ESAIM: M2AN 49 (2015) 1367–1398
DOI: 10.1051/m2an/2015017

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A CONVERGENT ADAPTIVE STOCHASTIC GALERKIN FINITE ELEMENT METHOD WITH QUASI-OPTIMAL SPATIAL MESHES *, **

Martin Eigel[1], Claude Jeffrey Gittelson[2], Christoph Schwab[3] and Elmar Zander[4]

Abstract. We analyze *a posteriori* error estimation and adaptive refinement algorithms for stochastic Galerkin Finite Element methods for countably-parametric, elliptic boundary value problems. A residual error estimator which separates the effects of gpc-Galerkin discretization in parameter space and of the Finite Element discretization in physical space in energy norm is established. It is proved that the adaptive algorithm converges. To this end, a contraction property of its iterates is proved. It is shown that the sequences of triangulations which are produced by the algorithm in the FE discretization of the active gpc coefficients are asymptotically optimal. Numerical experiments illustrate the theoretical results.

Mathematics Subject Classification. 65N30, 35R60, 47B80, 60H35, 65C20, 65N12, 65N22, 65J10.

Received December 27, 2013. Revised February 25, 2015.
Published online August 19, 2015.

1. Introduction

The efficient numerical solution of high-dimensional, parametric elliptic partial differential equations (PDEs for short) has attracted considerable attention in recent years, in particular in the context of uncertainty quantification (UQ), but also in connection with reduced basis approximation, optimization, and other computational techniques.

Depending on the particular goal of computation, numerical methods for parametric PDEs have particular advantages: we mention only the computation of ensemble averages (which take the form of integrals over the entire parameter space with respect to a probability measure on that space and which are treated by high-dimensional numerical integration), but also questions of optimization where a parsimonious, parametric

Keywords and phrases. Partial differential equations with random coefficients, generalized polynomial chaos, adaptive finite element methods, contraction property, residual *a posteriori* error estimation, uncertainty quantification.

* *Research supported in part by AFOSR, DOE, NNSA and NSF.*

** *Research supported in part by the European Research Council (ERC) under grant AdG247277 (to CS).*

[1] Weierstrass Institute, Mohrenstrasse 39, 10117 Berlin, Germany. martin.eigel@wias-berlin.de

[2] Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA. cgittels@purdue.edu

[3] Seminar for Applied Mathematics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland. schwab@sam.math.ethz.ch

[4] Institute of Scientific Computing, Technical University Braunschweig, 38092 Braunschweig, Germany. e.zander@tu-bs.de

numerical representation of the parametric solution with uniform, guaranteed accuracy on the entire parameter space is required.

A major issue in the design and analysis of efficient algorithms for these purposes has been the issue of *intrusive vs. nonintrusive* algorithms: the former are, roughly speaking, methods which require some degree of redesign of existing simulation code, whereas the latter rely on (possibly parallel) numerical solution with existing (sometime referred to as "legacy") code of the parametric PDEs in a number of (judiciously chosen) parameter values from a possibly infinite-dimensional parameter domain Γ. Examples include methods for numerical integration (*e.g.* [15, 19]) of mathematical expectations, and sparse, adaptive interpolation methods aiming at the adaptive computation of interpolants of the parametric PDE solution with uniform accuracy over the entire parameter spaces (*e.g.* [3, 4]).

As a rule, nonintrusive, collocation type methods are not amenable to reliable computable error bounds for the parametric surrogate solutions, likewise the results of approximate numerical integration; in order to ensure control of discretization errors in the context of UQ, therefore, the question of reliable or even guaranteed error bounds (in particular upper bounds) in the numerical solution of high-dimensional parametric PDE problems is of some interest. In the present paper, we continue our investigation [7] which analyzed intrusive so-called stochastic Galerkin discretizations of parametric elliptic PDEs. Here, approximations with respect to the parameter are achieved by *Galerkin projection* in mean square with respect to a probability measure π on the parameter domain Γ. Using Galerkin projections on generalized polynomial chaos bases on Γ instead of collocation of the parametric PDE problem requires modifications of the computational procedure which are, however, manageable in the context of Finite Element Methods (FEMs) for elliptic problems as we explained in [7]: most routines for generation of stiffness and mass matrices which are available in existing FE codes can be reused. In particular, due to the tensor product structure, the stiffness matrix corresponding to stochastic Galerkin discretization never needs to be formed explicitly, and efficient matrix-vector multiplications can be realized for the factored form of the matrix. Again, we refer to [7] for details on this. In that reference also the issue of *numerical a posteriori discretization error control has been addressed* and, in particular, reliable computable *a posteriori* error estimators for the (mean-square) discretization error have been derived. The possibility to treat high- or even infinite-dimensional problems efficiently by adaptive numerical methods is based on sparsity of coefficient sequences in polynomial chaos type expansions of the parametric solutions; we refer to [5] for sparsity results for the presently considered problems and to [10, 16] for a general introduction to the stochastic Galerkin FEM.

