{\frac{C_2}{1 + C_1(1 + C_3)}}. \tag{129}$$

We now proceed along the same lines as those of §7.3.

Lemma 27 (Dörfler marking: general data). *Let Assumption 4 hold and set $\mu := \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$. If $\mathscr{T}_* \geq \mathscr{T}$ and the corresponding Galerkin solution $U_* \in \mathbb{V}(\mathscr{T}_*)$ satisfy*

$$|||u - U_*|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}_*}(U_*) \leq \mu \big(|||u - U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U) \big), \tag{130}$$

then the refined set $\mathscr{R} = \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ satisfies the Dörfler property

$$\mathscr{E}_{\mathscr{T}}(U, \mathscr{R}) \geq \theta \, \mathscr{E}_{\mathscr{T}}(U, \mathscr{T}). \tag{131}$$

Proof. We split the proof into four steps.

1 In view of the global lower bound (94b)

$$C_2 \, \mathscr{E}^2_{\mathscr{T}}(U) \leq |||u - U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U)$$

and (130), we can write

$$\begin{aligned}(1 - 2\mu) C_2 \, \mathscr{E}^2_{\mathscr{T}}(U) &\leq (1 - 2\mu) \big(|||u - U|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}}(U) \big) \\ &\leq \big(|||u - U|||^2_{\Omega} - 2 \, |||u - U_*|||^2_{\Omega} \big) + \big(\operatorname{osc}^2_{\mathscr{T}}(U) - 2 \operatorname{osc}^2_{\mathscr{T}_*}(U_*) \big).\end{aligned}$$

2 Combining the Pythagoras orthogonality relation (92)

$$|||u - U|||^2_{\Omega} - |||u - U_*|||^2_{\Omega} = |||U - U_*|||^2_{\Omega} \, .$$

with the localized upper bound Lemma 24 yields

$$|||u - U|||^2_{\Omega} - 2 \, |||u - U_*|||^2_{\Omega} \leq |||U - U_*|||^2_{\Omega} \leq C_1 \, \mathscr{E}^2_{\mathscr{T}}(U, \mathscr{R}).$$

3 To deal with oscillation we decompose the elements of $\mathscr{T}$ into two disjoint sets: $\mathscr{R}$ and $\mathscr{T} \setminus \mathscr{R}$. In the former case, we have

$$\operatorname{osc}^2_{\mathscr{T}}(U, \mathscr{R}) - 2 \operatorname{osc}^2_{\mathscr{T}_*}(U_*, \mathscr{R}) \leq \operatorname{osc}^2_{\mathscr{T}}(U, \mathscr{R}) \leq \mathscr{E}^2_{\mathscr{T}}(U, \mathscr{R}),$$

because $\operatorname{osc}_{\mathscr{T}}(U,T) \le \mathscr{E}_{\mathscr{T}}(U,T)$ for all $T \in \mathscr{T}$. On the other hand, we use that $\mathscr{T} \setminus \mathscr{R} = \mathscr{T} \cap \mathscr{T}_*$ and apply Problem 46 in conjunction with Lemma 24 to arrive at

$$\operatorname{osc}^2_{\mathscr{T}}(U, \mathscr{T} \setminus \mathscr{R}) - 2\operatorname{osc}^2_{\mathscr{T}_*}(U_*, \mathscr{T} \setminus \mathscr{R}) \le C_3 \, |||U - U_*|||^2_{\Omega} \le C_1 C_3 \mathscr{E}^2_{\mathscr{T}}(U, \mathscr{R}).$$

Adding these two estimates gives

$$\operatorname{osc}^2_{\mathscr{T}}(U) - 2\operatorname{osc}^2_{\mathscr{T}_*}(U_*) \le (1 + C_1 C_3) \mathscr{E}^2_{\mathscr{T}}(U, \mathscr{R}).$$

4 Returning to 1 we realize that

$$(1 - 2\mu) C_2 \, \mathscr{E}^2_{\mathscr{T}}(U, \mathscr{T}) \le \big(1 + C_1(1 + C_3)\big) \, \mathscr{E}^2_{\mathscr{T}}(U, \mathscr{R}),$$

which is the asserted estimate (131) in disguise. □

Lemma 28 (Cardinality of $\mathscr{M}_k$: general data). *Let Assumptions 2 and 4 hold. If $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$, then*

$$\#\mathscr{M}_k \lesssim |u, f, \boldsymbol{A}|_s^{1/s} \big(|||u - U_k|||_{\Omega} + \operatorname{osc}_k(U_k) \big)^{-1/s} \quad \forall k \ge 0. \tag{132}$$

