

## QUASI-OPTIMAL CONVERGENCE RATE FOR AN ADAPTIVE FINITE ELEMENT METHOD\*

J. MANUEL CASCON<sup>†</sup>, CHRISTIAN KREUZER<sup>‡</sup>, RICARDO H. NOCHETTO<sup>§</sup>, AND  
KUNIBERT G. SIEBERT<sup>‡</sup>

**Abstract.** We analyze the simplest and most standard adaptive finite element method (AFEM), with any polynomial degree, for general second order linear, symmetric elliptic operators. As is customary in practice, the AFEM marks exclusively according to the error estimator and performs a minimal element refinement without the interior node property. We prove that the AFEM is a contraction, for the sum of the energy error and the scaled error estimator, between two consecutive adaptive loops. This geometric decay is instrumental to derive the optimal cardinality of the AFEM. We show that the AFEM yields a decay rate of the energy error plus oscillation in terms of the number of degrees of freedom as dictated by the best approximation for this combined nonlinear quantity.

**Key words.** error reduction, convergence, optimal cardinality, adaptive algorithm

**AMS subject classifications.** 65N30, 65N50, 65N15, 65N12, 41A25

**DOI.** 10.1137/07069047X

**1. Introduction.** Let  $\Omega$  be a bounded, polyhedral domain in  $\mathbb{R}^d$ ,  $d \geq 2$ . We consider a homogeneous Dirichlet boundary value problem for a self-adjoint second order elliptic partial differential equation (PDE)

$$(1.1) \quad \begin{aligned} \mathcal{L}u &:= -\operatorname{div}(\mathbf{A}\nabla u) + cu = f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

The choice of boundary condition is made for ease of presentation, since similar results are valid for other boundary conditions. Precise conditions on given data  $\mathbf{D} := (\mathbf{A}, c)$  and  $f$  of  $\mathcal{L}$  are stated in section 2.1. We refer to [5] for more general operators.

We analyze here a standard adaptive finite element method (AFEM) of the form

$$(1.2) \quad \text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}.$$

Even though adaptivity has been a fundamental tool of engineering and scientific computing for about three decades, the convergence analysis is rather recent. It started with Dörfler [8], who introduced a crucial marking, from now on called

---

\*Received by the editors May 3, 2007; accepted for publication (in revised form) January 15, 2008; published electronically DATE.

<http://www.siam.org/journals/sinum/x-x/69047.html>

<sup>†</sup>Departamento de Matemáticas, Universidad de Salamanca, 37008 Salamanca, Spain (casbar@usal.es, <http://matematicas.fis.usal.es/~casbar>). This author's research was partially supported by grants CGL2004-06171-C03-03/CLI, Ministerio de Ciencia y Tecnología (Spain), and SA078A05, Junta de Castilla y León (Spain), FEDER funds (European Union), and NSF grant DMS-0505454.

<sup>‡</sup>Institut für Mathematik, Universität Augsburg, Universitätsstraße 14, 86159 Augsburg, Germany (kreuzer@math.uni-augsburg.de, [http://www.math.uni-augsburg.de/de/prof/lam/mitarbeiter/christian\\_kreuzer](http://www.math.uni-augsburg.de/de/prof/lam/mitarbeiter/christian_kreuzer), siebert@math.uni-augsburg.de, <http://scicomp.math.uni-augsburg.de/siebert/>). The research of these authors was partially supported by DAAD grant "Efficient Finite Element Methods for Solid and Fluid Mechanics Computations."

<sup>§</sup>Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu, <http://www.math.umd.edu/~rhn>). This author's research was partially supported by NSF grant DMS-0505454.

Dörfler’s marking, and proved strict energy error reduction for the Laplacian provided the initial mesh  $\mathcal{T}_0$  satisfies a fineness assumption. Morin, Nochetto, and Siebert [17, 18] showed that such strict energy error reduction cannot be expected in general. By introducing the concept of data oscillation and the interior node property, they proved convergence of the AFEM without restrictions on  $\mathcal{T}_0$ . The latter result, however, is valid only for  $\mathbf{A}$  in (1.1) piecewise constant on  $\mathcal{T}_0$  and vanishing  $c$ . Inspired by the work by Chen and Feng [6], Mekchay and Nochetto [16] extended this result to general second order elliptic operators upon dealing with the new concept of *total error*, namely, the sum of the energy error plus oscillation. They proved that the AFEM is a contraction for the total error, a property that will turn out to be essential in this paper as well. Recently, Diening and Kreuzer [7] proved a similar property for the  $p$ -Laplacian but avoided marking for oscillation by using the fact that oscillation appearing in the lower bound is dominated by the estimator.

In this paper we go back to the basis of the AFEM philosophy and consider the simplest possible approach to adaptivity for (1.1). In section 2 we first state precisely the problem and describe the modules of the AFEM: **SOLVE** computes the Ritz–Galerkin approximation; i.e., we assume an exact solution and integration; **ESTIMATE** computes the standard residual estimator; **MARK** resorts to Dörfler marking based solely on the estimator; and **REFINE** utilizes bisectioning of elements with the minimal refinement condition that marked elements are bisected at least once. In this respect the AFEM is a really standard algorithm in that it avoids marking for oscillation and circumvents the interior node property of [16, 17, 18] for marked elements. We would like to stress that removing these ingredients is very important from a practical point of view. The interior node property enforces six bisections of marked elements in three space dimensions and thus increases the number of DOFs between two iterations drastically. Additionally, since oscillation is not used by the AFEM, it does not have to be computed. Although computing time for oscillation may be negligible, this strongly improves on implementation requirements.

The first convergence result for such a standard AFEM is due to Morin, Siebert, and Veiser [20]. They proved recently, for a larger problem class and more general marking strategies, plain convergence of the AFEM but without an error reduction property. By relying on Dörfler marking we are able to prove a stronger result for self-adjoint elliptic operators of the form (1.1), namely, contraction of the quasi-error and quasi-optimal cardinality of the AFEM.

To summarize the main results, let  $\{\mathcal{T}_k, \mathbb{V}_k, U_k, \eta_k, \text{osc}_k\}_{k \geq 0}$  be the sequence of meshes, finite element spaces, discrete solutions, estimators, and oscillations produced by the AFEM in the  $k$ th step. Even though the energy error is monotone, strict error reduction fails when  $U_{k+1} = U_k$ , i.e.,  $\|u - U_{k+1}\|_\Omega = \|u - U_k\|_\Omega$ ; see [17, 18] for further details. On the other hand, the residual estimator  $\eta_k = \eta_k(U_k)$  exhibits a strict reduction when  $U_{k+1} = U_k$  but no monotone behavior in general; this is shown in section 3. The new insight of this paper is that the sum of the energy error and the scaled estimator, the so-called *quasi error*  $(\|u - U_k\|_\Omega^2 + \gamma \eta_k^2)^{1/2}$ , is strictly reduced by the AFEM even though each term may not be. In fact, we prove in section 4 that the AFEM is a contraction for this new error notion (see Theorem 4.1).

**MAIN RESULT 1.** There exist constants  $\gamma > 0$  and  $0 < \alpha < 1$  such that

$$\|u - U_{k+1}\|_\Omega^2 + \gamma \eta_{k+1}^2 \leq \alpha^2 \left( \|u - U_k\|_\Omega^2 + \gamma \eta_k^2 \right).$$

Quasi-optimal convergence rates for the AFEM, expressing energy error decay in terms of the number of DOFs as dictated by nonlinear approximation theory, were

first proved by Binev, Dahmen, and DeVore [3]. They resorted to a crucial, but somewhat artificial, coarsening step. Coarsening was later removed by Stevenson [23], who developed an optimality theory for a much more realistic AFEM but still included an inner loop to deal with oscillation. Both papers [3, 23] are restricted to the Laplace operator and rely on suitable marking by oscillation and the interior node property.

To derive convergence rates we need to seek a suitable error quantity and an associated approximation class  $\mathbb{A}_s$ . Since all decisions of the AFEM in MARK are based on the estimator  $\eta_k$ , a decay rate for the true error is closely related to the quality of the estimator, which is described by the *global lower bound*

$$\eta_k^2 \preceq \|u - U_k\|_{\Omega}^2 + \text{osc}_k^2;$$

hereafter, *total error* refers to the square root of the right-hand side (see [16]). The lower bound states that the estimator is dominated by the error up to an oscillation term and overestimation may occur only when oscillation is large. Furthermore we observe that the total error is bounded again by the estimator—thanks to the upper bound  $\|u - U_k\|_{\Omega} \preceq \eta_k$  and  $\text{osc}_k \leq \eta_k$ —and the estimator itself is an upper bound for the quasi error, which is the quantity reduced by the AFEM:

$$\|u - U_k\|_{\Omega}^2 + \text{osc}_k^2 \approx \eta_k^2 \approx \|u - U_k\|_{\Omega}^2 + \gamma \eta_k^2.$$

In summary, the performance of the AFEM is intrinsically bonded to the total error, which measures both the approximability of  $u$  as well as of data, encoded in the oscillation term. For piecewise polynomial coefficients  $(\mathbf{A}, c)$ , oscillation reduces to approximation of the right-hand side  $f$  of (1.1). In the general case, approximation of data  $(\mathbf{A}, c)$  in  $\text{osc}_k$  couples in a nonlinear fashion with the discrete solution  $U_k$ . In section 5 we start with the definition of an approximation class  $\mathbb{A}_s$  based on the total error. Roughly speaking,  $(u, f, \mathbf{D})$  belongs to  $\mathbb{A}_s$  if the total error can be approximated within any tolerance  $\epsilon > 0$  with  $O(\epsilon^{-s})$  DOFs. We then prove the *quasi-optimal* convergence rate for the AFEM in terms of DOFs by assuming certain restrictions on the initial triangulation and the marking parameter  $\theta \in (0, \theta_*)$  (see Theorem 5.11).

**MAIN RESULT 2.** If  $(u, f, \mathbf{D}) \in \mathbb{A}_s$ , then there exists a constant  $C$  such that

$$\{\|u - U_k\|_{\Omega}^2 + \text{osc}_k^2\}^{1/2} \leq C (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s}.$$

Since the estimator is equivalent to the total error, we obtain the same decay rate for the estimator  $\eta_k$ . Besides the contraction property of the AFEM, crucial ingredients of the proof are a quasi-monotonicity property of oscillation as well as a localized upper bound, both proved in section 3. The latter refines a similar bound by Stevenson [23]. In contrast to former optimality proofs [3, 23], the present analysis stays within the class of conforming meshes. This is a necessary framework for the analysis when dealing with oscillation in the jump residual.

As already mentioned above, for piecewise polynomial coefficients  $(\mathbf{A}, c)$  one can separate approximation of  $u$  and data  $f$ . When doing so and marking with respect to two quantities such as the estimator and oscillation, the role of marking becomes critical for proving optimality. To shed light on this issue, we conclude this article by discussing in section 6 the simultaneous adaptive approximation of two functions with distinct asymptotic error decays. The discussion reveals that separate marking might be suboptimal and explains why the AFEM designed by Morin, Nochetto, and Siebert [17, 18] is altered in former optimality proofs [3, 23].

**2. Problem and adaptive finite element method.** We first introduce the underlying problem and state assumptions on given data. We then describe the refinement framework and the AFEM along with its modules.

**2.1. Weak formulation.** Let  $\Omega$  be a bounded, polyhedral domain in  $\mathbb{R}^d$ ,  $d \geq 2$ , that is triangulated by a conforming triangulation  $\mathcal{T}_0$ . We assume that data of (1.1) have the following properties:

- (a)  $\mathbf{A}: \Omega \mapsto \mathbb{R}^{d \times d}$  is piecewise Lipschitz over  $\mathcal{T}_0$  and is symmetric positive definite with eigenvalues in  $0 < a_* \leq a^* < \infty$ , i.e.,

$$a_*(x) |\xi|^2 \leq \mathbf{A}(x) \xi \cdot \xi \leq a^*(x) |\xi|^2 \quad \forall \xi \in \mathbb{R}^d, x \in \Omega;$$

- (b)  $c \in L^\infty(\Omega)$  is nonnegative, i.e.,  $c \geq 0$  in  $\Omega$ ;

- (c)  $f \in L^2(\Omega)$ .

Now we turn to the weak formulation of (1.1). For any set  $\omega \subset \mathbb{R}^d$  with a nonempty interior we denote by  $H^1(\omega)$  the usual Sobolev space of functions in  $L^2(\omega)$  whose first derivatives are also in  $L^2(\omega)$ , endowed with the norm

$$\|u\|_{H^1(\omega)} := \left( \|u\|_{L^2(\omega)}^2 + \|\nabla u\|_{L^2(\omega)}^2 \right)^{1/2}.$$

Moreover, we denote by  $\langle \cdot, \cdot \rangle_\omega$  the  $L^2(\omega)$  scalar product. Finally we let  $\mathbb{V} := H_0^1(\Omega)$  be the space of functions in  $H^1(\Omega)$  with vanishing trace on  $\partial\Omega$ . A weak solution of (1.1) is a function  $u$  satisfying

$$(2.1) \quad u \in \mathbb{V} : \quad \mathcal{B}[u, v] = \langle f, v \rangle_\Omega \quad \forall v \in \mathbb{V},$$

where the bilinear form is defined to be

$$\mathcal{B}[u, v] := \langle \mathbf{A} \nabla u, \nabla v \rangle_\Omega + \langle c u, v \rangle_\Omega \quad \forall u, v \in \mathbb{V}.$$

In view of the Poincaré–Friedrichs inequality, one has *coercivity* on  $\mathbb{V}$

$$\mathcal{B}[v, v] = \langle \mathbf{A} \nabla v, \nabla v \rangle_\Omega + \langle c v, v \rangle_\Omega \geq \int_\Omega a_* |\nabla v|^2 + c v^2 \geq c_B \|v\|_{H^1(\Omega)}^2,$$

and  $c_B$  depends only on data and  $\Omega$ . The bilinear form  $\mathcal{B}$  induces the so-called *energy norm*:

$$\|v\|_\Omega := \mathcal{B}[v, v]^{1/2} \quad \forall v \in \mathbb{V}.$$

Note that the bilinear form also fulfills the *local continuity*

$$\mathcal{B}[v, w] \leq \sqrt{C_B} \|v\|_\Omega \|w\|_{H^1(\omega)} \quad \forall v, w \in H_0^1(\Omega), \text{supp}(w) \subset \omega \subset \Omega,$$

with  $\|v\|_\omega := \langle \mathbf{A} \nabla v, \nabla v \rangle_\omega + \langle c v, v \rangle_\omega$ . This local continuity is essential in deriving *local lower bounds* in the a posteriori error analysis. Furthermore it implies continuity of  $\mathcal{B}[\cdot, \cdot]$  on  $H^1(\Omega)$  at once. Thanks to the coercivity and continuity of  $\mathcal{B}$ , the norm  $\|\cdot\|_\Omega$  is equivalent to  $\|\cdot\|_{H^1(\Omega)}$  on  $H_0^1(\Omega)$ . The existence and uniqueness of (2.1) thus follow from the Lax–Milgram theorem [9]. The restriction to a symmetric bilinear form and  $c \geq 0$  can be relaxed, provided (2.1) admits a unique solution. This extension will be studied in [5].