In the present work, we show that these error estimators have an intrinsic structure which allows to separate (in the sense of mean square with regard to the probability measure π in Γ and with respect to the natural energy inner product of the problem of interest) the contributions of the stochastic Galerkin discretization in the parameter domain as well as of the Finite Element discretization in the physical domain. With this separation at hand, we show that it is possible to design *adaptive refinement strategies in both the parameter domain Γ and the physical domain.* Also, we prove in the present paper convergence and certain optimality properties of such an adaptive refinement strategy. In particular, we show that the proposed strategy produces a sequence of finitely supported stochastic Galerkin FE solutions which converges in mean square with respect to π in Γ and with respect to the energy norm V in the physical domain, *and we establish that the FE mesh sequences generated by the proposed adaptive strategy for each of the gpc coefficients is, in a suitable sense, asymptotically optimal.*

The presented adaptive algorithm follows the `SOLVE-ESTIMATE-MARK-REFINE` paradigm which is well-established in the context of adaptive deterministic FEM. In particular, the convergence and optimality proofs are derived in the spirit of the seminal works [1, 2, 6, 17, 18, 20]. In these references, the fundamentally important concepts of oscillations, the contraction property and complexity classes were introduced.

As in [7], we consider here only an elementary, second order linearly elliptic problem in divergence form whose dependence on the parameter vector is affine. We hasten to add, however, that the principal conclusions of the present work also apply to more general, affine-parametric, linear elliptic problems, such as linear elasticity or Stokes, or parabolic evolution problems with parametric uncertainty as considered in [11].

The outline of the present paper is as follows: in Section 2, we specify the model problem and establish basic properties of its solution. Tensor product bases of FE bases and generalized polynomial chaos bases are introduced in Section 3. Section 4 then reviews the residual error estimator from [7] for the stochastic Galerkin truncation error, whereas Section 5 is devoted to computable error estimators for the spatial discretization error; here, we use a more or less standard residual error estimator, but remark that other error estimators can be used here as well. In Section 6, we present the adaptive stochastic Galerkin FEM algorithm. The algorithm is similar to the one proposed in [7], but differs from it in that a single finite element mesh is used for all active modes of the solution, as well as in several details which we have found to yield quantitative improvements in extensive numerical experiments which we performed since [7] (some of which are reported in the present paper's Sect. 9). Section 7 establishes the convergence of the adaptive algorithm (without rates), in particular the crucial contraction property. Section 8 establishes an optimality property of the iterates which are produced by the algorithm in the physical domain. The derivation of convergence and optimality mimics the results presented in [2] transferred to the parametric problem considered here. Finally, Section 9 contains several illustrative numerical examples.

2. Model problem

2.1. A parametric elliptic boundary value problem

For a bounded Lipschitz domain $D \subset \mathbb{R}^d$ and a function

$$a(y,x) = \bar{a}(x) + \sum_{m=1}^{\infty} y_m a_m(x), \quad x \in D, \tag{2.1}$$

depending on a sequence of scalar parameters y_m, and a function f on D, we consider the elliptic boundary value problem

$$\begin{cases} -\nabla\cdot(a\nabla u) = f & \text{in } D, \\ u = 0 & \text{on } \partial D. \end{cases} \tag{2.2}$$

For example, (2.1) may come from a Karhunen–Loève expansion of a random field. In order to ensure convergence in (2.1) and positivity of a, we assume $|y_m| \le 1$, *i.e.* $y := (y_m)_{m=1}^{\infty} \in \Gamma := [-1,1]^{\infty}$, and $\bar{a}, a_m \in W^{1,\infty}(D)$ with

$$\operatorname*{ess\,inf}_{x\in D} \bar{a}(x) > 0, \quad \sum_{m=1}^{\infty} \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(D)} \le \gamma < 1. \tag{2.3}$$

Let $V := H^1_0(D)$ with the $\bar{a}$-dependent norm $\|v\|_V := \sqrt{(v,v)_V}$ induced by the inner product

$$(w,v)_V := \int_D \bar{a}(x)\nabla w(x)\cdot\nabla v(x)\,\mathrm{d}x. \tag{2.4}$$

The parametric operator

$$A(y)\colon H^1_0(D) \to H^{-1}(D), \quad v \mapsto -\nabla\cdot(a(y)\nabla v), \quad y \in \Gamma, \tag{2.5}$$

can be expanded as

$$A(y) = \bar{A} + \sum_{m=1}^{\infty} y_m A_m, \quad y \in \Gamma, \tag{2.6}$$

with unconditional convergence in $\mathcal{L}(V,V^*)$ for the components

$$\bar{A}\colon H^1_0(D) \to H^{-1}(D), \quad v \mapsto -\nabla\cdot(\bar{a}\nabla v) \tag{2.7}$$

and

$$A_m \colon H_0^1(D) \to H^{-1}(D), \quad v \mapsto -\nabla \cdot (a_m \nabla v), \quad m \in \mathbb{N}. \tag{2.8}$$

The parametric operator equation

$$A(y)u(y) = f, \quad y \in \Gamma, \tag{2.9}$$

constitutes a weak formulation in space of the parametric boundary value problem (2.2).

2.2. Weak formulation

The weak formulation of (2.2) with respect to the parameter y requires a measure on the parameter domain $\Gamma = [-1,1]^\infty$. We consider symmetric product Borel measures; from a probabilistic point of view, this entails that the parameters y_m are independent and have symmetric distributions.