Proof. We split the proof into three steps.
1 We set $\varepsilon^2 := \mu \Lambda_2^{-1} \big(|||u - U_k|||^2_{\Omega} + \operatorname{osc}^2_k(U_k) \big)$ with $\mu = \frac{1}{2}\big(1 - \frac{\theta^2}{\theta_*^2}\big) > 0$ as in Lemma 27 and Λ_2 given Lemma 21. Since $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$, in view of Problem 51 there exists $\mathscr{T}_\varepsilon \in \mathbb{T}$ and $U_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that

$$|||u - U_\varepsilon|||^2_{\Omega} + \operatorname{osc}^2_\varepsilon(U_\varepsilon) \le \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u, f, \boldsymbol{A}|_s^{1/2} \, \varepsilon^{-1/s}.$$

Since $\mathscr{T}_\varepsilon$ may be totally unrelated to $\mathscr{T}_k$ we introduce the overlay $\mathscr{T}_* = \mathscr{T}_k \oplus \mathscr{T}_\varepsilon$.
2 We claim that the total error over $\mathscr{T}_*$ reduces by a factor μ relative to that one over $\mathscr{T}_k$. In fact, since $\mathscr{T}_* \ge \mathscr{T}_\varepsilon$ and so $\mathbb{V}(\mathscr{T}_*) \supset \mathbb{V}(\mathscr{T}_\varepsilon)$, we use Lemma 21 to obtain

$$\begin{aligned} |||u - U_*|||^2_{\Omega} + \operatorname{osc}^2_{\mathscr{T}_*}(U_*) &\le \Lambda_2 \Big(|||u - U_\varepsilon|||^2_{\Omega} + \operatorname{osc}^2_\varepsilon(U_\varepsilon) \Big) \\ &\le \Lambda_2 \varepsilon^2 = \mu \, \big(|||u - U_k|||^2_{\Omega} + \operatorname{osc}^2_k(U_k) \big). \end{aligned}$$

Upon applying Lemma 27 we conclude that the set $\mathscr{R} = \mathscr{R}_{\mathscr{T}_k \to \mathscr{T}_*}$ of refined elements satisfies a Dörfler marking (131) with parameter $\theta < \theta_*$.
3 According to Assumption 2, MARK selects a minimal set $\mathscr{M}_k$ satisfying this property. Therefore, we deduce

$$\#\mathscr{M}_k \le \#\mathscr{R} \le \#\mathscr{T}_* - \#\mathscr{T}_k \le \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \lesssim |u, f, \boldsymbol{A}|_s^{1/s} \, \varepsilon^{-1/s},$$

where we have employed Lemma 22 to account for the cardinality of the overlay. Finally, recalling the definition of ε we end up with the asserted estimate (132). □

Remark 15 (Blow-up of constant). The constant hidden in (132) blows up as $\theta \uparrow \theta_*$ because $\mu \downarrow 0$; see Problem 54.

We are ready to prove the main result of this section, which combines Theorem 9 and Lemma 28.

Theorem 10 (Quasi-optimality: general data). *Let Assumptions 2, 3 and 4 hold. If $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$, then AFEM gives rise to a sequence $(\mathscr{T}_k, \mathbb{V}_k, U_k)_{k=0}^{\infty}$ such that*

$$|||u - U_k|||_{\Omega} + \operatorname{osc}_k(U_k) \lesssim |u, f, \boldsymbol{A}|_s \, (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s} \quad \forall k \geq 1.$$

Proof. [1] Since no confusion arises, we use the notation $\operatorname{osc}_j = \operatorname{osc}_j(U_j)$ and $\mathscr{E}_j = \mathscr{E}_j(U_j)$. In light of Assumption 3, which yields Theorem 1, and (132) we have

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \lesssim \sum_{j=0}^{k-1} \#\mathscr{M}_j \lesssim |u, f, \boldsymbol{A}|_s^{1/s} \sum_{j=0}^{k-1} \left(|||u - U_j|||_{\Omega}^2 + \operatorname{osc}_j^2 \right)^{-1/(2s)}.$$

[2] Let $\gamma > 0$ be the scaling factor in the (contraction) Theorem 9. The lower bound (94b) along with $\operatorname{osc}_j \leq \mathscr{E}_j$ implies

$$|||u - U_j|||_{\Omega}^2 + \gamma \operatorname{osc}_j^2 \leq |||u - U_j|||_{\Omega}^2 + \gamma \mathscr{E}_j^2 \leq \left(1 + \frac{\gamma}{C_2}\right) \left(|||u - U_j|||_{\Omega}^2 + \operatorname{osc}_j^2 \right).$$

[3] Theorem 9 yields for $0 \leq j < k$

$$|||u - U_k|||_{\Omega}^2 + \gamma \mathscr{E}_k^2 \leq \alpha^{2(k-j)} \left(|||u - U_j|||_{\Omega}^2 + \gamma \mathscr{E}_j^2 \right),$$

whence

$$\#\mathscr{T}_k - \#\mathscr{T}_0 \lesssim |u, f, \boldsymbol{A}|_s^{1/s} \left(|||u - U_k|||_{\Omega}^2 + \gamma \mathscr{E}_k^2 \right)^{-1/(2s)} \sum_{j=0}^{k-1} \alpha^{(k-j)/s}.$$

Since $\sum_{j=0}^{k-1} \alpha^{(k-j)/s} = \sum_{j=1}^{k} \alpha^{j/s} < \sum_{j=1}^{\infty} \alpha^{j/s} < \infty$ because $\alpha < 1$, the assertion follows immediately. □

We conclude this section with several applications of Theorem 10.