**2.2. Refinement framework.** Refinement is based on shape-regular bisection of single elements. Any given simplex is subdivided into two subsimplices of the same size such that the minimal angle is uniformly bounded from below; we refer to [2, 13, 14, 15, 24, 25] or the monograph [21] and the references therein. In two dimensions, this is the newest vertex bisection. Bisectioning creates a unique *master forest*  $\mathbb{F}$  of binary trees with infinite depth, where each node is a simplex, its two successors are the two children created by bisection, and the roots of the binary trees are the elements of the initial conforming triangulation  $\mathcal{T}_0$ . The master forest  $\mathbb{F}$  contains full information of all possible subdivisions created from  $\mathcal{T}_0$  by bisection, i.e., information about vertices, neighbors, refinement edges, etc., for any simplex that can be generated.

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

- (a) all nodes in  $\mathcal{F} \setminus \mathcal{T}_0$  have a predecessor;
- (b) all nodes in  $\mathcal{F}$  have either two successors or none.

Any node of  $\mathcal{F}$  is thus uniquely connected with a node of the initial triangulation  $\mathcal{T}_0$ . Furthermore, any forest may have interior nodes, i.e., nodes with successors, and does have leaf nodes, i.e., nodes without successors. Finally, each forest  $\mathcal{F}$  corresponds one to one to a subdivision  $\mathcal{T}(\mathcal{F})$  of  $\Omega$  into simplices by defining  $\mathcal{T}(\mathcal{F})$  as the set of leaf nodes of  $\mathcal{F}$ . Note that such a subdivision may be nonconforming.

If  $\mathcal{F} \subset \mathcal{F}_*$  are two forests, we call  $\mathcal{T}_* = \mathcal{T}(\mathcal{F}_*)$  a refinement of  $\mathcal{T} = \mathcal{T}(\mathcal{F})$  and denote this by  $\mathcal{T} \leq \mathcal{T}_*$ . We define

$$\mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*} := \mathcal{T} \setminus (\mathcal{T}_* \cap \mathcal{T})$$

as the set of *refined elements*; i.e.,  $\mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$  is the set of leaf nodes of  $\mathcal{F}$  that are interior nodes of  $\mathcal{F}_*$ . The class of all *conforming* refinements by bisection of  $\mathcal{T}_0$  is

$$\mathbb{T} := \{\mathcal{T}(\mathcal{F}) \mid \mathcal{F} \subset \mathbb{F} \text{ is a forest and } \mathcal{T}(\mathcal{F}) \text{ is a conforming triangulation of } \Omega\}.$$

Given two forests  $\mathcal{F}, \mathcal{F}_* \subset \mathbb{F}$  corresponding to subdivisions  $\mathcal{T}$  and  $\mathcal{T}_*$  of  $\Omega$ , we define  $\mathcal{F} \cup \mathcal{F}_*$  to be the union of the nodes of  $\mathcal{F}$  and  $\mathcal{F}_*$ . Obviously,  $\mathcal{F} \cup \mathcal{F}_* \subset \mathbb{F}$  and  $\mathcal{T}_0 \subset \mathcal{F} \cup \mathcal{F}_*$ . By construction, all nodes of  $(\mathcal{F} \cup \mathcal{F}_*) \setminus \mathcal{T}_0$  have a predecessor, and all nodes of the union have either two successors or none: this implies that  $\mathcal{F} \cup \mathcal{F}_*$  is a forest. We call the resulting unique subdivision of  $\Omega$  the *overlay of  $\mathcal{T}$  and  $\mathcal{T}_*$* :

$$\mathcal{T} \oplus \mathcal{T}_* := \mathcal{T}(\mathcal{F} \cup \mathcal{F}_*).$$

For two conforming triangulations  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$  we prove in Lemma 3.7 below that  $\mathcal{T} \oplus \mathcal{T}_* \in \mathbb{T}$  is the smallest conforming refinement of  $\mathcal{T}, \mathcal{T}_*$  and satisfies

$$(2.2) \quad \#(\mathcal{T} \oplus \mathcal{T}_*) \leq \#\mathcal{T} + \#\mathcal{T}_* - \#\mathcal{T}_0.$$

We finally introduce some notations related to triangulations. For  $T \in \mathcal{T}$  we denote by  $h_T := |T|^{\frac{1}{d}}$  the local mesh size and by  $\omega_T$  the union of all elements in  $\mathcal{T}$  sharing one side with  $T$ . We further denote by  $\mathcal{S}$  the skeleton of  $\mathcal{T}$ , i.e., the union of the interelement sides, and for an interior side  $\sigma \in \mathcal{S}$  we let  $\omega_\sigma$  be the union of the two adjacent elements sharing  $\sigma$ .

Thanks to the properties of bisection, all constants depending on shape regularity of  $\mathcal{T} \in \mathbb{T}$  are uniformly bounded by a constant solely depending on  $\mathcal{T}_0$  [14, 25].

**2.3. The module SOLVE.** Given any conforming triangulation  $\mathcal{T} \in \mathbb{T}$ , we define the finite element space

$$\mathbb{V}(\mathcal{T}) := \{V \in \mathbb{V} \mid V|_T \in \mathbb{P}_n(T), T \in \mathcal{T}\},$$

where  $n \in \mathbb{N}$  is a fixed polynomial degree and  $\mathbb{P}_n$  denotes the space of all polynomials of degree  $\leq n$ . Since the continuity and coercivity of  $\mathcal{B}$  are inherited by any subspace of  $\mathbb{V}$ , the Lax–Milgram theorem implies the existence and uniqueness of the Ritz–Galerkin approximation in  $\mathbb{V}(\mathcal{T})$  defined by

$$(2.3) \quad U \in \mathbb{V}(\mathcal{T}) : \quad \mathcal{B}[U, V] = \langle f, V \rangle_\Omega \quad \forall V \in \mathbb{V}(\mathcal{T}).$$

We suppose that the module SOLVE outputs the exact Ritz–Galerkin solution on  $\mathcal{T}$ :

$$U = \text{SOLVE}(\mathcal{T});$$

i.e.,  $U$  is computed via exact linear algebra and exact integration. Optimal multilevel solvers can be incorporated as in [23], but quadrature is more delicate.

We observe that for any pair  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ , with  $\mathcal{T} \leq \mathcal{T}_*$ , there holds  $\mathbb{V}(\mathcal{T}) \subset \mathbb{V}(\mathcal{T}_*)$ ; i.e., the discrete spaces are *nested*. This turns out to be a crucial property in the subsequent analysis in that the following *orthogonality* relation holds:

$$(2.4) \quad \|u - U\|_\Omega^2 = \|u - U_*\|_\Omega^2 + \|U_* - U\|_\Omega^2.$$

This is a consequence of  $\mathcal{B}[u - U_*, U_* - U] = 0$  or, equivalently, that  $u - U_*$  and  $U_* - U$  are orthogonal with respect to the scalar product induced by  $\mathcal{B}[\cdot, \cdot]$ .

**2.4. The module ESTIMATE.** For  $\mathcal{T} \in \mathbb{T}$  and  $V \in \mathbb{V}(\mathcal{T})$  we define the *element residual* and *jump residual* for  $V$  by

$$R(V)|_T := (f + \mathcal{L}V)|_T, \quad T \in \mathcal{T}, \quad \text{and} \quad J(V)|_\sigma := (\llbracket \mathbf{A} \nabla V \rrbracket \cdot \boldsymbol{\nu})|_\sigma, \quad \sigma \in \mathcal{S}.$$

Hereafter,  $\llbracket q \rrbracket$  is the jump of  $q$  across an interior side  $\sigma$ , and  $\boldsymbol{\nu}$  denotes a unit normal vector associate to side  $\sigma$ . The error indicator for  $V$  on  $T \in \mathcal{T}$  is given by

$$(2.5) \quad \eta_{\mathcal{T}}^2(V, T) := h_T^2 \|R(V)\|_{L^2(T)}^2 + h_T \|J(V)\|_{L^2(\partial T \cap \Omega)}^2,$$

where we recall that  $h_T = |T|^{1/d}$ . We assume that, given a triangulation  $\mathcal{T}$  and the Ritz–Galerkin solution  $U \in \mathbb{V}(\mathcal{T})$ , the module ESTIMATE outputs the indicators

$$\{\eta_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}} = \text{ESTIMATE}(U, \mathcal{T}).$$

The lower bound involves oscillation, which we define next. For  $m \in \mathbb{N}_0$ , we denote by  $\Pi_m^p$  the  $L^p$ -best approximation operator onto the set of discontinuous polynomials of degree  $\leq m$  over either  $T \in \mathcal{T}$  or  $\sigma \in \mathcal{S}$ , depending on the context. If  $P_m^p := \text{id} - \Pi_m^p$ , then we define oscillation of  $V \in \mathbb{V}(\mathcal{T})$  to be

$$\text{osc}_{\mathcal{T}}^2(V, T) := h_T^2 \|P_{2n-2}^2 R(V)\|_{L^2(T)}^2 + h_T \|P_{2n-1}^2 J(V)\|_{L^2(\partial T \cap \Omega)}^2.$$

Finally, for any subset  $\mathcal{T}' \subset \mathcal{T}$  we set

$$\eta_{\mathcal{T}'}^2(V, \mathcal{T}') := \sum_{T \in \mathcal{T}'} \eta_{\mathcal{T}}^2(V, T) \quad \text{and} \quad \text{osc}_{\mathcal{T}'}^2(V, \mathcal{T}') := \sum_{T \in \mathcal{T}'} \text{osc}_{\mathcal{T}}^2(V, T).$$

*Remark 2.1.* Let  $\mathcal{T} \in \mathbb{T}$  and  $V \in \mathbb{V}(\mathcal{T})$  be given. We first observe that the indicator  $\eta_{\mathcal{T}}(V, T)$  *dominates* oscillation  $\text{osc}_{\mathcal{T}}(V, T)$ , i.e.,  $\text{osc}_{\mathcal{T}}(V, T) \leq \eta_{\mathcal{T}}(V, T)$  for all  $T \in \mathcal{T}$ . In addition, the definitions of the error indicator and oscillation are fully localized to  $T$ ; i.e., for any triangulation  $\mathcal{T}_* \in \mathbb{T}$ , with  $T \in \mathcal{T}_*$ , there holds

$\eta_{\mathcal{T}}(V, T) = \eta_{\mathcal{T}_*}(V, T)$  and  $\text{osc}_{\mathcal{T}}(V, T) = \text{osc}_{\mathcal{T}_*}(V, T)$ . Moreover, if  $\mathcal{T}_* \geq \mathcal{T}$  is any refinement of  $\mathcal{T}$ , then, by combining the monotonicity of local mesh sizes and properties of the local  $L^2$  projection, we deduce the *monotonicity* properties

$$\eta_{\mathcal{T}_*}(V, \mathcal{T}_*) \leq \eta_{\mathcal{T}}(V, \mathcal{T}) \quad \text{and} \quad \text{osc}_{\mathcal{T}_*}(V, \mathcal{T}_*) \leq \text{osc}_{\mathcal{T}}(V, \mathcal{T}) \quad \forall V \in \mathbb{V}(\mathcal{T}).$$

We now recall the well-known upper and lower bounds for the energy error in terms of the residual-type estimator [1, 16, 17, 18, 23, 27].

LEMMA 2.2 (global a posteriori upper and lower bounds). *Let  $u \in \mathbb{V}$  be the solution of (2.1),  $\mathcal{T} \in \mathbb{T}$ , and  $U \in \mathbb{V}(\mathcal{T})$  be the Ritz-Galerkin solution (2.3).*

*Then there exist a constant  $C_1$ , solely depending on  $\mathcal{T}_0$  and  $1/c_B$ , such that*

$$(2.6) \quad \|u - U\|_{\Omega}^2 \leq C_1 \eta_{\mathcal{T}}^2(U, \mathcal{T}),$$

*and a constant  $C_2$ , solely depending on  $\mathcal{T}_0$  and  $1/c_B$ , such that*

$$(2.7) \quad C_2 \eta_{\mathcal{T}}^2(U, \mathcal{T}) \leq \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}).$$

**2.5. The module MARK.** In the selection of elements we rely on Dörfler marking. Given a grid  $\mathcal{T}$ , the set of indicators  $\{\eta_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}}$ , and marking parameter  $\theta \in (0, 1]$ , we suppose that MARK outputs a subset of *marked elements*  $\mathcal{M} \subset \mathcal{T}$ , i.e.,

$$\mathcal{M} = \text{MARK}(\{\eta_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}}, \mathcal{T}, \theta),$$

such that  $\mathcal{M}$  satisfies Dörfler property

$$(2.8) \quad \eta_{\mathcal{T}}(U, \mathcal{M}) \geq \theta \eta_{\mathcal{T}}(U, \mathcal{T}).$$

**2.6. The module REFINE.** We suppose that a function REFINE is at our disposal that implements iterative or recursive bisection; see [2, 13, 14, 15, 24, 25]. When relying on recursive bisection, the distribution of refinement edges has to fulfill some compatibility conditions on  $\mathcal{T}_0$ . Given a fixed number  $b \geq 1$ , for any  $\mathcal{T} \in \mathbb{T}$  and a subset  $\mathcal{M} \subset \mathcal{T}$  of marked elements,

$$\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$$

outputs a conforming triangulation  $\mathcal{T}_* \in \mathbb{T}$ , where at least all elements of  $\mathcal{M}$  are bisected  $b$  times. In particular, this implies that  $\mathcal{M} \subset \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$ .

To ensure conformity of  $\mathcal{T}_*$  one usually has to refine additional elements in  $\mathcal{T} \setminus \mathcal{M}$  once or several times. In general, the number of these additionally refined elements is not controlled by  $\#\mathcal{M}$  that is,  $\#\mathcal{T}_* - \#\mathcal{T}$  cannot be bounded by  $C \#\mathcal{M}$  with a constant  $C$  independent of  $\mathcal{T}$ ;  $C$  may depend on the refinement level. On the other hand, by arguing with the entire sequence  $\{\mathcal{T}_k\}_{k \geq 0}$  of refinements, Binev, Dahmen, and DeVore showed in two dimensions that the *cumulative* number of elements added by conformity does not inflate the total number of marked elements [3, Theorem 2.4]. Stevenson generalized this result to higher dimensions [24, Theorem 6.1].