For each $m \in \mathbb{N}$, let π_m be a symmetric Borel probability measure[5] on $[-1,1]$; then

$$\pi := \bigotimes_{m=1}^{\infty} \pi_m \tag{2.10}$$

is a probability measure on Γ with the Borel σ-algebra. For the sake of clarity and ease of notation, we forbid the measures π_m from being finite convex combinations of Dirac measures, as this leads to finite instead of countably infinite gpc bases in Section 3.1 below.

Integrating (2.9) with respect to π over Γ leads to the weak formulation

$$\int_\Gamma \langle A(y)u(y), v(y)\rangle \,\mathrm{d}\pi(y) = \int_\Gamma \int_D f(x)v(y,x)\,\mathrm{d}x\,\mathrm{d}\pi(y) \quad \forall v \in L^2_\pi(\Gamma;V) \tag{2.11}$$

in the Lebesgue–Bochner space $L^2_\pi(\Gamma;V)$ of square-integrable, V-valued functions. The left hand side of (2.11) is a scalar product

$$(w,v)_{\mathcal{A}} := \int_\Gamma \langle A(y)w(y), v(y)\rangle \,\mathrm{d}\pi(y) = \int_\Gamma \int_D a(y,x)\nabla w(y,x)\cdot\nabla v(y,x)\,\mathrm{d}x\,\mathrm{d}\pi(y) \tag{2.12}$$

on $L^2_\pi(\Gamma;V)$, which induces the energy norm $\|\cdot\|_{\mathcal{A}}$ on this space. In particular, existence and uniqueness of the solution u of (2.11) are a consequence of the Riesz isomorphism, and u coincides with the solution of (2.9) for π-a.e. $y \in \Gamma$.

The operator

$$\mathcal{A} \colon L^2_\pi(\Gamma;V) \to L^2_\pi(\Gamma;V^*), \quad v \mapsto [y \mapsto A(y)v(y)] \tag{2.13}$$

allows (2.11) to be written succinctly as $\mathcal{A}u = f$, and the inner product (2.12) is $(w,v)_{\mathcal{A}} = \langle \mathcal{A}w, v\rangle$. Due to (2.6),

$$\mathcal{A} = \mathrm{id}_{L^2_\pi(\Gamma)} \otimes \bar{A} + \sum_{m=1}^{\infty} K_m \otimes A_m, \tag{2.14}$$

where[6] $K_m \colon L^2_\pi(\Gamma) \to L^2_\pi(\Gamma)$ refers to multiplication by y_m, which has operator norm at most 1 since $|y_m| \le 1$.

[5] *i.e.* π_m is invariant under the transformation $y_m \mapsto -y_m$.

[6] The tensor product $\otimes$ is meant with regards to the usual representation of the Bochner space $L^2_\pi(\Gamma;V)$ as the Hilbert tensor product space $L^2_\pi(\Gamma) \otimes V$, and similarly for V^* in place of V.

3. Galerkin approximation

3.1. Tensor product orthogonal polynomial basis

For each m, let $(P_n^m)_{n=0}^\infty$ denote an orthonormal polynomial basis of $L^2_{\pi_m}([-1,1])$ with $\deg(P_n^m) = n$. As a consequence of the symmetry of the measure π_m, such bases satisfy recursion formulas

$$\beta_n^m P_n^m(y_m) = y_m P_{n-1}^m(y_m) - \beta_{n-1}^m P_{n-2}^m(y_m), \quad n \geq 1, \tag{3.1}$$

with the initialization $P_0^m := 1$ and $\beta_0^m := 0$, and are unique *e.g.* if β_n^m are chosen as positive for all $n \geq 1$, which we assume.

In case of a uniform distribution $\mathrm{d}\pi_m(y_m) = \frac{1}{2}\,\mathrm{d}y_m$, the polynomials $(P_n^m)_{n=0}^\infty$ are Legendre polynomials, and $\beta_n^m = (4-n^{-2})^{-1/2}$. Alternatively, if $\mathrm{d}\pi_m(y_m) = \frac{1}{\pi}(1-y_m^2)^{-1/2}\,\mathrm{d}y_m$, then $(P_n^m)_{n=0}^\infty$ are Chebyshev polynomials of the first kind, with $\beta_1^m = 1/\sqrt{2}$ and $\beta_n^m = 1/2$ for $n \geq 2$. Further examples are tabulated *e.g.* in [9, 12].