Corollary 10 (Estimator decay). *Let Assumptions 2,, 3 and 4 be satisfied. If $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ then the estimator $\mathscr{E}_k(U_k)$ satisfies*

$$\mathscr{E}_k(U_k) \lesssim |u, f, \boldsymbol{A}|_s^{1/s} \, (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-s} \quad \forall k \geq 1.$$

Proof. Use (114) and Theorem 10. □

Corollary 11 (W_p^2-regularity with piecewise linear $\boldsymbol{A}$). *Let $d = 2$, the polynomial degree $n = 1$, $f \in L^2(\Omega)$, and let $\boldsymbol{A}$ be piecewise linear over $\mathscr{T}_0$. If $u \in W_p^2(\Omega; \mathscr{T}_0)$ for $p > 1$, then AFEM gives rise to a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ satisfying $\operatorname{osc}_k(U_k) = \|h_k(f - P_0 f)\|_{L^2(\Omega)}$ and for all $k \geq 1$*

$$|||u - U_k|||_{\Omega} + \operatorname{osc}_k(U_k) \lesssim \left(\|D^2 u\|_{L^p(\Omega; \mathscr{T}_0)} + \|f\|_{L^2(\Omega)} \right) (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-1/2}.$$

Proof. Combine Corollary 7 with Theorem 10. □

Corollary 12 (W_p^2-regularity with variable $\boldsymbol{A}$). *Assume the setting of Corollary 11, but let $\boldsymbol{A}$ be piecewise Lipschitz over the initial grid $\mathscr{T}_0$. Then AFEM gives rise to a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ satisfying for all $k \geq 1$*

$$|||u-U_k|||_{\Omega} + \operatorname{osc}_k(U_k) \lesssim \Big(\|D^2 u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)} \Big) (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-1/2}.$$

Proof. Combine Corollary 8 with Theorem 10. □

Corollary 13 (W_p^s-regularity with $s < 1/d$). *Let $d \geq 2$, $n = 1$, $1 < t < 2$, $p > 1$, $f \in L^2(\Omega)$, and $\boldsymbol{A} \in W^1_\infty(\Omega, \mathscr{T}_0)$ be piecewise Lipschitz. If $u \in W^t_p(\Omega;\mathscr{T}_0) \cap H^1_0(\Omega)$ is piecewise W^t_p over the initial mesh $\mathscr{T}_0$ with $t - \frac{d}{p} > 1 - \frac{d}{2}$, then AFEM gives rise to a sequence $\{\mathscr{T}_k, \mathbb{V}_k, U_k\}_{k=0}^{\infty}$ satisfying for all $k \geq 1$*

$$|||u-U_k|||_{\Omega} + \operatorname{osc}_k(U_k) \lesssim \Big(\|D^t u\|_{L^p(\Omega;\mathscr{T}_0)} + \|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W^1_\infty(\Omega;\mathscr{T}_0)} \Big) (\#\mathscr{T}_k - \#\mathscr{T}_0)^{-(t-1)/d}.$$

Proof. Combine Corollary 9 with Theorem 10. □

7.5 Extensions and Restrictions

We conclude with a brief discussion of extensions of the theory and some of its restrictions.

Optimal Complexity: Inexact Solvers, Quadrature, and Storage. We point out that we have never mentioned the notion of *complexity* so far. This is because complexity estimates entail crucial issues that we have ignored: inexact solvers to approximate the Galerkin solution; quadrature; and optimal storage. We comment on them now.

Multilevel solvers are known to deliver an approximate solution with cost proportional to the number of degrees of freedom. Even though the theory is well developed for uniform refinement, it is much less understood for adaptive refinement. This is due to the fact that the adaptive bisection meshes do not satisfy the so-called nested refinement assumption. Recently, Xu, Chen, and Nochetto [60] have bridged the gap between graded and quasi-uniform grids exploiting the geometric structure of bisection grids and a resulting new space decomposition. They designed and analyzed optimal additive and multiplicative multilevel methods for any dimension $d \geq 2$ and poynomial degree $n \geq 1$, thereby improving upon Wu and Chen [59]. The theories of §5 and §7 can be suitably modified to account for optimal iterative solvers; we refer to Stevenson [52].