LEMMA 2.3 (complexity of REFINE). *Assume that  $\mathcal{T}_0$  verifies condition (b) of section 4 in [24]. For  $k \geq 0$  let  $\{\mathcal{T}_k\}_{k \geq 0}$  be any sequence of refinements of  $\mathcal{T}_0$  where  $\mathcal{T}_{k+1}$  is generated from  $\mathcal{T}_k$  by  $\mathcal{T}_{k+1} = \text{REFINE}(\mathcal{T}_k, \mathcal{M}_k)$  with a subset  $\mathcal{M}_k \subset \mathcal{T}_k$ .*

*Then there exists a constant  $C_0$  solely depending on  $\mathcal{T}_0$  and  $b$  such that*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq C_0 \sum_{j=0}^{k-1} \#\mathcal{M}_j \quad \forall k \geq 1.$$

The original result is proved for a sequence of meshes generated by a loop

Mark  $\rightarrow$  Subdivision  $\rightarrow$  Complete,

where **Mark** selects certain elements, **Subdivision** bisects marked elements only once ( $b = 1$ ), and **Complete** adds some additional subdivisions to generate a conforming triangulation. To apply the above result when **REFINE** performs  $b > 1$  bisections, the set  $\mathcal{M}_k$  is to be understood as a sequence of *single* bisections recorded in sets  $\{\mathcal{M}_k(j)\}_{j=1}^b$ , which belong to intermediate triangulations between  $\mathcal{T}_k$  and  $\mathcal{T}_{k+1}$  with  $\#\mathcal{M}_k(j) \leq 2^{j-1}\#\mathcal{M}_k$ ,  $j = 1, \dots, b$ . Then Lemma 2.3 is a direct consequence of [24, Theorem 6.1] because

$$\sum_{j=1}^b \#\mathcal{M}_k(j) \leq \sum_{j=1}^b 2^{j-1}\#\mathcal{M}_k = (2^b - 1)\#\mathcal{M}_k.$$

**2.7. Adaptive algorithm.** In this section we now collect the modules described in the previous sections into the adaptive finite element method. In doing this, we replace the dependence on the actual triangulation  $\mathcal{T}$  by the iteration counter  $k \geq 0$ , for instance,  $\mathbb{V}_k = \mathbb{V}(\mathcal{T}_k)$ . The basic loop of AFEM is then given by the following iteration:

**AFEM**

Given the initial grid  $\mathcal{T}_0$  and marking parameter  $0 < \theta \leq 1$   
**set**  $k := 0$  **and iterate**

- (1)  $U_k = \text{SOLVE}(\mathcal{T}_k)$ ;
- (2)  $\{\eta_k(U_k, T)\}_{T \in \mathcal{T}_k} = \text{ESTIMATE}(U_k, \mathcal{T}_k)$ ;
- (3)  $\mathcal{M}_k = \text{MARK}(\{\eta_k(U_k, T)\}_{T \in \mathcal{T}_k}, \mathcal{T}_k, \theta)$ ;
- (4)  $\mathcal{T}_{k+1} = \text{REFINE}(\mathcal{M}_k, \mathcal{T}_k)$ ;  $k := k + 1$ .

We want to stress that the AFEM is really a *standard algorithm* in that it employs only the error indicators  $\{\eta_k(U_k, T)\}_{T \in \mathcal{T}_k}$  and does not use the oscillation indicators  $\{\text{osc}_k(U_k, T)\}_{T \in \mathcal{T}_k}$ . In addition, the AFEM relies only on a minimal refinement; i.e., an interior node property is not enforced for marked elements.

Several remarks are now in order:

- Recently, Morin, Siebert, and Veiser proved convergence of the above iteration for general marking strategies, including maximum and equidistribution besides Dörfler strategy [20]. The main result is a plain convergence result, i.e.,

$$\lim_{k \rightarrow \infty} \|u - U_k\|_{L^2(\Omega)} = \lim_{k \rightarrow \infty} \eta_k(U_k, \mathcal{T}_k) = 0.$$

This does not provide a strict total error reduction between two successive iterations, which is one key ingredient in the optimality proof of section 5.

- By using similar techniques as Diening and Kreuzer for the  $p$ -Laplacian [7], we prove in section 4 a contraction property for  $\|u - U_k\|_{L^2(\Omega)}^2 + \gamma \eta_k^2(U_k, \mathcal{T}_k)$  between two successive iterations, with  $\gamma > 0$  a suitable scaling factor; we observe that [7] still requires the interior node property. This in turn implies *linear convergence*. Crucial ingredients are the a posteriori *global upper bound* and *Dörfler marking*.



- We prove in section 5 that  $\mathcal{T}_k$  exhibits *quasi-optimal cardinality* provided that
  - (a) the marking parameter  $\theta$  satisfies  $\theta < \theta_*$ , where  $\theta_*$ , defined explicitly in Assumption 5.8, depends on the ratio  $\sqrt{C_2/C_1}$  of constants in (2.6) and (2.7);
  - (b)  $\mathcal{M}_k$  satisfies (2.8) with minimal cardinality;
  - (c)  $\mathcal{T}_{k+1}$  is the smallest refinement of  $\mathcal{T}_k$  such that  $\mathcal{M}_k \subset \mathcal{R}_{\mathcal{T}_k \rightarrow \mathcal{T}_{k+1}}$ .

The proof hinges on the a posteriori *lower bound* and a *localized upper bound*.

**3. Auxiliary results.** In this section we prove a quasi-reduction property of the estimator, which is instrumental for convergence. We also show a perturbation result of oscillation and a localized upper bound, both between two discrete functions, that play an essential role in the optimality proof. Finally we discuss the overlay of two meshes and thereby prove (2.2).

Throughout this paper we use the notation  $a \preceq b$  to indicate  $a \leq Cb$ , with generic constants  $C$  solely depending on shape regularity and thus on  $\mathcal{T}_0$ , the number of bisections  $b$  of marked elements, and the polynomial degree  $n$ . Dependence on data  $\mathbf{D} = (\mathbf{A}, c)$  and  $f$  of (1.1) is traced explicitly. We denote  $a \preceq b \preceq a$  by  $a \approx b$ .

**3.1. Reduction of the error estimator and oscillation.** We relate the error indicators and oscillation of two nested triangulations to each other. The link involves weighted maximum-norms of the coefficient functions  $\mathbf{D}$  or their oscillation. These results are the basis for our analyses in sections 4 and 5.

We start by defining the weighted maximum-norm of the coefficients  $\mathbf{D}$  and their oscillation. For  $m \in \mathbb{N}_0$ ,  $\mathcal{T} \in \mathbb{T}$ , and  $v \in L^\infty(\Omega)$ , we recall that  $\Pi_m^\infty v$  is the best  $L^\infty(\Omega)$ -approximation in the space of discontinuous polynomials of degree  $\leq m$ . We further set  $\Pi_{-1}^\infty v = 0$ ,  $P_m^\infty v = v - \Pi_m^\infty v$ , and

$$\begin{aligned} \eta_{\mathcal{T}}^2(\mathbf{D}, T) &:= h_T^2 \left( \|\operatorname{div} \mathbf{A}\|_{L^\infty(T)}^2 + h_T^{-2} \|\mathbf{A}\|_{L^\infty(\omega_T)}^2 + \|c\|_{L^\infty(T)}^2 \right), \\ \operatorname{osc}_{\mathcal{T}}^2(\mathbf{D}, T) &:= h_T^2 \left( \|P_{n-1}^\infty \operatorname{div} \mathbf{A}\|_{L^\infty(T)}^2 + h_T^{-2} \|P_n^\infty \mathbf{A}\|_{L^\infty(\omega_T)}^2 \right. \\ &\quad \left. + h_T^2 \|P_{n-2}^\infty c\|_{L^\infty(T)}^2 + \|P_{n-1}^\infty c\|_{L^\infty(T)}^2 \right); \end{aligned}$$

note that  $P_m^\infty$  is defined elementwise. For any subset  $\mathcal{T}' \subset \mathcal{T}$  we finally set

$$\eta_{\mathcal{T}}(\mathbf{D}, \mathcal{T}') := \max_{T \in \mathcal{T}'} \eta_{\mathcal{T}}(\mathbf{D}, T) \quad \text{and} \quad \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T}') := \max_{T \in \mathcal{T}'} \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, T).$$

*Remark 3.1* (monotonicity). The use of best approximation in  $L^\infty$  in the definitions of  $\eta_{\mathcal{T}}(\mathbf{D}, \mathcal{T})$  and  $\operatorname{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T})$  implies the following monotonicity property: for all  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ , with  $\mathcal{T} \leq \mathcal{T}_*$ , there holds

$$\eta_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \leq \eta_{\mathcal{T}}(\mathbf{D}, \mathcal{T}) \quad \text{and} \quad \operatorname{osc}_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \leq \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T}).$$

The following lemma allows us to avoid any smoothness assumption on the coefficients  $\mathbf{D}$  of the PDE.

**LEMMA 3.2** (implicit interpolation). *Let  $\omega$  be either a  $d$  or a  $(d-1)$  simplex. For  $\ell \in \mathbb{N}$  denote by  $\Pi_m^p: L^p(\omega, \mathbb{R}^\ell) \rightarrow \mathbb{P}_m(\omega, \mathbb{R}^\ell)$  the operator of best  $L^p$ -approximation in  $\omega$ , and set  $P_m^p = \operatorname{id} - \Pi_m^p$ . Then for all  $v \in L^\infty(\omega, \mathbb{R}^\ell)$ ,  $V \in \mathbb{P}_n(\omega, \mathbb{R}^\ell)$  and  $m \geq n$ , there holds*

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

*Proof.* Since  $V \in \mathbb{P}_n$  we obtain  $\Pi_{m-n}^\infty(v) \cdot V \in \mathbb{P}_m$ , and thus the orthogonality of the  $L^2$ -projection yields

$$\begin{aligned} \int_{\omega} (vV - \Pi_m^2(vV))(vV - \Pi_m^2(vV)) &= \int_{\omega} (vV - \Pi_m^2(vV))(v - \Pi_{m-n}^\infty(v)) \cdot V \\ &\leq \|P_m^2(vV)\|_{L^2(\omega)} \|P_{m-n}^\infty v\|_{L^\infty(\omega)} \|V\|_{L^2(\omega)}, \end{aligned}$$

which finishes the proof.  $\square$

**PROPOSITION 3.3** (local perturbation). *Let  $\mathcal{T} \in \mathbb{T}$ . For all  $T \in \mathcal{T}$  and for any pair of discrete functions  $V, W \in \mathbb{V}(\mathcal{T})$ , we have*

$$(3.1) \quad \eta_{\mathcal{T}}(V, T) \leq \eta_{\mathcal{T}}(W, T) + \bar{\Lambda}_1 \eta_{\mathcal{T}}(\mathbf{D}, T) \|V - W\|_{H^1(\omega_T)},$$

$$(3.2) \quad \text{osc}_{\mathcal{T}}(V, T) \leq \text{osc}_{\mathcal{T}}(W, T) + \bar{\Lambda}_1 \text{osc}_{\mathcal{T}}(\mathbf{D}, T) \|V - W\|_{H^1(\omega_T)},$$

where  $\omega_T$  is the union of elements in  $\mathcal{T}$  sharing a side with  $T$ . The constant  $\bar{\Lambda}_1 > 0$  depends only on the shape regularity of  $\mathcal{T}_0$  and the polynomial degree  $n$ .

*Proof.* We prove only (3.2); estimate (3.1) is somewhat simpler and can be derived similarly. Recall the definition of the element residual  $R(V) = f + \mathcal{L}(V)$  and the notation  $P_m^p = \text{id} - \Pi_m^p$ . Since the  $L^2$ -projection  $\Pi_m^2$  is linear, by adding and subtracting  $W$  and using the triangle inequality, we obtain

$$\text{osc}_{\mathcal{T}}(V, T) \leq \text{osc}_{\mathcal{T}}(W, T) + h_T \|P_{2n-2}^2 \mathcal{L}(E)\|_{L^2(T)} + h_T^{1/2} \|P_{2n-1}^2 J(E)\|_{L^2(\partial T)},$$

with  $E := V - W$ . It remains to show that the second and third terms are bounded by  $\text{osc}_{\mathcal{T}}(\mathbf{D}, T)$  times the local  $H^1$ -norm of  $E$ . For  $\mathcal{L}(E) = \text{div } \mathbf{A} \nabla E - cE$  we have

$$\|P_{2n-2}^2 \mathcal{L}(E)\|_{L^2(T)} \leq \|P_{2n-2}^2(\text{div } \mathbf{A} \nabla E)\|_{L^2(T)} + \|P_{2n-2}^2(cE)\|_{L^2(T)}.$$

We split the divergence term as

$$\text{div}(\mathbf{A} \nabla E) = \text{div } \mathbf{A} \cdot \nabla E + \mathbf{A} : D^2 E,$$

where  $D^2 E$  is the Hessian of  $E$ . By invoking Lemma 3.2 with  $\omega = T$  and observing that the polynomial degree of  $\nabla E$  is  $n - 1$ , we infer for the first term that

$$\|P_{2n-2}^2(\text{div } \mathbf{A} \cdot \nabla E)\|_{L^2(T)} \leq \|P_{n-1}^\infty \text{div } \mathbf{A}\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

Since  $D^2 E$  is a polynomial of degree  $\leq n - 2$ , by applying Lemma 3.2 again in conjunction with an inverse inequality, we have for the second term

$$\|P_{2n-2}^2(\mathbf{A} : D^2 E)\|_{L^2(T)} \preccurlyeq h_T^{-1} \|P_n^\infty \mathbf{A}\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

To analyze the reaction term we write

$$\|P_{2n-2}^2(cE)\|_{L^2(T)} \leq \|P_{2n-2}^2(c\Pi_0^2 E)\|_{L^2(T)} + \|P_{2n-2}^2(cP_0^2 E)\|_{L^2(T)}.$$

By applying again Lemma 3.2 we have for the first term

$$\|P_{2n-2}^2(c\Pi_0^2 E)\|_{L^2(T)} \leq \|P_{2n-2}^\infty c\|_{L^\infty(T)} \|E\|_{L^2(T)}$$

and for the last one

$$\|P_{2n-2}^2(cP_0^2 E)\|_{L^2(T)} \preccurlyeq h_T \|P_{n-2}^\infty c\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

We now deal with the jump residual. Let  $T' \in \mathcal{T}$  share an interior side  $\sigma$  with  $T$ . We write  $J(E) = ((\mathbf{A}\nabla E)|_T - (\mathbf{A}\nabla E)|_{T'}) \cdot \boldsymbol{\nu}$  and use the linearity of  $\Pi_{2n-1}^2$ , Lemma 3.2 with  $\omega = \sigma$ , and the inverse inequality  $\|\nabla E\|_{L^2(\sigma)} \preceq h_T^{-1/2} \|\nabla E\|_{L^2(T)}$  to deduce that

$$\|P_{2n-1}^2(\mathbf{A}\nabla E)|_T \cdot \boldsymbol{\nu}\|_{L^2(\sigma)} \preceq h_T^{-\frac{1}{2}} \|P_n^\infty \mathbf{A}\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

The same argument holds for  $T'$ , and, since  $\mathcal{T}$  is shape-regular, we can replace  $h_{T'}$  by  $h_T$ . Finally, collecting the above estimates for  $T$  and all of its neighbors yields the assertion (3.2).  $\square$

The following two corollaries are global forms of the above lemma.