Tensor products of the orthonormal polynomials P_n^m across all dimensions $m \in \mathbb{N}$ are indexed by the set

$$\mathcal{F} := \{\mu \in \mathbb{N}_0^\infty\,;\, \#\operatorname{supp}\mu < \infty\} \tag{3.2}$$

of finitely supported integer sequences, where $\operatorname{supp}(\mu) := \{m \in \mathbb{N}\,;\, \mu_m \neq 0\}$. For any $\mu \in \mathcal{F}$, the function $P_\mu := \bigotimes_{m=1}^\infty P_{\mu_m}^m$ is expressed as the finite product

$$P_\mu(y) = \prod_{m=1}^\infty P_{\mu_m}^m(y_m) = \prod_{m\in\operatorname{supp}\mu} P_{\mu_m}^m(y_m) \tag{3.3}$$

for $y = (y_m)_{m=1}^\infty \in \Gamma$ since $P_0^m = 1$ for all m due to the normalization of the measure π_m. The recursion (3.1) implies

$$y_m P_\mu(y) = \beta_{\mu_m+1}^m P_{\mu+\epsilon_m}(y) + \beta_{\mu_m}^m P_{\mu-\epsilon_m}(y), \quad y \in \Gamma, \tag{3.4}$$

where $\epsilon_m := (\delta_{mn})_{n=1}^\infty$ denotes the Kronecker sequence for the coordinate m, and we set $P_\mu := 0$ if any $\mu_m < 0$.

The tensorized polynomials $(P_\mu)_{\mu\in\mathcal{F}}$ form an orthonormal basis of $L^2_\pi(\Gamma)$. Equation (3.4) indicates the representation of the multiplication operator K_m in this basis.

Lemma 3.1. *The map* $\mathrm{K}_m\colon \ell^2(\mathcal{F}) \to \ell^2(\mathcal{F})$ *given by* $(c_\mu)_{\mu\in\mathcal{F}} \mapsto (\beta_{\mu_m+1}^m c_{\mu+\epsilon_m} + \beta_{\mu_m}^m c_{\mu-\epsilon_m})_{\mu\in\mathcal{F}}$ *has operator norm at most one.*

Proof. Due to (3.4), K_m is the representation of multiplication by y_m in the orthonormal basis $(P_\mu)_{\mu\in\mathcal{F}}$. By Parseval's identity, the operator norm of K_m on $\ell^2(\mathcal{F})$ coincides with that of K_m on $L^2_\pi(\Gamma)$, and this is at most 1 since $|y_m| \leq 1$. □

For any subset $\Lambda \subset \mathcal{F}$, we define $\operatorname{supp}(\Lambda) \subset \mathbb{N}$ as the set of active dimensions in Λ,

$$\operatorname{supp}\Lambda := \bigcup_{\mu\in\Lambda} \operatorname{supp}\mu. \tag{3.5}$$

The *boundary* of Λ is the infinite set

$$\partial\Lambda := \{\nu \in \mathcal{F}\setminus\Lambda\,;\, \exists m \in \mathbb{N}\colon \nu - \epsilon_m \in \Lambda \ \vee\ \nu + \epsilon_m \in \Lambda\}. \tag{3.6}$$

Restricting m in (3.6) to the support $\operatorname{supp}(\Lambda)$ leads to the *active boundary*

$$\partial^\circ\Lambda := \{\nu \in \mathcal{F}\setminus\Lambda\,;\, \exists m \in \operatorname{supp}\Lambda\colon \nu - \epsilon_m \in \Lambda \ \vee\ \nu + \epsilon_m \in \Lambda\}, \tag{3.7}$$

which is a finite set with cardinality at most $2(\#\operatorname{supp}\Lambda)\#\Lambda$ if Λ is finite.

A set $\Lambda \subset \mathcal{F}$ is *monotone*[7] if $\mu - \epsilon_m \in \Lambda$ for all $\mu \in \Lambda$ and $m \in \operatorname{supp}(\mu)$. If Λ is monotone, then $\partial\Lambda$ and $\partial^\circ\Lambda$ consist only of $\nu = \mu + \epsilon_m$ with $\mu \in \Lambda$, and consequently the cardinality of $\partial^\circ\Lambda$ is at most $(\#\operatorname{supp}\Lambda)\#\Lambda$. Neither our algorithm nor our convergence analysis require the monotonicity of the index sets of active gpc coefficients, nor does the presently proposed algorithm necessarily generate monotone sets. However, it is easily modified in order to ensure monotonicity, as indicated below. As shown in ([4], Thm. 4.3), for parametric diffusion problems under consideration here, the constraint of monotonicity and nestedness on the sets of active gpc coefficients does not reduce the N-term gpc approximation rate. Monotonicity of active index sets is desireable algorithmically, as it entails a tree structure for the sets of active indices where the boundary of the set is easily accessible.

3.2. Polynomial expansion

The expansion of the solution u of (2.11) with respect to the basis $(P_\mu)_{\mu\in\mathcal{F}}$ of $L^2_\pi(\Gamma)$ has the form

$$u(y,x) = \sum_{\mu\in\mathcal{F}} u_\mu(x) P_\mu(y), \tag{3.8}$$

with coefficients u_μ in $V = H^1_0(D)$ and convergence in $L^2_\pi(\Gamma;V)$. The vector of coefficients $(u_\mu)_{\mu\in\mathcal{F}} \in \ell^2(\mathcal{F};V)$ is determined by the infinite coupled system

$$\bar{A}u_\mu + \sum_{m=1}^{\infty} A_m(\beta^m_{\mu_m+1} u_{\mu+\epsilon_m} + \beta^m_{\mu_m} u_{\mu-\epsilon_m}) = f\delta_{\mu 0} \qquad \forall \mu \in \mathcal{F}. \tag{3.9}$$

The coefficients β^m_n in this system are the coefficients in the recursion formula (3.1).