Quadrature is a very delicate issue in a purely a posteriori context, that is without a priori knowledge of the functions involved. Even if we were to replace both data f and $\boldsymbol{A}$ by piecewise polynomials so that quadrature would be simple, we would need to account for the discrepancy in adequate norms between exact and approximate data, again a rather delicate matter. This issue is to a large extend open.

Optimal storage is an essential, but often disregarded, aspect of a complexity analysis. For instance, ALBERTA is an excellent library for AFEM but does not have optimal storage capabilities [50].

Non-Residual Estimators. The cardinality analysis of this section extends to estimators other than the residual; we refer to Cascón and Nochetto [15] and Kreuzer and Siebert [35]. They include the hierarchical, Zienkiewicz-Zhu [2, 27, 55, 58], and Braess-Schoerbel estimators, as well as those based on the solution of local problems [12, 42]. Even though the contraction property of Theorem 9 is no longer valid between consecutive iterates, it is true after a fixed number of iterations, which is enough for the arguments in Proposition 4 and Theorem 10 to apply. The resulting error estimates possess constants proportional to this gap.

Nonconforming Meshes. Since REFINE exhibits optimal complexity for admissible nonconforming meshes, according to §6.4, and this is the only ingredient where nonconformity might play a role, the theory of this section extends. We refer to Bonito and Nochetto [9].

Discontinuous Galerkin Methods (dG). The study of cardinality for adaptive dG methods is rather technical. This is in part due to the fact that key Lemmas 26 and 28 hinge on mesh overlay, which in turn does not provide control of the level of refinement. This makes it difficult to compare broken energy norms

$$|||v|||^2_{\mathscr{T}} = \|\boldsymbol{A}^{1/2}\nabla v\|^2_{L^2(\Omega;\mathscr{T})} + \|h^{-1/2}\,[\![v]\!]\,\|^2_{L^2(\Sigma)},$$

which contain jump terms with negative powers of the mesh-size over the scheleton Σ of $\mathscr{T}$. Consequently, the monotonicity of energy norms used in Lemmas 26 and 28 is no longer true!

To circumvent this difficulty, Bonito and Nochetto [9] resorted to continuous finite elements $\mathbb{V}^0(\mathscr{T})$ over the (admissible nonconforming) mesh $\mathscr{T}$, which have the same degree as their discontinuous counterpart $\mathbb{V}(\mathscr{T})$. This leads to a cardinality theory very much in the spirit of this section. However, it raises the question whether discontinuous elements deliver a better asymptotic rate over admissible nonconforming meshes. Since this result is of intrinsic interest, we report it now.

Lemma 29 (Equivalence of classes). *Let $\mathbb{A}_s$ be the approximation class using discontinuous elements of degree $\le n$ and $\mathbb{A}^0_s$ be the continuous counterpart. Then, for $0 < s \le n/d$, total errors are equivalent on the same mesh, whence $\mathbb{A}_s = \mathbb{A}^0_s$.*

Proof. We use the notation of Problem 11. Since $\mathbb{V}^0(\mathscr{T}) \subset \mathbb{V}(\mathscr{T})$, the inclusion $\mathbb{A}^0_s \subset \mathbb{A}_s$ is obvious. To prove the converse, we let $(u, f, \boldsymbol{A}) \in \mathbb{A}_s$ and, for $N > \#\mathscr{T}_0$, let $\mathscr{T}_* \in \mathbb{T}_N$ be an admissible nonconforming grid and $U_* \in \mathbb{V}(\mathscr{T}_*)$ be so that

$$|||u-U_*|||_{\mathscr{T}_*}+\operatorname{osc}_{\mathscr{T}_*}(U_*)=\inf_{\mathscr{T}\in\mathbb{T}_N}\inf_{V\in\mathbb{V}(\mathscr{T})}\Big(|||u-V|||_{\mathscr{T}}+\operatorname{osc}_{\mathscr{T}}(V)\Big)\lesssim N^{-s}.$$

Let $I_{\mathscr{T}}:\mathbb{V}(\mathscr{T})\to\mathbb{V}^0(\mathscr{T})$ be the interpolation operator of Problem 11. Since $I_{\mathscr{T}_*}U_*\in\mathbb{V}^0(\mathscr{T}_*)$, if we were able to prove

$$|||u-I_{\mathscr{T}_*}U_*|||_{\mathscr{T}_*}+\operatorname{osc}_{\mathscr{T}_*}(I_{\mathscr{T}_*}U_*)\lesssim N^{-s},$$

then $(u,f,\boldsymbol{A})\in\mathbb{A}_s^0$. Using the triangle inequality, we get

$$|||u-I_{\mathscr{T}_*}U_*|||_{\mathscr{T}_*}\le\|\boldsymbol{A}^{1/2}\nabla(u-U_*)\|_{L^2(\Omega;\mathscr{T}_*)}+\|\boldsymbol{A}^{1/2}\nabla(U_*-I_{\mathscr{T}_*}U^*)\|_{L^2(\Omega;\mathscr{T}_*)},$$