**COROLLARY 3.4** (estimator reduction). *For  $\mathcal{T} \in \mathbb{T}$  and  $\mathcal{M} \subset \mathcal{T}$  let  $\mathcal{T}_* \in \mathbb{T}$  be given by  $\mathcal{T}_* := \text{REFINE}(\mathcal{T}, \mathcal{M})$ . If  $\Lambda_1 := (d+1)\bar{\Lambda}_1^2/c_B$ , with  $\bar{\Lambda}_1$  from Proposition 3.3 and  $\lambda := 1 - 2^{-\frac{b}{d}} > 0$ , then there holds for all  $V \in \mathbb{V}(\mathcal{T})$  and  $V_* \in \mathbb{V}(\mathcal{T}_*)$  and any  $\delta > 0$*

$$\begin{aligned} \eta_{\mathcal{T}_*}^2(V_*, \mathcal{T}_*) &\leq (1 + \delta) \{ \eta_{\mathcal{T}}^2(V, \mathcal{T}) - \lambda \eta_{\mathcal{T}}^2(V, \mathcal{M}) \} \\ &\quad + (1 + \delta^{-1}) \Lambda_1 \eta_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0) \|V_* - V\|_{\Omega}^2. \end{aligned}$$

*Proof.* By applying Proposition 3.3 with  $V_*, V \in \mathbb{V}(\mathcal{T}_*)$  over  $T \in \mathcal{T}_*$  and using Young's inequality with parameter  $\delta$ , we derive

$$\eta_{\mathcal{T}_*}^2(V_*, T) \leq (1 + \delta) \eta_{\mathcal{T}_*}^2(V, T) + (1 + \delta^{-1}) \bar{\Lambda}_1^2 \eta_{\mathcal{T}_*}^2(\mathbf{D}, T) \|V_* - V\|_{H^1(\omega_T)}^2.$$

By summing over all elements  $T \in \mathcal{T}_*$ , using the finite overlap property of patches  $\omega_T$ , and the equivalence of the  $H^1$ -norm and the energy norm in  $\Omega$ , we have the following direct by-product:

$$\eta_{\mathcal{T}_*}^2(V_*, \mathcal{T}_*) \leq (1 + \delta) \eta_{\mathcal{T}_*}^2(V, \mathcal{T}_*) + (1 + \delta^{-1}) \Lambda_1 \eta_{\mathcal{T}_*}^2(\mathbf{D}, \mathcal{T}_*) \|V_* - V\|_{\Omega}^2.$$

For a marked element  $T \in \mathcal{M}$ , we set  $\mathcal{T}_{*,T} := \{T' \in \mathcal{T}_* \mid T' \subset T\}$ . Since  $V \in \mathbb{V}(\mathcal{T})$  and  $\mathbf{A}$  jumps only across sides of  $\mathcal{T}_0$ , we see that  $J(V) = 0$  on sides of  $\mathcal{T}_{*,T}$  in the interior of  $T$ . We then obtain

$$\sum_{T' \in \mathcal{T}_{*,T}} \eta_{\mathcal{T}_*}^2(V, T') \leq 2^{-\frac{b}{d}} \eta_{\mathcal{T}}^2(V, T),$$

because refinement by bisection implies that  $h_{T'} = |T'|^{\frac{1}{d}} \leq (2^{-b}|T|)^{\frac{1}{d}} \leq 2^{-\frac{b}{d}} h_T$  for all  $T' \in \mathcal{T}_{*,T}$ . For an element  $T \in \mathcal{T} \setminus \mathcal{M}$ , instead, Remark 2.1 yields  $\eta_{\mathcal{T}_*}(V, T) \leq \eta_{\mathcal{T}}(V, T)$ . Hence, by summing over all  $T \in \mathcal{T}_*$ , we arrive at

$$\eta_{\mathcal{T}_*}^2(V, \mathcal{T}_*) \leq \eta_{\mathcal{T}}^2(V, \mathcal{T} \setminus \mathcal{M}) + 2^{-\frac{b}{d}} \eta_{\mathcal{T}}^2(V, \mathcal{M}) = \eta_{\mathcal{T}}^2(V, \mathcal{T}) - \lambda \eta_{\mathcal{T}}^2(V, \mathcal{M}).$$

The assertion finally follows from the monotonicity  $\eta_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \leq \eta_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)$  stated in Remark 3.1.  $\square$

**COROLLARY 3.5** (perturbation of oscillation). *Let  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ , with  $\mathcal{T} \leq \mathcal{T}_*$ , and let  $\Lambda_1 = (d+1)\bar{\Lambda}_1^2/c_B$  be as in Corollary 3.4. For all  $V \in \mathbb{V}(\mathcal{T})$  and  $V_* \in \mathbb{V}(\mathcal{T}_*)$ , we have*

$$\text{osc}_{\mathcal{T}}^2(V, \mathcal{T} \cap \mathcal{T}_*) \leq 2 \text{osc}_{\mathcal{T}_*}^2(V_*, \mathcal{T} \cap \mathcal{T}_*) + 2\Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0) \|V - V_*\|_{\Omega}^2.$$

*Proof.* Remark 2.1 yields  $\text{osc}_{\mathcal{T}}(V, T) = \text{osc}_{\mathcal{T}_*}(V, T)$  for all  $T \in \mathcal{T} \cap \mathcal{T}_*$ , whence

$$\text{osc}_{\mathcal{T}}^2(V, T) \leq 2 \text{osc}_{\mathcal{T}_*}^2(V_*, T) + 2\bar{\Lambda}_1^2 \text{osc}_{\mathcal{T}_*}^2(\mathbf{D}, T) \|V - V_*\|_{H^1(\omega_T)}^2,$$

by (3.2) and Young's inequality. We now sum over  $T \in \mathcal{T} \cap \mathcal{T}_*$  and use the equivalence of  $H^1$  and energy norms in  $\Omega$ , as well as the monotonicity property  $\text{osc}_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \leq \text{osc}_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)$  stated in Remark 3.1, to prove the assertion.  $\square$

**3.2. Localized upper bound.** To prove the optimality of the AFEM, we need a localized upper bound for the distance between two nested solutions. This slightly improves a similar result by Stevenson [23] in the sense that the error can be estimated here by using only the indicators of refined elements without a buffer layer.

LEMMA 3.6 (localized upper bound). *For  $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ , with  $\mathcal{T} \leq \mathcal{T}_*$ , let  $\mathcal{R} = \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$  be the set of refined elements. Let  $U \in \mathbb{V}(\mathcal{T})$  and  $U_* \in \mathbb{V}(\mathcal{T}_*)$  be the discrete solutions of (2.1). Then the following localized upper bound is valid:*

$$\|U_* - U\|_{\Omega}^2 \leq C_1 \eta_{\mathcal{T}}^2(U, \mathcal{R}).$$

*Proof.* Let  $\Omega_* = \bigcup\{T : T \in \mathcal{R}\}$  be the union of refined elements, and denote by  $\Omega_i$ ,  $i = 1, \dots, I$ , the connected components of its interior. For  $i \in \{1, \dots, I\}$  we set  $\mathcal{T}_i = \{T \in \mathcal{T} : T \subset \Omega_i\}$  and let  $\mathbb{V}(\mathcal{T}_i)$  be the restriction of  $\mathbb{V}(\mathcal{T})$  to  $\Omega_i$ . By using the Scott–Zhang interpolation operator  $P_i : H^1(\Omega_i) \rightarrow \mathbb{V}(\mathcal{T}_i)$  over the triangulation  $\mathcal{T}_i$ , the following interpolation estimate holds true for all  $v \in H^1(\mathcal{T}_i)$ :

$$(3.3) \quad \sum_{T \in \mathcal{T}_i} h_T^{-2} \|v - P_i v\|_{L_2(T)}^2 + h_T^{-1} \|v - P_i v\|_{L_2(\partial T)}^2 \preceq \|\nabla v\|_{L_2(\Omega_i)}^2.$$

The constant hidden in “ $\preceq$ ” does not depend on  $\Omega_i$  but only on the shape-regularity of the underlying triangulation  $\mathcal{T}_i$  and thus solely on  $\mathcal{T}_0$  [22]. By construction,  $P_i$  preserves conforming boundary values, i.e.,  $P_i v = v$  on  $\partial\Omega_i$  whenever  $v = V$  on  $\partial\Omega_i$  for some  $V \in \mathbb{V}(\mathcal{T}_i)$ .

For the error  $E_* := U_* - U \in \mathbb{V}(\mathcal{T}_*)$  we construct an approximation  $V \in \mathbb{V}(\mathcal{T})$  by setting

$$V := \begin{cases} P_i E_* & \text{in } \Omega_i, \\ E_* & \text{elsewhere.} \end{cases}$$

Since  $E_*$  has conforming boundary values on  $\partial\Omega_i$  in  $\mathbb{V}(\mathcal{T}_i)$ , we conclude that  $V$  is continuous in  $\Omega$  implying that  $V \in \mathbb{V}(\mathcal{T})$  and  $V$  is an  $H^1$ -stable approximation to  $E_*$ .

Nesting of spaces  $\mathbb{V}(\mathcal{T}) \subset \mathbb{V}(\mathcal{T}_*)$  implies Galerkin orthogonality  $\mathcal{B}[E_*, E_*] = \mathcal{B}[E_*, E_* - V]$ . In conjunction with  $E_* - V = 0$  in  $\Omega \setminus \Omega_*$  we thus obtain by standard arguments

$$\mathcal{B}[E_*, E_*] = \sum_{T \in \mathcal{R}} \langle R(U), E_* - V \rangle_T + \frac{1}{2} \langle J(U), E_* - V \rangle_{\partial T} \preceq \eta_{\mathcal{T}}(U, \mathcal{R}) \|\nabla E_*\|_{L^2(\Omega_*)}.$$

In the last step we have used (3.3) on all connected components  $\Omega_i$  of  $\Omega_*$  with  $v = E_*$ . This, in conjunction with the coercivity of  $\mathcal{B}$ , proves the proposition.  $\square$

**3.3. Overlay of meshes.** We finish the auxiliary results with a counting argument for the overlay  $\mathcal{T}_1 \oplus \mathcal{T}_2$  of two triangulations  $\mathcal{T}_1$  and  $\mathcal{T}_2$  from [23]. For the convenience of the reader we include a proof of this crucial result. As a consequence of the following lemma we see that for two conforming triangulations  $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$  the overlay is the smallest conforming triangulation  $\mathcal{T} \in \mathbb{T}$ , with  $\mathcal{T}_1, \mathcal{T}_2 \leq \mathcal{T}$ .

LEMMA 3.7 (overlay of meshes). *For  $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$  the overlay  $\mathcal{T} := \mathcal{T}_1 \oplus \mathcal{T}_2$  is conforming, i.e.,  $\mathcal{T} \in \mathbb{T}$ , and satisfies*

$$\#\mathcal{T} \leq \#\mathcal{T}_1 + \#\mathcal{T}_2 - \#\mathcal{T}_0.$$

*Proof.* Assume that  $\mathcal{T}$  contains a nonconforming vertex  $z$ . Then there exist  $T_1, T_2 \in \mathcal{T}$  with a common edge such that  $z$  is a vertex of  $T_1$  and  $z \in T_2$  but  $z$  is no

vertex of  $T_2$ . Without loss of generality let  $T_1 \in \mathcal{T}_1$ . Since  $\mathcal{T}_1$  is conforming, there exists a  $T' \in \mathcal{T}_1$ ,  $T' \subset T_2$  such that  $z$  is a vertex of  $T'$ . Hence,  $T_2$  cannot be a leaf node of  $\mathcal{F}(\mathcal{T})$ , i.e.,  $T_2 \notin \mathcal{T}$ , a contradiction.

For  $T \in \mathcal{T}_0$  and  $i = 1, 2$  we denote by  $\mathcal{F}_i(T) \subset \mathcal{F}(\mathcal{T})$  the binary trees with root  $T$  corresponding to  $\mathcal{T}_i$  and let  $\mathcal{T}_i(T)$  be the triangulation given by the leaf nodes of  $\mathcal{F}_i(T)$ . Since  $\mathcal{T}(T) \subset \mathcal{T}_1(T) \cup \mathcal{T}_2(T)$ , we infer that  $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T)$ . We now show that  $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T) - 1$  by distinguishing two cases.

*Case 1:*  $\mathcal{T}_1(T) \cap \mathcal{T}_2(T) \neq \emptyset$ . Then there exists  $T' \in \mathcal{T}_1(T) \cap \mathcal{T}_2(T)$ , and so  $T' \in \mathcal{T}(T)$ . By counting  $T'$  only once in  $\#(\mathcal{T}_1(T) \cup \mathcal{T}_2(T))$  we get  $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T) - 1$ .

*Case 2:*  $\mathcal{T}_1(T) \cap \mathcal{T}_2(T) = \emptyset$ . Then there exists  $T' \in \mathcal{T}_1(T)$  (resp.,  $T' \in \mathcal{T}_2(T)$ ) so that  $T' \notin \mathcal{T}(T)$ , for otherwise  $T' \in \mathcal{T}_2(T)$  (resp.,  $T' \in \mathcal{T}_1(T)$ ), thereby contradicting the assumption. We obtain again  $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T) - 1$ .

Finally, since  $\mathcal{T}_i = \bigcup_{T \in \mathcal{T}_0} \mathcal{T}_i(T)$ , the assertion follows by adding over  $T \in \mathcal{T}_0$ .  $\square$

**4. Contraction property of the AFEM.** We now prove that the AFEM is a contraction with respect to the sum of the energy error plus the scaled error estimator, the so-called *quasi error*. Consequently, the quasi error is reduced by a fixed rate at every step. This can be motivated heuristically as follows: in light of (2.4) the energy error  $\|u - U_k\|_\Omega$  decreases strictly, unless  $U_{k+1} = U_k$ , in which case the estimator  $\eta_k(U_k, \mathcal{T}_k)$  does according to Corollary 3.4. The theorem below makes this observation quantitative.