For any subset $\Lambda \subset \mathcal{F}$, the Galerkin projection of u onto

$$\mathcal{V}(\Lambda) := \left\{ v_\Lambda(y,x) = \sum_{\mu\in\Lambda} v_{\Lambda,\mu}(x) P_\mu(y)\,;\ v_{\Lambda,\mu} \in V\ \forall \mu \in \Lambda \right\} \subset L^2_\pi(\Gamma;V) \tag{3.10}$$

is the unique $u_\Lambda \in \mathcal{V}(\Lambda)$ satisfying

$$\int_\Gamma \langle A(y)u_\Lambda(y), v_\Lambda(y)\rangle \,\mathrm{d}\pi(y) = \int_\Gamma \int_D f(x) v_\Lambda(y,x)\,\mathrm{d}x\,\mathrm{d}\pi(y) \quad \forall v_\Lambda \in \mathcal{V}(\Lambda). \tag{3.11}$$

If Λ is finite, then the sequence of coefficients $(u_{\Lambda,\mu})_{\mu\in\Lambda} \in V^\Lambda = \prod_{\mu\in\Lambda} V$ of u_Λ is determined by the finite system

$$\bar{A}u_{\Lambda,\mu} + \sum_{m=1}^{\infty} A_m(\beta^m_{\mu_m+1} u_{\Lambda,\mu+\epsilon_m} + \beta^m_{\mu_m} u_{\Lambda,\mu-\epsilon_m}) = f\delta_{\mu 0} \qquad \forall \mu \in \Lambda, \tag{3.12}$$

where $u_{\Lambda,\nu} := 0$ for $\nu \in \mathcal{F}\setminus\Lambda$. The infinite sum in (3.12) can be restricted to the finite set $\operatorname{supp}(\Lambda)$ since $u_{\Lambda,\mu\pm\epsilon_m} = 0$ for all $m \in \mathbb{N}\setminus\operatorname{supp}(\Lambda)$.

3.3. Finite element approximation

We discretize (2.11) further by restricting to a finite element space $V_p(\mathcal{T})$ of continuous piecewise polynomial functions of a fixed degree p on a conforming simplicial mesh $\mathcal{T}$ of D. For any finite set $\Lambda \subset \mathcal{F}$,

$$\mathcal{V}_p(\Lambda,\mathcal{T}) := \left\{ v_N(y,x) = \sum_{\mu\in\Lambda} v_{N,\mu}(x) P_\mu(y)\,;\ v_{N,\mu} \in V_p(\mathcal{T})\ \forall \mu \in \Lambda \right\} \subset \mathcal{V}(\Lambda) \tag{3.13}$$

[7]Monotone sets are sometimes termed *lower sets* or *downward closed sets*.

is a finite-dimensional subspace of $L^2_\pi(\Gamma;V)$, and the Galerkin approximation of u in $\mathcal{V}_p(\Lambda,\mathcal{T})$ is the unique $u_N \in \mathcal{V}_p(\Lambda,\mathcal{T})$ satisfying

$$\int_\Gamma \langle A(y)u_N(y), v_N(y)\rangle \,\mathrm{d}\pi(y) = \int_\Gamma \int_D f(x)v_N(y,x)\,\mathrm{d}x\,\mathrm{d}\pi(y) \quad \forall v_N \in \mathcal{V}_p(\Lambda,\mathcal{T}). \tag{3.14}$$

The sequence of coefficients $(u_{N,\mu})_{\mu\in\Lambda} \in V_p(\mathcal{T})^\Lambda = \prod_{\mu\in\Lambda} V_p(\mathcal{T})$ constitutes the finite element approximation of the system (3.12), determined by

$$\langle \bar{A}u_{N,\mu}, v_N\rangle + \sum_{m=1}^\infty \langle A_m(\beta^m_{\mu_m+1}u_{N,\mu+\epsilon_m} + \beta^m_{\mu_m}u_{N,\mu-\epsilon_m}), v_N\rangle = \langle f\delta_{\mu 0}, v_N\rangle \tag{3.15}$$

for all $v_N \in V_p(\mathcal{T})$ and all $\mu \in \Lambda$, where $u_{N,\nu} := 0$ for $\nu \in \mathcal{F}\setminus\Lambda$.

More specifically, we consider meshes resulting from refinements of a prescribed conforming simplicial mesh $\mathcal{T}_{\mathrm{init}}$ of D. For each cell $T \in \mathcal{T}_{\mathrm{init}}$, let a sequence of bisections of T into uniformly shape regular simplices be prescribed, and let $\mathbb{T}$ consist of all conforming simplicial meshes of D attainable through these bisections. We assume $\mathcal{T} \in \mathbb{T}$.

We denote the set of facets of the mesh $\mathcal{T}$ by $\mathcal{S} = \mathcal{S}(\mathcal{T})$, which are divided into interior facets $\mathcal{S}\cap D$ and exterior facets $\mathcal{S}\cap\partial D$. For any cell $T \in \mathcal{T}$, the set $\mathcal{S}\cap\partial T$ consists of the facets of $\mathcal{T}$ in the boundary of T. Similarly, for any $T \in \mathcal{T}$, $\partial T \cap D$ denotes the facets in the boundary of T in the interior of D.