because $[\![u-I_{\mathscr{T}_*}U_*]\!]$ vanish on Σ. Problem 11 implies the estimate

$$\|\boldsymbol{A}^{1/2}\nabla(U_*-I_{\mathscr{T}_*}U^*)\|_{L^2(\Omega;\mathscr{T}_*)}\lesssim\|h^{-1/2}\,[\![U_*]\!]\,\|_{L^2(\Sigma_*)}\le|||u-U_*|||_{\mathscr{T}_*},$$

whence

$$|||u-I_{\mathscr{T}_*}U_*|||_{\mathscr{T}_*}\lesssim|||u-U_*|||_{\mathscr{T}_*}.$$

Since $\|\boldsymbol{A}^{1/2}\nabla(U_*-I_{\mathscr{T}_*}U_*)\|_{L^2(\Omega;\mathscr{T}_*)}\lesssim|||U_*-I_{\mathscr{T}_*}U_*|||_{\mathscr{T}_*}$, the oscillation term can be treated similarly. In fact, Problem 46 adapted to discontinuous functions yields

$$\operatorname{osc}_{\mathscr{T}_*}(I_{\mathscr{T}_*}U_*)\lesssim\operatorname{osc}_{\mathscr{T}_*}(U_*)+|||u-U_*|||_{\mathscr{T}_*}.$$

Coupling the two estimates above, we end up with

$$|||u-I_{\mathscr{T}_*}U_*|||_{\mathscr{T}_*}+\operatorname{osc}_{\mathscr{T}_*}(I_{\mathscr{T}_*}U_*)\lesssim|||u-U_*|||_{\mathscr{T}_*}+\operatorname{osc}_{\mathscr{T}_*}(U_*)\lesssim N^{-s}.$$

Therefore, $(u,f,\boldsymbol{A})\in\mathbb{A}_s^0$ as desired. □

7.6 Notes

The theory presented in this section is rather recent. It started with the breakthrough (112) by Stevenson [52] for vanishing oscillation. If f is variable and $\boldsymbol{A}$ is piecewise constant, then Stevenson extended this idea upon adding an inner loop to handle data oscillation to the usual AFEM. This idea does not extend to the model problem (87) with variable $\boldsymbol{A}$, because the oscillation then depends on the Galerkin solution.

The next crucial step was made by Cascón, Kreuzer, Nochetto, and Siebert [14], who dealt with the notion of total error of §7.1, as previously done by Mekchay and Nochetto [37], and introduced the nonlinear approximation class $\mathbb{A}_s$ of §7.2. They derived the convergence rates of §7.4.

The analysis for nonconforming meshes is due to Bonito and Nochetto [9], who developed this theory in the context of dG methods for which they also derived convergence rates. The study of non-residual estimators is due to Kreuzer and Siebert [35] and Cascón and Nochetto [15].

The theory is almost exclusively devoted to the energy norm, except for the L^2-analysis of Demlow and Stevenson [21], who proved an optimal convergence rate for mildly varying graded meshes. Convergence rates have been proved for Raviart-Thomas mixed FEM by Chen, Holst, and Xu [18].

7.7 Problems

Problem 51 (Alternative definition of $\mathbb{A}_s$). Show that $(v, f, \boldsymbol{A}) \in \mathbb{A}_s$ if and only there exists a constant $\Lambda > 0$ such that for all $\varepsilon > 0$ there exist $\mathscr{T}_\varepsilon \geq \mathscr{T}_0$ conforming and $V_\varepsilon \in \mathbb{V}(\mathscr{T}_\varepsilon)$ such that

$$|||v - V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2(V_\varepsilon) \leq \varepsilon^2 \qquad \text{and} \qquad \#\mathscr{T}_\varepsilon - \#\mathscr{T}_0 \leq \Lambda^{1/s}\, \varepsilon^{-1/s};$$

in this case $|v, f, \boldsymbol{A}|_s \leq \Lambda$. Hint: Let $\mathscr{T}_\varepsilon$ be minimal for $|||v - V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2(V_\varepsilon) \leq \varepsilon^2$. This means that for all $\mathscr{T} \in \mathbb{T}$ such that $\#\mathscr{T} = \#\mathscr{T}_\varepsilon - 1$ we have $|||v - V_\varepsilon|||_\Omega^2 + \operatorname{osc}_{\mathscr{T}_\varepsilon}^2(V_\varepsilon) > \varepsilon$.