**THEOREM 4.1** (contraction property). *Let  $\theta \in (0, 1]$ , and let  $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k \geq 0}$  be the sequence of meshes, finite element spaces, and discrete solutions produced by the AFEM.*

*Then there exist constants  $\gamma > 0$  and  $0 < \alpha < 1$ , depending solely on the shape-regularity of  $\mathcal{T}_0$ ,  $b$ , and the marking parameter  $0 < \theta \leq 1$ , such that*

$$\|u - U_{k+1}\|_\Omega^2 + \gamma \eta_{k+1}^2(U_{k+1}, \mathcal{T}_{k+1}) \leq \alpha^2 \left( \|u - U_k\|_\Omega^2 + \gamma \eta_k^2(U_k, \mathcal{T}_k) \right).$$

*Proof.* For convenience, we use the notation

$$\begin{aligned} e_k &:= \|u - U_k\|_\Omega, & E_k &:= \|U_{k+1} - U_k\|_\Omega, \\ \eta_k &:= \eta_k(U_k, \mathcal{T}_k), & \eta_k(\mathcal{M}_k) &:= \eta_k(U_k, \mathcal{M}_k), & \eta_0(\mathbf{D}) &:= \eta_0(\mathbf{D}, \mathcal{T}_0). \end{aligned}$$

We combine the orthogonality (2.4) with Corollary 3.4 to write

$$e_{k+1}^2 + \gamma \eta_{k+1}^2 \leq e_k^2 - E_k^2 + (1 + \delta) \gamma (\eta_k^2 - \lambda \eta_k^2(\mathcal{M}_k)) + (1 + \delta^{-1}) \gamma \Lambda_1 \eta_0^2(\mathbf{D}) E_k^2.$$

We choose  $\gamma$  dependent on  $\delta$  to be

$$(4.1) \quad \gamma := \frac{1}{(1 + \delta^{-1}) \Lambda_1 \eta_0^2(\mathbf{D})} \quad \Leftrightarrow \quad \gamma(1 + \delta) = \frac{\delta}{\Lambda_1 \eta_0^2(\mathbf{D})}$$

to obtain

$$e_{k+1}^2 + \gamma \eta_{k+1}^2 \leq e_k^2 + (1 + \delta) \gamma \eta_k^2 - (1 + \delta) \lambda \gamma \eta_k^2(\mathcal{M}_k).$$

By invoking Dörfler marking (2.8), we deduce that

$$e_{k+1}^2 + \gamma \eta_{k+1}^2 \leq e_k^2 + (1 + \delta) \gamma \eta_k^2 - (1 + \delta) \lambda \theta^2 \gamma \eta_k^2.$$

We rewrite this inequality as follows with any  $\beta \in (0, 1)$ :

$$e_{k+1}^2 + \gamma \eta_{k+1}^2 \leq e_k^2 + (1 + \delta) \gamma \eta_k^2 - \beta(1 + \delta) \lambda \theta^2 \gamma \eta_k^2 - (1 - \beta)(1 + \delta) \lambda \theta^2 \gamma \eta_k^2,$$

apply the upper bound (2.6), and replace  $\gamma$  according to (4.1) to obtain

$$e_{k+1}^2 + \gamma \eta_{k+1}^2 \leq \alpha_1^2(\delta, \beta) e_k^2 + \gamma \alpha_2^2(\delta, \beta) \eta_k^2,$$

with

$$\alpha_1^2(\delta, \beta) := 1 - \beta \frac{\lambda \theta^2}{C_1 \Lambda_1 \eta_0^2(\mathbf{D})} \delta, \quad \alpha_2^2(\delta, \beta) := (1 + \delta)(1 - (1 - \beta) \lambda \theta^2).$$

Now choosing  $\delta > 0$  small enough yields

$$\alpha^2 := \max\{\alpha_1^2, \alpha_2^2\} < 1,$$

which is the desired result.  $\square$

*Remark 4.2* (ingredients for convergence). We stress that this new proof of linear convergence relies exclusively on the upper bound (2.6), the orthogonality relation (2.4), the error estimator reduction property of Corollary 3.4, and the Dörfler marking (2.8) for the estimator. It does not need any marking due to oscillation, which turns out to be problematic for optimality in light of the discussion in section 6. Equality (2.4) is used only to cancel, via (4.1), the contribution involving  $\eta_0(\mathbf{D}, \mathcal{T}_0)$  in Corollary 3.4. Its role is much less prominent than in [8, 16, 17, 18, 23].

Moreover, we use neither the lower bound (2.7) nor a discrete lower bound for proving convergence. The latter hinges on the rather demanding *interior node property*: every element of  $\mathcal{M}_k$ , as well as its adjacent elements, contains a node of  $\mathcal{T}_{k+1}$  in their interior as well as in the interior of their common sides. However, the global lower bound (2.7) will be instrumental to prove optimality in section 5.

The treatment of oscillation and the interior node property is essential in [16, 17, 18] and so in [3, 23]. Our new approach simplifies the analysis and directly applies to any polynomial degree  $n \geq 1$ .

*Remark 4.3* (optimal contraction factor  $\alpha$ ). Consider

$$D := \{(\delta, \beta) \in \mathbb{R}_+ \times [0, 1] : 0 \leq \alpha_1^2(\delta, \beta), \alpha_2^2(\delta, \beta) \leq 1\}.$$

This set is nonempty, according to the proof of Theorem 4.1, closed, and bounded; thus  $D$  is compact. Since  $\alpha_1^2$  and  $\alpha_2^2$  are continuous functions in  $D$ ,  $\alpha^2 = \max\{\alpha_1^2, \alpha_2^2\}$  attains its minimum in  $D$ . It turns out that  $\alpha^2|_{\partial D} = 1$  by definition of  $D$  and  $\alpha^2 < 1$  in the interior of  $D$  from the proof of Theorem 4.1. Consequently,  $\alpha^2$  attains an absolute minimum smaller than 1 and satisfies

$$\alpha^2 = \alpha_1^2 = \alpha_2^2.$$

To see this assume that  $\alpha_1^2 < \alpha_2^2$ , and decrease the value of  $\delta$  slightly. Since  $\alpha_1^2$  increases whereas  $\alpha_2^2$  decreases, this yields a contradiction to the minimality of  $\alpha^2$ . The case  $\alpha_1^2 > \alpha_2^2$  is similar. In principle, this optimal value of  $\alpha^2$  can be computed explicitly.

*Remark 4.4* (range of  $\gamma$  and  $\alpha$ ). We see from (4.1) that  $\gamma \approx \eta_0^{-2}(\mathbf{D}, \mathcal{T}_0)$  provided that  $\eta_0(\mathbf{D}, \mathcal{T}_0)$  is large; this provides a lower bound for  $\gamma$ . An upper bound results from the condition  $\alpha_1^2 > 0$  and  $\delta \leq 1$ , namely,

$$\alpha_1^2 = 1 - \frac{(1 + \delta)\beta\lambda\theta^2}{C_1} \gamma \quad \Rightarrow \quad \gamma < \frac{C_1}{2\beta\lambda\theta^2} \approx C_1.$$

On the other hand, it is clear from the definitions of  $\alpha_1$  and  $\alpha_2$  that  $\alpha$  deteriorates if either  $\eta_0(\mathbf{D}, \mathcal{T}_0)$  increases or  $\theta$  decreases. In fact, take  $\beta = \frac{1}{2}$  and  $\delta = \frac{1}{2}\lambda\theta^2$  to get

$$\alpha_1^2 = 1 - \frac{1}{C_1\Lambda_1\eta_0^2(\mathbf{D}, \mathcal{T}_0)} \left(\frac{\lambda\theta^2}{2}\right)^2, \quad \alpha_2^2 = 1 - \left(\frac{\lambda\theta^2}{2}\right)^2.$$

Thus by the definition of  $\alpha$  there exist constants  $c, C > 0$  not depending on  $\theta$ , with

$$1 - C\theta^4 \leq \alpha^2 \leq 1 - c\theta^4.$$

**5. Quasi-optimal cardinality of the AFEM.** Building on fundamental ideas of Binev, Dahmen, and DeVore [3] and Stevenson [23], used in the first optimality proofs for variants of the AFEM for the Poisson equation, we prove now the optimal cardinality of a *standard* AFEM for the general symmetric elliptic problem (1.1). We thus improve and extend the results of [3, 23].

**5.1. Approximation class.** Procedure MARK selects elements for refinement based on information provided by the error indicators  $\eta_{\mathcal{T}}(U, T)$  for all  $T \in \mathcal{T}$ . Hence, a decay rate of the AFEM in terms of DOFs is closely related to the behavior of the estimator  $\eta_{\mathcal{T}}(U, \mathcal{T})$ . In order to link the estimator to the energy error  $\|u - U\|_{\Omega}$ , we observe that, according to the upper bound (2.6)

$$\|u - U\|_{\Omega}^2 \leq C_1 \eta_{\mathcal{T}}^2(U, \mathcal{T}),$$

a decay rate for the estimator directly implies a decay rate for the error. However, the error may decay faster than the estimator whenever overestimation occurs. The quality of the upper bound is described by the global lower bound (2.7)

$$C_2 \eta_{\mathcal{T}}^2(U, \mathcal{T}) \leq \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}).$$

Hence, the quality of the estimator strongly depends on the resolution of data  $(\mathbf{D}, f)$  as well as the interaction of  $\mathbf{D}$  with the discrete solution  $U$ , encoded in the oscillation term  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$ . The latter cannot be removed as the following example shows.

*Remark 5.1* (role of oscillation). The presence of  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$  in the lower bound is the price to pay for having a simple and computable estimator  $\eta_{\mathcal{T}}(U, \mathcal{T})$ . To see that oscillation cannot be removed, we let  $\epsilon = 2^{-K}$  for  $K$  integer and extend the function  $\frac{1}{2}x(\epsilon - |x|)$  defined on  $(-\epsilon, \epsilon)$  to a  $2\epsilon$ -periodic  $C^1$  function  $u$  on  $\Omega = (-1, 1)$ . We then consider  $u$  for data  $\mathbf{D} = (1, 0)$  to be the exact solution, whence the corresponding right-hand side  $f$  is  $2\epsilon$ -periodic and piecewise constant with values  $+1$  and  $-1$  that change at multiples of  $\epsilon$ .

Let  $\mathcal{T}$  be a uniform mesh with mesh size  $h = 2^{-k}$ , with  $k \ll K$ . We consider piecewise linear approximation ( $n = 1$ ), so that  $f$  is orthogonal to the space of piecewise constants, i.e.,  $\Pi_0^2 f = 0$ . It is easy to verify that  $U = 0$  and

$$\|u - U\|_{\Omega} = \|u\|_{\Omega} = \frac{\epsilon}{\sqrt{6}} = \frac{2^{-K}}{\sqrt{6}} \ll 2^{-k} = h = \|hf\|_{\Omega} = \text{osc}_{\mathcal{T}}(U, \mathcal{T}) = \eta_{\mathcal{T}}(U, \mathcal{T}).$$

Therefore, the ratio between  $\|u - U\|_{\Omega}$  and  $\eta_{\mathcal{T}}(U, \mathcal{T})$  can be made arbitrarily small by increasing  $K/k$ , and  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$  accounts for the discrepancy. Moreover, if we were to express the oscillation in  $H^{-1}(\Omega)$  as suggested in [3, 23], namely,  $\|f - \Pi_0^2 f\|_{H^{-1}(\Omega)} = \|f\|_{H^{-1}(\Omega)} \approx \epsilon$ , then  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$  would overestimate this quantity. This would correspond to evaluating the estimator in  $H^{-1}(\Omega)$ , which is impractical.

According to Remark 2.1 the estimator dominates oscillation, i.e.,  $\text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \leq \eta_{\mathcal{T}}^2(U, \mathcal{T})$ , which in combination with the upper bound yields

$$\|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \leq (C_1 + 1) \eta_{\mathcal{T}}^2(U, \mathcal{T}).$$

We thus realize that

$$\eta_{\mathcal{T}}(U, \mathcal{T}) \approx \left( \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \right)^{1/2},$$

where the right-hand side is the so-called *total error*. Since all decisions of the AFEM are based on the estimator, a decay rate for the AFEM can be characterized only by properties of the total error, and this has to be reflected in the definition of a suitable approximation class. By invoking the upper bound once again, we also see that the total error is equivalent to the quasi error

$$\|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \approx \|u - U\|_{\Omega}^2 + \eta_{\mathcal{T}}^2(U, \mathcal{T}),$$

which is strictly reduced by the AFEM. Finally, the total error satisfies Cea's lemma.

LEMMA 5.2 (quasi optimality of the total error). *Let  $u$  be the solution of (2.1), and for  $\mathcal{T} \in \mathbb{T}$  let  $U \in \mathbb{V}(\mathcal{T})$  be the Ritz–Galerkin approximation of (2.3).*

*Then there exists a constant  $C_{\mathbf{D}}$  depending only on data  $\mathbf{D}$  and shape-regularity of  $\mathcal{T}_0$  such that*

$$\|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \leq C_{\mathbf{D}} \inf_{V \in \mathbb{V}(\mathcal{T})} \left( \|u - V\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V, \mathcal{T}) \right).$$

*Proof.* For  $\epsilon > 0$  choose  $V_{\epsilon} \in \mathbb{V}(\mathcal{T})$ , with

$$\|u - V_{\epsilon}\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V_{\epsilon}, \mathcal{T}) \leq (1 + \epsilon) \inf_{V \in \mathbb{V}(\mathcal{T})} \left( \|u - V\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V, \mathcal{T}) \right).$$

Corollary 3.5 with  $\mathcal{T}_* = \mathcal{T}$ ,  $V = U$ , and  $V_* = V_{\epsilon}$  yields

$$\text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \leq 2 \text{osc}_{\mathcal{T}}^2(V_{\epsilon}, \mathcal{T}) + 2 \Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0) \|U - V_{\epsilon}\|_{\Omega}^2.$$

Since  $U$  is the Galerkin solution,  $\|u - U\|_{\Omega}^2 + \|U - V_{\epsilon}\|_{\Omega}^2 = \|u - V_{\epsilon}\|_{\Omega}^2$ , whence

$$\begin{aligned} \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) &\leq (1 + 2 \Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0)) \|u - V_{\epsilon}\|_{\Omega}^2 + 2 \text{osc}_{\mathcal{T}}^2(V_{\epsilon}, \mathcal{T}) \\ &\leq (1 + \epsilon) C_{\mathbf{D}}^2 \inf_{V \in \mathbb{V}(\mathcal{T})} \left( \|u - V\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V, \mathcal{T}) \right), \end{aligned}$$

with  $C_{\mathbf{D}} = \max \{2, 1 + 2 \Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0)\}$ , and the assertion follows from  $\epsilon \rightarrow 0$ .  $\square$

After collecting the above observations we are ready to define an approximation class  $\mathbb{A}_s$  based on the total error. Let  $\mathbb{T}_N \subset \mathbb{T}$  be the set of all possible conforming triangulations generated from  $\mathcal{T}_0$  with at most  $N$  elements more than  $\mathcal{T}_0$ :

$$\mathbb{T}_N := \{\mathcal{T} \in \mathbb{T} \mid \#\mathcal{T} - \#\mathcal{T}_0 \leq N\}.$$

The quality of the best approximation to the total error in the set  $\mathbb{T}_N$  is given by

$$\sigma(N; v, f, \mathbf{D}) := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{V \in \mathbb{V}(\mathcal{T})} \left( \|v - V\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V, \mathcal{T}) \right)^{1/2},$$



and for  $s > 0$  we define the nonlinear approximation class  $\mathbb{A}_s$  to be

$$\mathbb{A}_s := \left\{ (v, f, \mathbf{D}) \mid |v, f, \mathbf{D}|_s := \sup_{N>0} (N^s \sigma(N; v, f, \mathbf{D})) < \infty \right\}.$$

Thanks to Lemma 5.2, the solution  $u$  of (1.1) with data  $(\mathbf{D}, f)$  satisfies

$$(5.1) \quad \sigma(N; u, f, \mathbf{D}) \approx \inf_{\mathcal{T} \in \mathbb{T}_N} \{ \eta_{\mathcal{T}}(U, \mathcal{T}) : U = \text{SOLVE}(\mathcal{T}) \}.$$

An important pending issue is the characterization of  $\mathbb{A}_s$ , which is beyond the scope of this paper. Nevertheless, we present a preliminary discussion now. We start by stressing that  $\mathbb{A}_s$  is not a typical approximation class for functions, as known from approximation theory, because there is a nonlinear interaction between data  $\mathbf{D} = (\mathbf{A}, c)$  and  $U$  through  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$ . To avoid such an interaction, we first assume that coefficients  $\mathbf{A}$  and  $c$  are piecewise polynomials of degree at most  $n$  and  $n-2$  over  $\mathcal{T}_0$ , respectively. In this case, oscillation reduces to

$$\text{osc}_{\mathcal{T}}(U, \mathcal{T}) = \|h(f - \Pi_{2n-2}^2 f)\|_{L^2(\Omega)} \quad \forall \mathcal{T} \in \mathbb{T},$$

which is independent of  $U$ . We have the following characterization of  $\mathbb{A}_s$  in terms of standard approximation classes [3, 4, 23]:

$$\begin{aligned} \mathcal{A}_s &:= \left\{ v \in \mathbb{V} \mid \sup_{N>0} (N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{V \in \mathbb{V}(\mathcal{T})} \|v - V\|_{\Omega}) < \infty \right\}, \\ \bar{\mathcal{A}}_s &:= \left\{ g \in L^2(\Omega) \mid \sup_{N>0} (N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \|h(g - \Pi_{2n-2}^2 g)\|_{L^2(\Omega)}) < \infty \right\}. \end{aligned}$$

LEMMA 5.3 (equivalence of classes). *If data  $\mathbf{D} = (\mathbf{A}, c)$  are piecewise polynomials over  $\mathcal{T}_0$ , as specified above, then  $(u, f, \mathbf{D}) \in \mathbb{A}_s$  if and only if  $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$ .*

*Proof.* It is obvious that  $(u, f, \mathbf{D}) \in \mathbb{A}_s$  implies that  $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$ . In order to prove the reverse implication, let  $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$  and  $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}_N$  be meshes so that  $\|u - U\|_{\Omega} \lesssim N^{-s}$  and  $\|f - \Pi_{2n-2}^2 f\|_{L^2(\Omega)} \lesssim N^{-s}$ . According to Lemma 3.7, the overlay  $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2$  is a mesh with  $\#\mathcal{T} - \#\mathcal{T}_0 \leq 2N$ , whence  $\mathcal{T} \in \mathbb{T}_{2N}$  and

$$\|u - U\|_{\Omega} + \text{osc}_{\mathcal{T}}(U, \mathcal{T}) \lesssim N^{-s}.$$

This implies that  $(u, f, \mathbf{D}) \in \mathbb{A}_s$ , as asserted.  $\square$

A similar approximability condition  $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$  is imposed in [3, 23] to obtain quasi-optimal cardinality in terms of DOFs. In contrast to our approach, however, the two classes are handled separately (see section 6.2 for separate marking).