We define local mesh size parameters by $h_T := |T|^{1/d}$ for $T \in \mathcal{T}$, and the resulting piecewise constant function $h_\mathcal{T}$ on $\mathcal{T}$ taking the value $h_\mathcal{T}(x) = h_T$ for $x \in T$.

The set $\mathbb{T}$ is partially ordered by the relation $\mathcal{T}_1 \preceq \mathcal{T}_2$ denoting that $\mathcal{T}_2$ is finer than $\mathcal{T}_1$, *i.e.* $\mathcal{T}_2$ can be obtained from $\mathcal{T}_1$ through a suitable refinement. Furthermore, for any $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$, the overlay $\mathcal{T} := \mathcal{T}_1 \oplus \mathcal{T}_2$ is the coarsest mesh in $\mathbb{T}$ with $\mathcal{T}_1 \preceq \mathcal{T}_1 \oplus \mathcal{T}_2$ and $\mathcal{T}_2 \preceq \mathcal{T}_1 \oplus \mathcal{T}_2$. By ([2], Lem. 3.7), the cardinality of $\mathcal{T}_1 \oplus \mathcal{T}_2$ is bounded by

$$\#(\mathcal{T}_1 \oplus \mathcal{T}_2) \le \#\mathcal{T}_1 + \#\mathcal{T}_2 - \#\mathcal{T}_0 \tag{3.16}$$

where $\mathcal{T}_0$ is any mesh $\mathcal{T}_0 \in \mathbb{T}$ with $\mathcal{T}_0 \preceq \mathcal{T}_1$ and $\mathcal{T}_0 \preceq \mathcal{T}_2$, *e.g.* $\mathcal{T}_0 = \mathcal{T}_{\mathrm{init}}$.

4. Estimation of the truncation error

4.1. Expansion of the residual

The residual $\mathcal{R}(w_\Lambda) \in L^2_\pi(\Gamma;V^*)$ of any approximation w_Λ of u in $\mathcal{V}(\Lambda)$ is

$$\mathcal{R}(w_\Lambda) := f - \mathcal{A}w_\Lambda = \mathcal{A}(u - w_\Lambda). \tag{4.1}$$

It can be expanded as $\mathcal{R}(w_\Lambda) = \sum_{\nu\in\mathcal{F}} r_\nu(w_\Lambda)P_\nu$ with convergence in $L^2_\pi(\Gamma;V^*)$ for the coefficients

$$r_\nu(w_\Lambda) = f\delta_{\nu 0} - \bar{A}w_{\Lambda,\nu} - \sum_{m=1}^\infty A_m(\beta^m_{\nu_m+1}w_{\Lambda,\nu+\epsilon_m} + \beta^m_{\nu_m}w_{\Lambda,\nu-\epsilon_m}), \quad \nu \in \mathcal{F}, \tag{4.2}$$

i.e.

$$\langle r_\nu(w_\Lambda), v\rangle = \int_D f\delta_{\nu 0}v - \sigma_\nu(w_\Lambda)\cdot\nabla v\,\mathrm{d}x \qquad \forall v \in V \tag{4.3}$$

for

$$\sigma_\nu(w_\Lambda) := \bar{a}\nabla w_{\Lambda,\nu} + \sum_{m=1}^\infty a_m\nabla(\beta^m_{\nu_m+1}w_{\Lambda,\nu+\epsilon_m} + \beta^m_{\nu_m}w_{\Lambda,\nu-\epsilon_m}), \quad \nu \in \mathcal{F}. \tag{4.4}$$

Noting that $r_\nu(w_\Lambda)$ is nonzero only for ν in $\Lambda\cup\partial\Lambda$, we have the decomposition $\mathcal{R}(w_\Lambda) = \mathcal{R}_\Lambda(w_\Lambda) + \mathcal{R}_{\partial\Lambda}(w_\Lambda)$ for

$$\mathcal{R}_\Xi(w_\Lambda) := \sum_{\nu\in\Xi} r_\nu(w_\Lambda)P_\nu, \quad \Xi \subset \mathcal{F}, \tag{4.5}$$

and consequently

$$\|\mathcal{R}(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)} = \|\mathcal{R}_\Lambda(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)} + \|\mathcal{R}_{\partial\Lambda}(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)}. \tag{4.6}$$

Lemma 4.1. *For any $w_\Lambda \in \mathcal{V}(\Lambda)$,*

$$\|w_\Lambda - u\|^2_{\mathcal{A}} \geq \frac{1}{1+\gamma}\big(\|\mathcal{R}_\Lambda(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)} + \|\mathcal{R}_{\partial\Lambda}(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)}\big), \tag{4.7}$$

$$\|w_\Lambda - u\|^2_{\mathcal{A}} \leq \frac{1}{1-\gamma}\big(\|\mathcal{R}_\Lambda(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)} + \|\mathcal{R}_{\partial\Lambda}(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)}\big). \tag{4.8}$$