Problem 52 (Lemma 22). Prove that the overlay $\mathscr{T} = \mathscr{T}_1 \oplus \mathscr{T}_2$ is conforming and

$$\#\mathscr{T} \leq \#\mathscr{T}_1 + \#\mathscr{T}_2 - \#\mathscr{T}_0.$$

Hint: for each $T \in \mathscr{T}_0$, consider two cases $\mathscr{T}_1(T) \cap \mathscr{T}_2(T) \neq \emptyset$ and $\mathscr{T}_1(T) \cap \mathscr{T}_2(T) = \emptyset$, where $\mathscr{T}_i(T)$ is the portion of the mesh $\mathscr{T}_i$ contained in T.

Problem 53 (Lemma 24). Prove that if $\mathscr{T}, \mathscr{T}_* \in \mathbb{T}$ satisfy $\mathscr{T}_* \geq \mathscr{T}$, $\mathscr{R} := \mathscr{R}_{\mathscr{T} \to \mathscr{T}_*}$ is the refined set to go from $\mathscr{T}$ to $\mathscr{T}_*$, and $U \in \mathbb{V}$, $U_* \in \mathbb{V}_*$ are the corresponding Galerkin solutions, then

$$|||U_* - U|||_\Omega^2 \leq C_1\, \mathscr{E}_{\mathscr{T}}^2(U, \mathscr{R}).$$

To this end, write the equation fulfilled by $U_* - U \in \mathbb{V}_*$ and use as a test function the local quasi-interpolant $I_{\mathscr{T}}(U_* - U)$ of $U_* - U$ introduced in Proposition 2.

Problem 54 (Explicit dependence on θ and s). Trace the dependence on θ and s, as $\theta \to \theta_*$ and $s \to 0$, in the hidden constants in Lemma 28 and Theorem 10.

Problem 55 (Asymptotic decay of oscillation). Let $\boldsymbol{A} \in W_\infty^1(\Omega; \mathscr{T}_0)$ be piecewise Lipschitz over the initial grid $\mathscr{T}_0$ and $f \in L^2(\Omega)$. Show that

$$\inf_{\mathscr{T} \in \mathbb{T}_N} \operatorname{osc}_{\mathscr{T}}(U) \lesssim \Big(\|f\|_{L^2(\Omega)} + \|\boldsymbol{A}\|_{W_\infty^1(\Omega; \mathscr{T}_0)} \Big) N^{-1/d}$$

is attained with uniform meshes.

Problem 56 (Faster decay of data oscillation). Let $d = 2$ and $n = 1$. Let f be piecewise W_1^1 over the initial mesh $\mathscr{T}_0$, namely $f \in W_1^1(\Omega; \mathscr{T}_0)$. Show that

$$\inf_{\mathscr{T}\in\mathbb{T}_N} \|h_{\mathscr{T}}(f-P_0 f)\|_{L^2(\Omega)} \lesssim \|f\|_{W_1^1(\Omega;\mathscr{T}_0)} N^{-1},$$

using the thresholding algorithm of §1.6. Therefore, data oscillation decays twice as fast as the energy error asymptotically on suitably graded meshes.

Problem 57 (Faster decay of coefficient oscillation). Consider the coefficient oscillation weighted locally by the energy of the discrete solution U:

$$\eta_{\mathscr{T}}^2(\boldsymbol{A},U) = \sum_{T\in\mathscr{T}} \operatorname{osc}_{\mathscr{T}}^2(\boldsymbol{A},T)\|\nabla U\|_{L^2(\omega_T)}^2,$$

where $\operatorname{osc}_{\mathscr{T}}(\boldsymbol{A},T)$ is defined in Problem 45. Let $d=2, n=1, p>2$, and $\boldsymbol{A}\in W_p^2(\Omega;\mathscr{T}_0)$ be piecewise in W_p^2 over the initial grid $\mathscr{T}_0$. Use the thresholding algorithm of §1.6 to show that $\eta_{\mathscr{T}}(\boldsymbol{A},U)$ decays with a rate twice as fast as the energy error:

$$\inf_{\mathscr{T}\in\mathbb{T}_N} \eta_{\mathscr{T}}(\boldsymbol{A},U) \lesssim \|\boldsymbol{A}\|_{W_p^2(\Omega;\mathscr{T}_0)} \|\nabla U\|_{L^2(\Omega)} N^{-1}.$$

Problem 58 (Faster decay of oscillation). Combine Problems 29, 56 and 57 for $d=2, n=1$ and $p>2$ to prove that if $f\in W_1^1(\Omega;\mathscr{T}_0)$ and $\boldsymbol{A}\in W_p^2(\Omega;\mathscr{T}_0)$, then the oscillation $\operatorname{osc}_{\mathscr{T}}(U,\mathscr{T})$ decays with a rate twice as fast as the energy error:

$$\inf_{\mathscr{T}\in\mathbb{T}_N} \operatorname{osc}_{\mathscr{T}}(U) \lesssim \Big(\|f\|_{W_1^1(\Omega;\mathscr{T}_0)} + \|\boldsymbol{A}\|_{W_p^2(\Omega;\mathscr{T}_0)}\Big) N^{-1}.$$