We now turn to the nonlinear interaction encoded in  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$ . We first show, by using a priori arguments, that  $(u, f, \mathbf{D}) \in \mathbb{A}_s$  for some  $s > 0$ .

LEMMA 5.4 (decay rate of oscillation). *The oscillation term  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$  has a decay rate of order at least  $-1/d$ , i.e.,*

$$(5.2) \quad \inf_{\mathcal{T} \in \mathbb{T}_N} \{ \text{osc}_{\mathcal{T}}(U, \mathcal{T}) \mid U = \text{SOLVE}(\mathcal{T}) \} \lesssim N^{-1/d}.$$

Therefore,  $(u, f, \mathbf{D}) \in \mathbb{A}_{r/d}$  for some  $r \in (0, 1]$  depending on the regularity of  $\mathbf{D}$  and  $\Omega$ . Furthermore, if  $u \in \mathcal{A}_s$  for some  $s \in (0, 1/d]$ , then  $(u, f, \mathbf{D}) \in \mathbb{A}_s$ .

*Proof.* Let  $\mathcal{T} \in \mathbb{T}$  be generated by quasi-uniform refinement of  $\mathcal{T}_0$  with  $\#\mathcal{T} \approx N$ . By applying (3.2) with  $V = U$  and  $W = 0$ , we obtain for the oscillation term

$$\text{osc}_{\mathcal{T}}(U, \mathcal{T}) \leq h_{\mathcal{T}} \|P_{2n-2}^2 f\|_{\mathcal{T}} + \bar{\Lambda}_1 \text{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T}) \|U\|_{H^1(\omega_{\mathcal{T}})} \lesssim h_{\mathcal{T}} (\|f\|_{\mathcal{T}} + \|U\|_{H^1(\omega_{\mathcal{T}})})$$

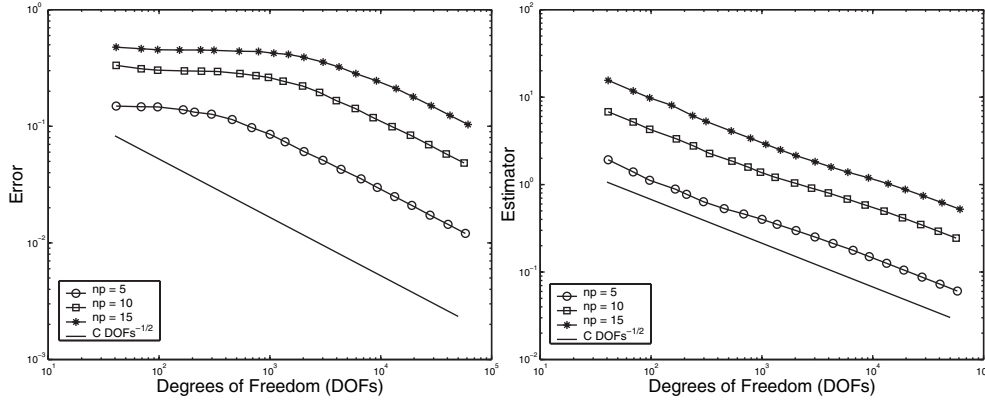


FIG. 5.1. Decay of the energy error (left) and the estimator (right) for the smooth solution  $u_S$  of section 6.1 with frequencies 5, 10, and 15. The energy error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. This behavior is allowed by  $\mathbb{A}_s$ .

by the definition of oscillation  $\text{osc}_{\mathcal{T}}(\mathbf{D}, T)$  and  $\|P_n^\infty \mathbf{A}\|_{L^\infty(T)} \leq h_T \|\nabla \mathbf{A}\|_{L^\infty(T)}$ . Quasi-uniform refinement yields the relation  $h_T \approx N^{-1/d}$  for all  $T \in \mathcal{T}$  which in turn implies (5.2). In the same vein, regularity  $u \in H^{1+r}(\Omega)$  for some  $0 < r \leq 1$  in any dimension  $d$  [11, Theorem 3] implies that a  $\|u - U\|_\Omega \lesssim N^{-r/d}$  by standard a priori estimates. Combining these estimates yields a decay rate  $-r/d$  for the total error.

Finally, let  $u \in \mathcal{A}_s$  and  $\mathcal{T}_1 \in \mathbb{T}_N$  be such that  $\|u - U\|_\Omega \lesssim N^{-s}$ . In view of (5.2), we find another grid  $\mathcal{T}_2 \in \mathbb{T}_N$  with  $\text{osc}_{\mathcal{T}}(U, \mathcal{T}) \lesssim N^{-1/d} \leq N^{-s}$ . According to Lemma 3.7, the overlay  $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2$  is a mesh with  $\#\mathcal{T} - \#\mathcal{T}_0 \leq 2N$ , i.e.,  $\mathcal{T} \in \mathbb{T}_{2N}$ . The decay rate of total error is  $N^{-s}$  over  $\mathcal{T}$ , whence  $(u, f, \mathbf{D}) \in \mathbb{A}_s$ .  $\square$

*Remark 5.5* (error notion). Lemma 5.4 states that for  $s \leq 1/d$  oscillation exhibits a decay faster than the energy error; so the situation described in Remark 5.1 shows that 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 that  $\text{osc}_k(U_k, \mathcal{T}_k) = \|h(f - \Pi_0^2 f)\|_\Omega$ , 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,  $\eta_k(U_k, \mathcal{T}_k) = \text{osc}_k(U_k, \mathcal{T}_k) = \|hf\|_\Omega$  reduces strictly for  $k \leq K$  but overestimates  $\|u - U_k\|_\Omega$ . On the other hand,  $\text{osc}_k(U_k, \mathcal{T}_k) = 0$ , and  $\eta_k(U_k, \mathcal{T}_k) \approx \|u - U_k\|_\Omega$  dictates the asymptotics for  $k > K$ . This is consistent with Lemma 5.4 and  $(u, f, \mathbf{D}) \in \mathbb{A}_1$ . However, 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  $\|u - U_k\|_\Omega$ .

In practice, this effect is typically less dramatic because  $f$  is not orthogonal to  $\mathbb{V}(\mathcal{T}_k)$ . Figure 5.1 displays the behavior of the AFEM for the smooth solution  $u_S$  of section 6.1 with frequencies 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 the approximation class  $\mathbb{A}_s$ , which can be characterized just by the estimator; compare (5.1).

According to Lemma 5.4, the new approximation class  $\mathbb{A}_s$  is completely determined by the standard approximation class  $\mathcal{A}_s$  for  $s \leq 1/d$ . Furthermore, in practice oscillation is often much smaller than the true error, whence the asymptotic rate  $N^{-s}$  would be determined by  $u \in \mathcal{A}_s$  for any  $s > 0$ . We point out the upper bound  $s \leq n/d$  for polynomial degree  $n \geq 1$  in dimension  $d$ , which can be easily seen for

uniform refinement; thus  $s = 1/d$  is the highest attainable order with piecewise linears  $n = 1$ . A near characterization of  $\mathcal{A}_s$  for  $d = 2$  in terms of Besov spaces can be found in [3, 4]:  $u \in \mathcal{A}_s$  implies that  $u \in B_p^{2s+1}(L^p(\Omega))$  for  $p = \frac{2}{2s+1}$  [3, Theorem 9.3];  $u \in B_p^{2s+1}(L^p(\Omega))$  for  $p > \frac{2}{2s+1}$  implies that  $u \in \mathcal{A}_s$  [3, Theorem 9.1]. Note that  $p < 1$  for  $s > 1/2$ .

To motivate the case  $n = 1$ ,  $d = 2$ ,  $s = 1/2$ , we consider the following heuristic argument. Assume that a mesh  $\mathcal{T} \in \mathbb{T}$  equidistributes the quantity  $\Lambda = \|D^2u\|_{L^1(\mathcal{T})}$ . Polynomial interpolation theory then yields

$$\|\nabla(u - I_{\mathcal{T}}u)\|_{\mathcal{T}} \preccurlyeq \|D^2u\|_{L^1(\omega_{\mathcal{T}})} \approx \Lambda \quad \forall T \in \mathcal{T},$$

and thus

$$\|\nabla(u - I_{\mathcal{T}}u)\|_{\Omega}^2 = \sum_{T \in \mathcal{T}} \|\nabla(u - I_{\mathcal{T}}u)\|_T^2 \preccurlyeq \Lambda^2 \#\mathcal{T}.$$

Since  $\|D^2u\|_{L^1(\Omega)} \approx \Lambda \#\mathcal{T}$ , we deduce the error decay

$$\|\nabla(u - I_{\mathcal{T}}u)\|_{\Omega} \preccurlyeq \|D^2u\|_{L^1(\Omega)} (\#\mathcal{T})^{-1/2}.$$

What this argument does not show is the existence of such a mesh  $\mathcal{T} \in \mathbb{T}$  created by bisection from  $\mathcal{T}_0$ . This explains the need for a strict inequality in the characterization above. However, we still have the following optimal result.

**LEMMA 5.6** (membership in  $\mathbb{A}_{1/2}$ ). *Let  $d = 2$ ,  $n = 1$ , and  $u$  be the solution of (1.1). Let  $\Omega$ ,  $\mathbf{A}$ ,  $c$ , and  $f$  satisfy the assumptions in section 2.1, and let  $\mathbf{A}$  be globally Lipschitz. Then  $(u, \mathbf{D}, f) \in \mathbb{A}_{1/2}$ .*

*Proof.* By applying [10, Theorem 5.2.2], we deduce that  $u \in W_p^2(\Omega)$  for some  $p > 1$ . The assertion follows upon combining [3, Theorem 9.1] for  $s = 1/2$  with Lemma 5.4.  $\square$

A complete characterization of  $\mathbb{A}_s$  for general  $d$  and  $n$  is still missing. It is important to realize that the nonlinear interaction between data  $\mathbf{D}$  and  $U$  must be accounted for, thereby leading to a new concept of approximation class  $\mathbb{A}_s$ , which generalizes those in [3, 4, 23].

*Remark 5.7* (conforming vs. nonconforming meshes). In contrast to [3, 23], our approach relies exclusively on conforming triangulations. When  $\mathbf{D} = (\mathbf{I}, 0)$  and  $n = 1$ , the approximation class  $\mathbb{A}_s$  is the same regardless of conformity [3]. The situation is quite different, however, in dealing with oscillation of a jump residual unless the depth of the nonconforming refinement is restricted beforehand.

Assume that  $(u, f, \mathbf{D}) \in \mathbb{A}_s$  for some  $s > 0$ , where typically  $s \leq n/d$ . We now prove that the approximation  $U_k$  generated by the AFEM converges to  $u$  with the same rate  $(\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s}$  as the best approximation described by  $\mathbb{A}_s$  up to a multiplicative constant. We need to count elements marked by the estimator (the cardinality of  $\mathcal{M}_k$ ) as well as those added to keep mesh conformity (see Lemma 2.3). To this end, we impose more stringent requirements than for convergence of the AFEM.

*Assumption 5.8* (optimality). We assume the following properties of the AFEM:

(a) The marking parameter  $\theta$  satisfies  $\theta \in (0, \theta_*)$ , with

$$\theta_*^2 = \frac{C_2}{1 + C_1(1 + 2\Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0))}.$$

(b) Procedure MARK selects a set  $\mathcal{M}_k$  of marked elements with *minimal* cardinality.

- (c) The distribution of refinement edges on  $\mathcal{T}_0$  satisfies condition (b) of section 4 in [24].

The limit value  $\theta_*$  depends on the ratio  $\sqrt{C_2/C_1} \leq 1$ , which quantifies the quality of estimator  $\eta_{\mathcal{T}_k}(U_k, \mathcal{T}_k)$ , as well as the oscillation  $\text{osc}_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)$  of coefficients of the PDE on  $\mathcal{T}_0$ .

**5.2. Cardinality of  $\mathcal{M}_k$ .** The following lemma establishes a link between non-linear approximation theory and the AFEM through the Dörfler marking strategy. Roughly speaking we prove that, if an approximation satisfies a suitable total error reduction from  $\mathcal{T}$  to  $\mathcal{T}_* \geq \mathcal{T}$ , the error indicators of the coarser solutions must satisfy a Dörfler property on the set  $\mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$ . In other words, Dörfler marking and total error reduction are intimately connected.