Proof. By the Riesz representation theorem in $L^2_\pi(\Gamma;V^*)$,

$$\|u - w_\Lambda\|^2_{\mathcal{A}} = \sup_{v\in L^2_\pi(\Gamma;V)} \frac{|\langle \mathcal{A}(u-w_\Lambda), v\rangle|^2}{\|v\|^2_{\mathcal{A}}} = \sup_{v\in L^2_\pi(\Gamma;V)} \frac{|\langle \mathcal{R}(w_\Lambda), v\rangle|^2}{\|v\|^2_{\mathcal{A}}},$$

and $(1-\gamma)\|v\|^2_{L^2_\pi(\Gamma;V)} \leq \|v\|^2_{\mathcal{A}} \leq (1+\gamma)\|v\|^2_{L^2_\pi(\Gamma;V)}$ due to (2.3). The assertion follows with (4.6). □

The component $\|\mathcal{R}_\Lambda(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)}$ of (4.6) can be interpreted as an interior residual in the sense that it gauges the distance of w_Λ to u_Λ.

Lemma 4.2. *For any $w_\Lambda \in \mathcal{V}(\Lambda)$,*

$$\frac{1}{1+\gamma}\|\mathcal{R}_\Lambda(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)} \leq \|w_\Lambda - u_\Lambda\|^2_{\mathcal{A}} \leq \frac{1}{1-\gamma}\|\mathcal{R}_\Lambda(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)}. \tag{4.9}$$

Proof. For any $v_\Lambda \in \mathcal{V}(\Lambda)$,

$$\langle \mathcal{A}(u_\Lambda - w_\Lambda), v_\Lambda\rangle = \langle \mathcal{A}(u - w_\Lambda), v_\Lambda\rangle = \langle \mathcal{R}(w_\Lambda), v_\Lambda\rangle = \langle \mathcal{R}_\Lambda(w_\Lambda), v_\Lambda\rangle.$$

The assertion follows as in the proof of Lemma 4.1 using

$$\|u_\Lambda - w_\Lambda\|_{\mathcal{A}} = \sup_{v_\Lambda\in\mathcal{V}(\Lambda)} \frac{|\langle \mathcal{A}(u_\Lambda - w_\Lambda), v\rangle|}{\|v_\Lambda\|_{\mathcal{A}}} = \sup_{v_\Lambda\in\mathcal{V}(\Lambda)} \frac{|\langle \mathcal{R}_\Lambda(w_\Lambda), v_\Lambda\rangle|}{\|v\|_{\mathcal{A}}}.$$ □

Remark 4.3. Using Lemma 4.2, a statement similar to that of Lemma 4.1 for the Galerkin projection $w_\Lambda = u_N$ in a subspace of $\mathcal{V}(\Lambda)$ could be derived by means of Galerkin orthogonality

$$\|u_N - u\|^2_{\mathcal{A}} = \|u_N - u_\Lambda\|^2_{\mathcal{A}} + \|u_\Lambda - u\|^2_{\mathcal{A}}, \tag{4.10}$$

with each term on the right corresponding to one component of the residual. However, this leads to $\mathcal{R}_{\partial\Lambda}(u_\Lambda)$ in place of $\mathcal{R}_{\partial\Lambda}(u_N)$, which is less accessible.

We estimate the two terms of (4.6) separately, beginning with $\mathcal{R}_{\partial\Lambda}(w_\Lambda)$.

4.2. Upper bounds for the tail of the residual

Let $\Lambda \subset \mathcal{F}$ be a finite set. For any $w_\Lambda \in \mathcal{V}(\Lambda)$ and any $\nu \in \partial\Lambda$, let

$$\zeta_\nu(w_\Lambda) := \sum_{m=1}^{\infty} \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(D)} \big(\beta^m_{\nu_m+1}\|w_{\Lambda,\nu+\epsilon_m}\|_V + \beta^m_{\nu_m}\|w_{\Lambda,\nu-\epsilon_m}\|_V\big). \tag{4.11}$$

The sum in (4.11) is a finite sum over $\operatorname{supp}(\Lambda)$ since all other terms are zero. For any subset $\Delta \subset \partial\Lambda$, let

$$\zeta(w_\Lambda, \Delta) := \left(\sum_{\nu\in\Delta} \zeta_\nu(w_\Lambda)^2\right)^{1/2}. \tag{4.12}$$

Lemma 4.4. *If $0 \in \Lambda$, then for any $w_\Lambda \in \mathcal{V}(\Lambda)$,*

$$\|\mathcal{R}_{\partial\Lambda}(w_\Lambda)\|_{L^2_\pi(\Gamma;V^*)} \le \zeta(w_\Lambda, \partial\Lambda). \tag{4.13}$$

Proof. By Parseval's identity,

$$\|\mathcal{R}_{\partial\Lambda}(w_\Lambda)\|^2_{L^2_\pi(\Gamma;V^*)} = \sum_{\nu\in\partial\Lambda} \|r_\nu(w_\Lambda)\|^2_{V^*}.$$