References

1. M. AINSWORTH AND D. W. KELLY, *A posteriori error estimators and adaptivity for finite element approximation of the non-homogeneous Dirichlet problem*, Adv. Comput. Math., 15 (2001), pp. 3–23.
2. M. AINSWORTH AND J.T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience, 2000.
3. I. BABUŠKA, R.B. KELLOGG, AND J. PITKÄRANTA, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.
4. I. BABUŠKA AND A. MILLER, *A feedback finite element method with a posteriori error estimation. I. The finite element method and some basic properties of the a posteriori error estimator*, Comput. Methods Appl. Mech. Engrg. 61 (1) (1987), pp. 1–40.
5. I. BABUŠKA AND W. RHEINBOLDT, *Error estimates for adaptive finite element computations* SIAM J. Numer. Anal., 15 (1978), pp. 736–754.
6. E. BÄNSCH, P. MORIN, AND R. H. NOCHETTO, *An adaptive Uzawa FEM for the Stokes problem: convergence without the inf-sup condition*, SIAM J. Numer. Anal., 40 (2002), pp. 1207–1229 (electronic).
7. P. BINEV, W. DAHMEN, AND R. DEVORE, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.
8. P. BINEV, W. DAHMEN, R. DEVORE, AND P. PETRUSHEV, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 391–416. Dedicated to the memory of Vassil Popov on the occasion of his 60th birthday.
9. A. BONITO AND R.H. NOCHETTO, *Quasi-optimal convergence rate for an adaptive discontinuous Galerkin method*, (submitted).
10. D. BRAESS, *Finite Elements. Theory, fast solvers, and applications in solid mechanics*, 2nd edition. Cambridge University Press (2001).
11. S. BRENNER AND L.R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer Texts in Applied Mathematics 15 (2008).
12. C. CARSTENSEN AND S. A. FUNKEN, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21 (1999), pp. 1465–1484.
13. C. CARSTENSEN AND R.H.W. HOPPE, *Error reduction and convergence for an adaptive mixed finite element method*, Math. Comp. 75 (2006), 1033–1042.
14. J. M. CASCÓN, C. KREUZER, R. H. NOCHETTO, AND K. G. SIEBERT, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal., 46 (2008), pp. 2524–2550.
15. J. M. CASCÓN AND R. H. NOCHETTO, *Convergence and quasi-optimality for AFEM based on non-residual a posteriori error estimators*, (in preparation).
16. J. M. CASCÓN, R. H. NOCHETTO, AND K. G. SIEBERT, *Design and convergence of AFEM in $H(\mathrm{div})$*, Math. Models Methods Appl. Sci., 17 (2007), pp. 1849–1881.
17. Z. CHEN AND J. FENG, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2006), pp. 1167–1042.
18. L. CHEN, M. HOLST, AND J. XU, *Convergence and optimality of adaptive mixed finite element methods*, Math. Comp. 78 (2009), 35–53.
19. P.G. CIARLET, *The Finite Element Method for Elliptic Problems,* Classics in Applied Mathematics, 40, SIAM (2002).
20. A. DEMLOW, *Convergence of an adaptive finite element method for controlling local energy errors*, (submitted).
21. A.. DEMLOW AND R.P. STEVENSON, *Convergence and quasi-optimality of an adaptive finite element method for controlling L_2 errors*, (submitted).
22. R.A. DEVORE, *Nonlinear approximation*, A. Iserles (ed.) Acta Numerica, vol. 7, pp. 51–150. Cambridge University Press (1998).
23. L. DIENING AND CH. KREUZER, *Convergence of an adaptive finite element method for the p-Laplacian equation*, SIAM J. Numer. Anal., 46 (2008), pp. 614–638.