LEMMA 5.9 (optimal marking). *Assume that the marking parameter  $\theta$  verifies (a) of Assumption 5.8. Let  $\mathcal{T} \in \mathbb{T}$  and  $U \in \mathbb{V}(\mathcal{T})$  be the discrete solution of (2.3). Set  $\mu := \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$ , and let  $\mathcal{T}_* \in \mathbb{T}$  be any refinement of  $\mathcal{T}$ , i.e.,  $\mathcal{T} \leq \mathcal{T}_*$ , such that the discrete solution  $U_* \in \mathbb{V}(\mathcal{T}_*)$  satisfies*

$$(5.3) \quad \|u - U_*\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*) \leq \mu \left\{ \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \right\}.$$

Then the set  $\mathcal{R} := \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$  satisfies the Dörfler property

$$(5.4) \quad \eta_{\mathcal{T}}(U, \mathcal{R}) \geq \theta \eta_{\mathcal{T}}(U, \mathcal{T}).$$

*Proof.* We first combine the lower bound (2.7) with (5.3) to obtain

$$(5.5) \quad \begin{aligned} (1 - 2\mu) C_2 \eta_{\mathcal{T}}^2(U, \mathcal{T}) &\leq (1 - 2\mu) \left( \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \right) \\ &\leq \|u - U\|_{\Omega}^2 - \|u - U_*\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) - 2 \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*). \end{aligned}$$

We estimate separately the error and oscillation terms. We invoke the orthogonality (2.4) and the localized upper bound of Lemma 3.6 to arrive at

$$(5.6) \quad \|u - U\|_{\Omega}^2 - \|u - U_*\|_{\Omega}^2 = \|U_* - U\|_{\Omega}^2 \leq C_1 \eta_{\mathcal{T}}^2(U, \mathcal{R}).$$

For the oscillation terms we argue according to whether an element  $T \in \mathcal{T}$  belongs to the set of refined elements  $\mathcal{R}$  or not. For  $T \in \mathcal{R}$  we use the dominance

$$\text{osc}_{\mathcal{T}}^2(U, T) \leq \eta_{\mathcal{T}}^2(U, T)$$

of Remark 2.1. For  $T \in \mathcal{T} \cap \mathcal{T}_*$ , Corollary 3.5 with  $V = U$  and  $V_* = U_*$  yields

$$\text{osc}_{\mathcal{T}}^2(U, \mathcal{T} \cap \mathcal{T}_*) - 2 \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T} \cap \mathcal{T}_*) \leq 2\Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0) \|U_* - U\|_{\Omega}^2.$$

By combining these two estimates with (5.6) we infer that

$$(5.7) \quad \text{osc}_{\mathcal{T}}^2(U, \mathcal{T}) - 2 \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*) \leq (1 + 2C_1\Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0)) \eta_{\mathcal{T}}^2(U, \mathcal{R}).$$

By collecting (5.5), (5.6), and (5.7), we finally deduce that

$$\eta_{\mathcal{T}}^2(U, \mathcal{R}) \geq \frac{(1 - 2\mu) C_2}{1 + C_1(1 + 2\Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0))} \eta_{\mathcal{T}}^2(U, \mathcal{T}) = \theta^2 \eta_{\mathcal{T}}^2(U, \mathcal{T})$$

in light of the definitions of  $\theta_*$ ,  $\theta$ , and  $\mu$ . This concludes the proof.  $\square$

The key to relate the best mesh with AFEM triangulations is the fact that procedure MARK selects the marked set  $\mathcal{M}_k$  with *minimal* cardinality. This crucial idea, due to Stevenson [23], is used next.

LEMMA 5.10 (cardinality of  $\mathcal{M}_k$ ). *Assume that the marking parameter  $\theta$  verifies (a) and procedure MARK satisfies (b) of Assumption 5.8. Let  $u$  be the solution of (1.1), and let  $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k \geq 0}$  be the sequence of meshes, finite element spaces, and discrete solutions produced by the AFEM.*

*If  $(u, f, \mathbf{D}) \in \mathbb{A}_s$ , then the following estimate is valid:*

$$\#\mathcal{M}_k \leq \left(1 - \frac{\theta^2}{\theta_*^2}\right)^{-\frac{1}{2s}} |u, f, \mathbf{D}|_s^{\frac{1}{s}} C_{\mathbf{D}}^{\frac{1}{2s}} \left\{ \|u - U_k\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k) \right\}^{-\frac{1}{2s}}.$$

*Proof.* We set  $\epsilon^2 := \mu C_{\mathbf{D}}^{-1} (\|u - U_k\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k))$ , where  $\mu = \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$  and  $C_{\mathbf{D}}$  is the constant in Lemma 5.2. Since  $(u, f, \mathbf{D}) \in \mathbb{A}_s$ , there exists a  $\mathcal{T}_{\epsilon} \in \mathbb{T}$  and a  $V_{\epsilon} \in \mathbb{V}(\mathcal{T}_{\epsilon})$  such that

$$(5.8) \quad \#\mathcal{T}_{\epsilon} - \#\mathcal{T}_0 \leq |u, f, \mathbf{D}|_s^{1/s} \epsilon^{-1/s},$$

$$(5.9) \quad \|u - V_{\epsilon}\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_{\epsilon}}^2(V_{\epsilon}, \mathcal{T}_{\epsilon}) \leq \epsilon^2.$$

Let  $\mathcal{T}_* := \mathcal{T}_{\epsilon} \oplus \mathcal{T}_k$  be the overlay of  $\mathcal{T}_{\epsilon}$  and  $\mathcal{T}_k$ , and let  $U_* \in \mathbb{V}(\mathcal{T}_*)$  be the discrete solution of (2.3) on  $\mathcal{T}_*$ . To show that there is a reduction by a factor  $\mu$  of the total error between  $U_*$  and  $U_k$ , since  $\mathcal{T}_* \geq \mathcal{T}_{\epsilon}$ , we argue as in Lemma 5.2 and obtain

$$\begin{aligned} \|u - U_*\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*) &\leq C_{\mathbf{D}} \left\{ \|u - V_{\epsilon}\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_{\epsilon}}^2(V_{\epsilon}, \mathcal{T}_{\epsilon}) \right\} \\ &\leq C_{\mathbf{D}} \epsilon^2 = \mu \left\{ \|u - U_k\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k) \right\}. \end{aligned}$$

Hence, we deduce from Lemma 5.9 that the subset  $\mathcal{R} := \mathcal{R}_{\mathcal{T}_k \rightarrow \mathcal{T}_*} \subset \mathcal{T}_k$  verifies the Dörfler property (5.4) for  $\theta < \theta_*$ . The fact that procedure MARK selects a subset  $\mathcal{M}_k \subset \mathcal{T}_k$  with minimal cardinality satisfying the same property translates into

$$(5.10) \quad \#\mathcal{M}_k \leq \#\mathcal{R} \leq \#\mathcal{T}_* - \#\mathcal{T}_k \leq \#\mathcal{T}_{\epsilon} - \#\mathcal{T}_0,$$

where Lemma 3.7 has been employed in the last step. Finally, by combining (5.10), (5.8), and the definition of  $\epsilon$ , we end up with

$$\#\mathcal{M}_k \leq \#\mathcal{T}_{\epsilon} - \#\mathcal{T}_0 \leq \mu^{-\frac{1}{2s}} |u, f, \mathbf{D}|_s^{\frac{1}{s}} C_{\mathbf{D}}^{\frac{1}{2s}} \left\{ \|u - U_k\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k) \right\}^{-\frac{1}{2s}},$$

which is the asserted estimate.  $\square$

**5.3. Quasi optimality.** As a consequence of the previous estimates and the fact that the AFEM is a contraction for the quasi error, we obtain quasi optimality of the total error.

THEOREM 5.11 (quasi optimality). *Let Assumption 5.8 be satisfied by the AFEM. Let  $u$  be the solution of (1.1), and let  $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k \geq 0}$  be the sequence of meshes, finite element spaces, and discrete solutions produced by the AFEM.*

*Let  $(u, f, \mathbf{D}) \in \mathbb{A}_s$  and  $\Theta(\theta, s) := \theta^{-4s} (1 - \frac{\theta^2}{\theta_*^2})^{-1/2}$  describe the asymptotics of the AFEM as  $\theta \rightarrow \theta_*$ , 0 or  $s \rightarrow 0$ . Then there exists a constant  $C$ , depending on data, the refinement depth  $b$ , and  $\mathcal{T}_0$ , but independent of  $s$ , such that*

$$\left\{ \|u - U_k\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k) \right\}^{1/2} \leq C \Theta(\theta, s) |u, f, \mathbf{D}|_s (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s}.$$

*Proof.* By combining Lemmas 2.3 and 5.10 we deduce that

$$(5.11) \quad \#\mathcal{T}_k - \#\mathcal{T}_0 \preccurlyeq \sum_{j=0}^{k-1} \#\mathcal{M}_j \preccurlyeq M \sum_{j=0}^{k-1} \left\{ \|u - U_j\|_{\Omega}^2 + \text{osc}_j^2(U_j, \mathcal{T}_j) \right\}^{-\frac{1}{2s}},$$

with  $M := (1 - \frac{\theta^2}{\theta_0^2})^{-\frac{1}{2s}} |u, f, \mathbf{D}|_s^{\frac{1}{s}} C_{\mathbf{D}}^{\frac{1}{2s}}$ . We infer from the lower bound (2.7) that

$$(5.12) \quad \begin{aligned} \|u - U_j\|_{\Omega}^2 + \gamma \text{osc}_j^2(U_j, \mathcal{T}_j) &\leq \|u - U_j\|_{\Omega}^2 + \gamma \eta_j^2(U_j, \mathcal{T}_j) \\ &\leq \left(1 + \frac{\gamma}{C_2}\right) \left\{ \|u - U_j\|_{\Omega}^2 + \text{osc}_j^2(U_j, \mathcal{T}_j) \right\}. \end{aligned}$$

On the other hand, the linear rate  $\alpha = \alpha(\theta) < 1$  of convergence of Theorem 4.1 for the sum of the energy error and the scaled error estimator implies that for  $0 \leq j \leq k-1$

$$(5.13) \quad \|u - U_k\|_{\Omega}^2 + \gamma \eta_k^2(U_k, \mathcal{T}_k) \leq \alpha^{2(k-j)} \left\{ \|u - U_j\|_{\Omega}^2 + \gamma \eta_j^2(U_j, \mathcal{T}_j) \right\}.$$

We combine (5.11), (5.12), and (5.13) to obtain

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \preccurlyeq M \left(1 + \frac{\gamma}{C_2}\right)^{\frac{1}{2s}} \left\{ \|u - U_k\|_{\Omega}^2 + \gamma \eta_k^2(U_k, \mathcal{T}_k) \right\}^{-\frac{1}{2s}} \sum_{j=1}^k \alpha^{\frac{j}{s}}.$$

Since  $\alpha < 1$ , the geometric series is bounded by the constant  $S_{\theta} = \alpha^{1/s}(1 - \alpha^{1/s})^{-1}$ . By recalling that the element residual dominates the oscillation, we end up with

$$(5.14) \quad \#\mathcal{T}_k - \#\mathcal{T}_0 \preccurlyeq S_{\theta} M \left(1 + \frac{\gamma}{C_2}\right)^{\frac{1}{2s}} \left\{ \|u - U_k\|_{\Omega}^2 + \gamma \text{osc}_k^2(U_k, \mathcal{T}_k) \right\}^{-\frac{1}{2s}}.$$

To examine the asymptotics as  $\theta, s \rightarrow 0$ , hidden in  $S_{\theta}$ , we use Remark 4.4 to get  $\alpha \leq (1-t)^{\frac{1}{2}}$ , with  $t = c\theta^4$ , and observe that  $\lim_{t \rightarrow 0} \frac{t}{1-(1-t)^{\frac{1}{2s}}} = 2s$ . Therefore

$$S_{\theta}^s \leq \frac{\alpha}{t^s} \left( \frac{t}{1-(1-t)^{1/2s}} \right)^s \approx \alpha s^s \theta^{-4s},$$

whence  $S_{\theta}^s M^s \approx \Theta(\theta, s) |u, f, \mathbf{D}|_s C_{\mathbf{D}}^{1/2}$ . By raising (5.14) to the  $s$ th power and re-ordering, we obtain the desired estimate for the total error with  $C > 0$  independent of  $s$ .  $\square$

The same decay rate is obtained for the error estimator due to its equivalence to the total error.

**COROLLARY 5.12.** *Under the same requirements as in Theorem 5.11, there exists a constant  $C$  such that*

$$\eta_k \leq C \Theta(\theta, s) |u, f, \mathbf{D}|_s (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s}.$$

The constant  $C$  may differ from the one in Theorem 5.11.

**6. The critical role of marking for optimality.** We conclude the article by revisiting former optimality proofs by Binev, Dahmen, and DeVore [3] and Stevenson [23], which are based on the AFEM designed by Morin, Nochetto, and Siebert [17, 18], from now on called MNS. This is an adaptive loop of the form (1.2), which marks

separately for both the estimator and the oscillation in each iteration, which in turn yields a contraction for the energy error. It turns out that separate marking appears to be problematic for proving optimality, and thus variants of MNS are studied instead [3, 23]. The key issue is the delicate choice of marking parameters for the estimator and oscillation. To shed light on this intrinsic difficulty we first present some numerical experiments, then discuss the effect of separate marking, and finally comment on marking for an optimal AFEM.

**6.1. Separate marking and MNS.** The procedure ESTIMATE of MNS calculates both the error and oscillation indicators  $\{\eta_k(U_k, T), \text{osc}_k(U_k, T)\}_{T \in \mathcal{T}_k}$ , and the procedure MARK uses Dörfler marking for both the estimator and oscillation. More precisely, the routine MARK of MNS is of the form: *given parameters*  $0 < \theta_{\text{est}}, \theta_{\text{osc}} < 1$ ,

$$(6.1a) \quad \text{mark any subset } \mathcal{M}_k \subset \mathcal{T}_k \text{ such that } \eta_k(U_k, \mathcal{M}_k) \geq \theta_{\text{est}} \eta_k(U_k, \mathcal{T}_k);$$

$$(6.1b) \quad \text{if necessary enlarge } \mathcal{M}_k \text{ to satisfy } \text{osc}_k(U_k, \mathcal{M}_k) \geq \theta_{\text{osc}} \text{osc}_k(U_k, \mathcal{T}_k).$$

Since oscillation is generically of higher order than the estimator, the issue at stake is whether elements added by oscillation, even though immaterial relative to the error, could ruin the optimality of MNS observed in experiments [17, 18, 16]. If  $\eta_k(U_k, \mathcal{T}_k)$  has large indicators in a small area, then Dörfler marking for the estimator (6.1a) could select a set  $\mathcal{M}_k$  with a small number of elements relative to  $\mathcal{T}_k$ . However, if  $\text{osc}_k(U_k, \mathcal{T}_k)$  were globally distributed in  $\mathcal{T}_k$ , then MNS would require additional marking of a large percentage of all elements to satisfy (6.1b); i.e.,  $\#\mathcal{M}_k$  could be large relative to  $\#\mathcal{T}_k$ .

To explore this idea computationally, we consider a simple modification of the two-dimensional Example 5.3 in [12, 17]

$$-\text{div } \mathbf{A} \nabla u = f \quad \text{in } \Omega, \quad u = g \quad \text{on } \Omega,$$

where  $\Omega = (-1, 1)^2$  and  $\mathbf{A} := a_i \mathbf{I}$  is a piecewise constant (checkerboard pattern), with  $a_1 = a_3$  in the first and third quadrants and  $a_2 = a_4$  in the second and fourth quadrants. The exact solution  $u$  is given as a sum:  $u = u_K + u_S$ . Function  $u_K \in H^1(\Omega)$  is the weak solution of  $\text{div}(\mathbf{A} \nabla u_K) = 0$  for parameters [12, 17, 18]

$$a_1 = a_3 = 161.4476387975881, \quad a_2 = a_4 = 1;$$

the singularity  $u_K \approx r^{0.1}$  at the origin is so extreme that it typically leads to marking of a handful of elements per step to satisfy (6.1a) (see [17, 18] for details). Function  $u_S$  is given in each quadrant by

$$u_S(x, y) = 10^{-2} a_i^{-1} (x^2 + y^2) \sin^2(4\pi x) \sin^2(4\pi y), \quad 1 \leq i \leq 4,$$

is smooth, and is of comparable magnitude with  $u_K$ , while the corresponding  $f = -\text{div } \mathbf{A} \nabla u_S$  exhibits an increasing amount of data oscillation away from the origin. Procedure MARK takes the usual value of  $\theta_{\text{est}} = 0.5$  [8, 17, 18, 21], and procedure REFINESUBDIVIDES all elements in  $\mathcal{M}_k$  by using two bisections.

The behavior of MNS for several values of  $\theta_{\text{osc}}$  is depicted in Figure 6.1. We can visualize the sensitivity of MNS with respect to parameter  $\theta_{\text{osc}}$ . For values of  $\theta_{\text{osc}} \leq 0.4$  the rate of convergence appears to be quasi-optimal. However, beyond this threshold the curves for both the error and the estimator flatten out, thereby showing lack of optimality. The threshold value  $\theta_{\text{osc}} = 0.4$ , even though consistent with the practice of MNS, is tricky to find in general since it is problem-dependent.

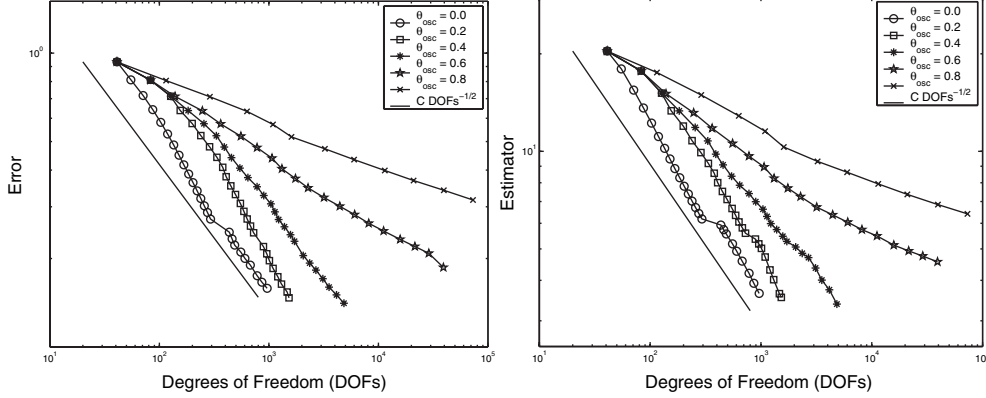


FIG. 6.1. Decay of the error (left) and the estimator (right) vs. DOFs for  $\theta_{\text{est}} = 0.5$  and values  $\theta_{\text{osc}} = 0.0, 0.2, 0.4, 0.6,$  and  $0.8$ . For values of  $\theta_{\text{osc}} \leq 0.4$  the rate of convergence is quasi-optimal, but for  $\theta_{\text{osc}} > 0.4$  the curves flatten out, thereby showing lack of optimality.

Lemma 6.1 below provides theoretical insight on this matter. Altogether this indicates that an optimality proof for MNS is out of reach without restrictions on  $\theta_{\text{osc}}$  and makes marking by oscillation (6.1b) questionable.

**6.2. Analysis of separate marking.** In this section we focus on the adaptive approximation of two given functions in an idealized scenario. We show that separate marking, similar to (6.1), may lead to suboptimal meshes in general. However, a suitable choice of marking parameters may restore optimality. The experiments of section 6.1 confirm this theoretical insight in the nonidealized case of MNS.

For the discussion, we assume that we have two functions  $u_i$ ,  $i = 1, 2$ , and have access to their local approximation error

$$e_{\mathcal{T}}(u_i; T) = |u_i - I_{\mathcal{T}}u_i|_{i;T} \quad \forall T \in \mathcal{T}$$

and global error  $e_{\mathcal{T}}^2(u_i) = \sum_{T \in \mathcal{T}} e_{\mathcal{T}}^2(u_i; T)$ ; hereafter  $|\cdot|_i$  are unspecified norms, and  $I_{\mathcal{T}}$  is a local interpolation operator over  $\mathcal{T} \in \mathbb{T}$ . We define the *total error* to be

$$e_{\mathcal{T}}^2 := e_{\mathcal{T}}^2(u_1) + e_{\mathcal{T}}^2(u_2)$$

and are interested in its asymptotic decay. If  $\mathcal{T} = \mathcal{T}_k$ , then we denote  $e_k = e_{\mathcal{T}_k}$ .

To explore the use of (6.1), we examine the effect of separate marking for  $e_k(u_i)$  on a sequence of meshes  $\mathcal{T}_k^i$  for  $i = 1, 2$ . We put ourselves in an idealized, but plausible, situation governed by the following three simplifying assumptions:

(6.2a) *Independence:*  $\mathcal{T}_k^1$  and  $\mathcal{T}_k^2$  are generated from  $\mathcal{T}_0$  and are independent of each other;

(6.2b) *Marking:* Separate Dörfler marking with parameters  $\theta_i \in (0, 1)$  implies that  $e_k(u_i) \approx \alpha_i^k$  on  $\mathcal{T}_k^i$ , with  $\alpha_i \in (0, 1)$ ;

(6.2c) *Approximability:*  $e_k(u_i) \approx (\#\mathcal{T}_k^i - \#\mathcal{T}_0)^{-s_i}$ , with  $s_1 \leq s_2$  maximal.

We are interested in the decay of the total error  $e_k$  on the overlay  $\mathcal{T}_k := \mathcal{T}_k^1 \oplus \mathcal{T}_k^2$ . This scenario is a simplification of the more realistic approximation of  $u_1$  and  $u_2$  with separate Dörfler marking on the same sequence of grids  $\mathcal{T}_k$  but avoids the complicated interaction of the two marking procedures.



LEMMA 6.1 (separate marking). *Let assumptions (6.2) be satisfied.*

*Then the decay of the total error  $e_k$  on the overlay  $\mathcal{T}_k = \mathcal{T}_k^1 \oplus \mathcal{T}_k^2$  for separate marking is always suboptimal except when  $\alpha_1$  and  $\alpha_2$  satisfy*

$$\alpha_2 \leq \alpha_1 \leq \alpha_2^{s_1/s_2}.$$

*Proof.* Assumption (6.2b) on the average reduction rate implies for the total error that

$$(6.3) \quad e_k \approx e_k(u_1) + e_k(u_2) \approx \max\{e_k(u_1), e_k(u_2)\} \approx \max\{\alpha_1^k, \alpha_2^k\}.$$

Combining (6.2b) and (6.2c) yields  $\alpha_i^k \approx (\#\mathcal{T}_k^i - \#\mathcal{T}_0)^{-s_i}$ , whence

$$(6.4) \quad \#\mathcal{T}_k^1 - \#\mathcal{T}_0 \approx \alpha_1^{-k/s_1} = \beta^k \alpha_2^{-k/s_2} \approx \beta^k (\#\mathcal{T}_k^2 - \#\mathcal{T}_0),$$

with  $\beta = \alpha_1^{-1/s_1} \alpha_2^{1/s_2}$ . In view of (2.2), this gives for the overlay  $\mathcal{T}_k = \mathcal{T}_k^1 \oplus \mathcal{T}_k^2$

$$(6.5) \quad \#\mathcal{T}_k - \#\mathcal{T}_0 \approx \begin{cases} \#\mathcal{T}_k^1 - \#\mathcal{T}_0, & \beta \geq 1, \\ \#\mathcal{T}_k^2 - \#\mathcal{T}_0, & \beta < 1. \end{cases}$$

The optimal decay of total error  $e_k$  corresponds to  $e_k \approx (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}$  because  $s_1 \leq s_2$ . In analyzing the relation of  $e_k$  to the number of elements  $\#\mathcal{T}_k$  in the overlay  $\mathcal{T}_k$ , we distinguish three cases and employ (6.3), (6.4), and (6.5).

*Case 1:*  $\alpha_1 < \alpha_2$ . We note that  $\alpha_1 < \alpha_2$  and  $s_1 \leq s_2$  yields  $\beta \geq 1$ . We thus deduce that

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_2^k = (\alpha_2/\alpha_1)^k \alpha_1^k \\ &\approx (\alpha_2/\alpha_1)^k (\#\mathcal{T}_k^1 - \#\mathcal{T}_0)^{-s_1} \approx (\alpha_2/\alpha_1)^k (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}. \end{aligned}$$

Since  $\alpha_2/\alpha_1 > 1$ , the approximation of  $e_k$  on  $\mathcal{T}_k$  is suboptimal.

*Case 2:*  $\alpha_1 \geq \alpha_2$  and  $\beta < 1$ . We obtain

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathcal{T}_k^1 - \#\mathcal{T}_0)^{-s_1} \\ &\approx \beta^{-ks_1} (\#\mathcal{T}_k^2 - \#\mathcal{T}_0)^{-s_1} \approx \beta^{-ks_1} (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}, \end{aligned}$$

whence the approximation of the total error on  $\mathcal{T}_k$  is again suboptimal.

*Case 3:*  $\alpha_1 \geq \alpha_2$  and  $\beta \geq 1$ . We infer that

$$e_k \approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathcal{T}_k^1 - \#\mathcal{T}_0)^{-s_1} \approx (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}$$

and that  $\mathcal{T}_k$  exhibits optimal cardinality. This exceptional case corresponds to the assertion and concludes the proof.  $\square$

We learn from Lemma 6.1 that separate marking requires a critical choice of parameters  $\theta_i$  to obtain optimal cardinality of grids with respect to the total error  $e_k$ . By revisiting MNS in light of Lemma 6.1, we could identify the estimator  $\eta_k$  with the error  $e_k(u_1)$  and oscillation  $\text{osc}_k$  with the error  $e_k(u_2)$ . We observe that  $\text{osc}_k \leq \eta_k$  combined with (6.2b) implies that  $\alpha_2 \leq \alpha_1$  and that  $\text{osc}_k$  is generically of higher order than  $\eta_k$ , thereby yielding  $s_1 < s_2$ .

We wonder whether or not the optimality condition  $\alpha_1 \leq \alpha_2^{s_1/s_2}$  is valid. Note that  $\alpha_2^{s_1/s_2}$  increases as the gap between  $s_1$  and  $s_2$  increases. Since the oscillation reduction estimate of [17] reveals that  $\alpha_2$  increases as  $\theta_{\text{osc}}$  decreases, we see that

separate marking may be optimal for a wide range of marking parameters  $\theta_{\text{est}}, \theta_{\text{osc}}$ ; this is confirmed by the numerical experiments in section 6.1 even though it is unclear whether  $\eta_k$  and  $\text{osc}_k$  satisfy (6.2). However, choosing marking parameters  $\theta_{\text{est}}, \theta_{\text{osc}}$  is rather tricky in practice because neither the explicit dependence of average reduction rates  $\alpha_1, \alpha_2$  on  $\theta_{\text{est}}, \theta_{\text{osc}}$  nor the optimal exponents  $s_1, s_2$  are known. In contrast to MNS, the standard AFEM of section 2.7 marks solely according to the estimator  $\eta_{\mathcal{T}}(U, \mathcal{T})$  and thus avoids separate marking.

We conclude by reviewing how an alternative optimal AFEM for the Poisson problem compensates for the lack of optimality of separate marking. Binev, Dahmen, and DeVore [3] added a coarsening step to MNS to prove optimality. Veerer [26] was the first to mark oscillation relative to the error estimator to prove convergence of the AFEM for the  $p$ -Laplacian. More recently, Stevenson [23] resorted to a similar marking to prove optimality of the AFEM at the expense of an additional inner loop to reduce data oscillation. Since for the more general elliptic PDE (1.1) oscillation  $\text{osc}_{\mathcal{T}}(U, \mathcal{T})$  interacts nonlinearly with the discrete solution  $U$ , it is not clear how to follow the approaches in [3, 23] with separate marking.

**Acknowledgments.** We thank the editor and the anonymous referees for their constructive comments that helped to improve the presentation.

#### REFERENCES

- [1] M. AINSWORTH AND J. T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.
- [2] E. BÄNSCH, *Local mesh refinement in 2 and 3 dimensions*, IMPACT Comput. Sci. Engrg., 3 (1991), pp. 181–191.
- [3] P. BINEV, W. DAHMEN, AND R. DEVORE, *Adaptive finite element methods with convergence rate*, Numer. Math., 97 (2004), pp. 219–268.
- [4] P. BINEV, W. DAHMEN, R. DEVORE, AND P. PETRUSHEV, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 1001–1026.
- [5] J. M. CASCON, C. KREUZER, R. H. NOCHETTO, AND K. G. SIEBERT, *Quasi-optimal AFEM for General Elliptic Operators*, manuscript.
- [6] Z. CHEN AND J. FENG, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2004), pp. 1167–1193.
- [7] L. DIENING AND C. KREUZER, *Linear Convergence of an Adaptive Finite Element Method for the  $p$ -Laplacian Equation*, Preprint 03/2007, Institut für Mathematik, Universität Freiburg, 2007.
- [8] W. DÖRFLER, *A convergent adaptive algorithm for Poisson’s equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [9] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
- [10] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
- [11] F. JOCHMANN, *An  $H^s$  regularity result for the gradient of solutions to elliptic equations with mixed boundary conditions*, J. Math. Anal. Appl., 238 (1999), pp. 429–450.
- [12] R. B. KELLOGG, *On the Poisson equation with intersecting interfaces*, Appl. Anal., 4 (1975), pp. 101–129.
- [13] I. KOSSACZKÝ, *A recursive approach to local mesh refinement in two and three dimensions*, J. Comput. Appl. Math., 55 (1994), pp. 275–288.
- [14] J. MAUBACH, *Local bisection refinement for  $n$ -simplicial grids generated by reflection*, SIAM J. Sci. Comput., 16 (1995), pp. 210–227.
- [15] W. F. MITCHELL, *A comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math. Software, 15 (1989), pp. 326–347.
- [16] 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.
- [17] 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.
- [18] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.

- [19] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Local problems on stars: a posteriori error estimation, convergence, and performance*, Math. Comp., 72 (2003), pp. 1067–1097.
- [20] P. MORIN, K.G. SIEBERT, AND A. VEESER, *A basic convergence result for conforming adaptive finite elements*, Math. Models Methods Appl., to appear.
- [21] A. SCHMIDT AND K. G. SIEBERT, *Design of Adaptive Finite Element Software: The Finite Element Toolbox ALBERTA*, Lect. Notes Comput. Sci. Eng. 42, Springer-Verlag, Berlin, 2005.
- [22] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
- [23] R. STEVENSON, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., DOI 10.1007/s10208-005-0183-0, (2006).
- [24] R. STEVENSON, *The Completion of Locally Refined Simplicial Partitions Created by Bisection*, Preprint 1336, Department of Mathematics, Utrecht University, 2005; corrected version, 2006.
- [25] C. T. TRAXLER, *An algorithm for adaptive mesh refinement in  $n$  dimensions*, Computing, 59 (1997), pp. 115–137.
- [26] A. VEESER, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math., 92 (2002), pp. 743–770.
- [27] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Technique*, Wiley–Teubner, Chichester, 1996.