24. W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
25. W. DÖRFLER AND M. RUMPF, *An adaptive strategy for elliptic problems including a posteriori controlled boundary approximation*, Math. Comp., 67 (1998), pp. 1361–1382.
26. T. DUPONT AND L.R. SCOTT, *Polynomial approximation of functions in Sobolev spaces*, Math. Comp., 34 (1980), pp. 441–463.
27. F. FIERRO AND A. VEESER, *A posteriori error estimators, gradient recovery by averaging, and superconvergence*, Numer. Math., 103 (2006), pp. 267-298.
28. E.M. GARAU, P. MORIN, C. ZUPPA, *Convergence of adaptive finite element methods for eigenvalue problems*, Math. Models Methods Appl. Sci. 19 (2009), pp. 721–747.
29. F. GASPOZ AND P. MORIN, *Approximation classes for adaptive higher order finite element approximation*, (in preparation), (2010).
30. P. GRISVARD, *Elliptic Problems in Nonsmooth Domains, Monographs and Studies in Mathematics*, vol. 24. Pitman (Advanced Publishing Program), Boston, MA (1985).
31. M. HOLST, G. TSOGTGEREL, AND Y. ZHU, *Local convergence of adaptive methods for nonlinear partial differential equations*, preprint arXiv:math.NA/1001.1382 (2009).
32. R. H. W. HOPPE, G. KANSCHAT, AND T. WARBURTON, *Convergence analysis of an adaptive interior penalty discontinuous Galerkin method*, SIAM J. Numer. Anal., 47 (2009), pp. 534–550.
33. O.A. KARAKASHIAN AND F. PASCAL *Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems*, SIAM J. Numer. Anal. 45 (2007), 641–665.
34. R.B. KELLOGG, *On the Poisson equation with intersecting interfaces*, Applicable Anal., 4 (1974/75), 101–129.
35. CH. KREUZER AND K.G. SIEBERT, *Decay rates of adaptive finite elements with Dörfler marking*, (submitted).
36. J.M. MAUBACH, *Local bisection refinement for n-simplicial grids generated by reflection*, SIAM J. Sci. Comput. 16 (1995), 210–227.
37. K. MEKCHAY AND R. H. NOCHETTO, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827 (electronic).
38. K. MEKCHAY, P. MORIN, AND R. H. NOCHETTO, *AFEM for the Laplace-Beltrami operator on graphs: design and conditional contraction property*, Math. Comp. (to appear).
39. W.F. MITCHELL, *Unified multilevel adaptive finite element methods for elliptic problems*, Ph.D. thesis, Department of Computer Science, University of Illinois, Urbana (1988)
40. P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
41. P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Convergence of adaptive finite element methods*, SIAM Review, 44 (2002), pp. 631–658.
42. P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Local problems on stars: a posteriori error estimators, convergence, and performance*, Math. Comp., 72 (2003), pp. 1067–1097 (electronic).
43. P. MORIN, K. G. SIEBERT, AND A. VEESER, *Convergence of finite elements adapted for weak norms*, in Applied and Industrial Mathematics in Italy II, Vinvenzo Cutello, Giorgio Fotia, Luigia Puccio eds, Series on Advances in Mathematics for Applied Sciences, 75 (2007), pp. 468–479.
44. P. MORIN, K. G. SIEBERT, AND A. VEESER, *A basic convergence result for conforming adaptive finite elements*, Math. Mod. Meth. Appl. Sci., 5 (2008), pp. 707–737.
45. R.H. NOCHETTO, K. G. SIEBERT, AND A. VEESER, *Theory of adaptive finite element methods: an introduction*, in *Multiscale, Nonlinear and Adaptive Approximation*, Springer, 2009, pp. 409–542.
46. R. H. NOCHETTO, A. SCHMIDT, K. G. SIEBERT, AND A. VEESER, *Pointwise a posteriori error estimates for monotone semi-linear equations*, Numer. Math., 104 (2006), pp. 515-538.
47. R. SACCHI AND A. VEESER, *Locally efficient and reliable a posteriori error estimators for Dirichlet problems*, Math. Models Methods Appl., 16 (2006), pp. 319–346.

48. L.R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
49. K.G. SIEBERT, *A convergence proof for adaptive finite elements without lower bound*, Preprint Universität Duisburg-Essen and Universität Freiburg No. 1/2009.
50. K.G. SIEBERT, *Mathematically founded design of adaptive finite element software*, in Multiscale and Adaptivity: Modeling, Numerics and Applications, CIME-EMS Summer School in Applied Mathematics, G. Naldi and G. Russo eds., Springer (2010).
51. K.G. SIEBERT AND A. VEESER, *A unilaterally constrained quadratic minimization with adaptive finite elements*, SIAM J. Optim., 18 (2007), pp. 260–289.
52. R. STEVENSON, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.
53. R. STEVENSON, *The completion of locally refined simplicial partitions created by bisection*, Math. Comput., 77 (2008), pp. 227–241.
54. C.T. TRAXLER, *An algorithm for adaptive mesh refinement in n dimensions*, Computing 59 (1997), 115–137.
55. A. VEESER, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math. 92 (2002), pp. 743–770.
56. A. VEESER AND R. VERFÜRTH, *Explicit upper bounds for dual norms of residuals*, SIAM J. Numer. Anal., 47 (2009), pp. 2387–2405.
57. R. VERFÜRTH, *A posteriori error estimators for the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.
58. R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Adv. Numer. Math. John Wiley, Chichester, UK (1996).
59. H. WU AND Z. CHEN, *Uniform convergence of multigrid V-cycle on adaptively refined finite element meshes for second order elliptic problems*, Science in China: Series A Mathematics, 49 (2006), pp. 1–28.
60. J. XU, L. CHEN, AND R.H. NOCHETTO, *Adaptive multilevel methods on graded bisection grids*, in *Multiscale, Nonlinear and Adaptive Approximation*, Springer, 2009, pp. 599–659.