Since $\nu \neq 0$, (4.3) and the Cauchy–Schwarz and triangle inequalities lead to

$$\|r_\nu(w_\Lambda)\|_{V^*} = \sup_{v\in V} \frac{1}{\|v\|_V} \left| \int_D \sigma_\nu(w_\Lambda)\cdot\nabla v \,\mathrm{d}x \right| \le \zeta_\nu(w_\Lambda). \qquad \square$$

Due to the infinite cardinality of $\partial\Lambda$, $\zeta(w_\Lambda, \partial\Lambda)$ is defined as an infinite sum in (4.12). However, for $\nu \in \partial\Lambda \setminus \partial^\circ\Lambda$, *i.e.* $\nu = \mu + \epsilon_m$ with $\mu \in \Lambda$ and $m \in \mathbb{N} \setminus \operatorname{supp}(\Lambda)$,

$$\zeta_\nu(w_\Lambda) = \left\| \frac{a_m}{\bar a} \right\|_{L^\infty(D)} \beta_1^m \|w_{\Lambda,\mu}\|_V. \tag{4.14}$$

Summing these terms over all inactive dimensions m leads to the lumped error indicator

$$\begin{aligned} \bar\zeta_\mu(w_\Lambda, \Lambda) &:= \left(\sum_{m\in\mathbb{N}\setminus\operatorname{supp}\Lambda} \zeta_{\mu+\epsilon_m}(w_\Lambda)^2 \right)^{1/2} \\ &= \|w_{\Lambda,\mu}\|_V \left(\sum_{m\in\mathbb{N}\setminus\operatorname{supp}\Lambda} \left(\left\| \frac{a_m}{\bar a} \right\|_{L^\infty(D)} \beta_1^m \right)^2 \right)^{1/2} \end{aligned} \tag{4.15}$$

for $\mu \in \Lambda$. The infinite sum remaining in $\bar\zeta_\mu(w_\Lambda, \Lambda)$ is independent of w_Λ and μ, depending only on $\operatorname{supp}(\Lambda)$; we assume that it can be computed. Then $\zeta(w_\Lambda, \partial\Lambda)$ is represented by the finite sum

$$\zeta(w_\Lambda, \partial\Lambda)^2 = \sum_{\nu\in\partial^\circ\Lambda} \zeta_\nu(w_\Lambda)^2 + \sum_{\mu\in\Lambda} \bar\zeta_\mu(w_\Lambda, \Lambda)^2. \tag{4.16}$$

4.3. Lipschitz continuity of the error indicator

The error indicator $\zeta(w_\Lambda, \partial\Lambda)$ depends Lipschitz-continuously on the approximation w_Λ in $\mathcal{V}(\Lambda)$.

Lemma 4.5. *For all $v_\Lambda, w_\Lambda \in \mathcal{V}(\Lambda)$,*

$$|\zeta(v_\Lambda, \partial\Lambda) - \zeta(w_\Lambda, \partial\Lambda)| \le \gamma \|v_\Lambda - w_\Lambda\|_{L^2_\pi(\Gamma;V)}. \tag{4.17}$$

Proof. Let $e_\Lambda := v_\Lambda - w_\Lambda \in \mathcal{V}(\Lambda)$. For any $\nu \in \partial\Lambda$,

$$\left|\zeta_\nu(v_\Lambda)^2 - \zeta_\nu(w_\Lambda)^2\right| = \left|\zeta_\nu(v_\Lambda) - \zeta_\nu(w_\Lambda)\right| \left(\zeta_\nu(v_\Lambda) + \zeta_\nu(w_\Lambda)\right) \le \zeta_\nu(e_\Lambda) s_\nu$$

with $s_\nu := \zeta_\nu(v_\Lambda) + \zeta_\nu(w_\Lambda)$. Appropriately rearranging terms and applying the Cauchy–Schwarz inequality, Lemma 3.1 and (2.3),

$$\begin{aligned} \sum_{\nu\in\partial\Lambda} \zeta_\nu(e_\Lambda) s_\nu &\le \sum_{\mu\in\Lambda} \|e_{\Lambda,\mu}\|_V \left[\sum_{m=1}^\infty \left\| \frac{a_m}{\bar a} \right\|_{L^\infty(D)} \left(\beta^m_{\mu_m+1} s_{\mu+\epsilon_m} + \beta^m_{\mu_m} s_{\mu-\epsilon_m} \right) \right] \\ &\le \gamma \left(\sum_{\mu\in\Lambda} \|e_{\Lambda,\mu}\|_V^2 \right)^{1/2} \left(\sum_{\nu\in\partial\Lambda} s_\nu^2 \right)^{1/2}, \end{aligned}$$

and $(\sum_{\nu\in\partial\Lambda} s_\nu^2)^{1/2} \leq \zeta(v_\Lambda,\partial\Lambda)+\zeta(w_\Lambda,\partial\Lambda)$ by the triangle inequality. The error indicator ζ satisfies

$$\begin{aligned}\big|\zeta(v_\Lambda,\partial\Lambda)-\zeta(w_\Lambda,\partial\Lambda)\big|\big(\zeta(v_\Lambda,\partial\Lambda)+\zeta(w_\Lambda,\partial\Lambda)\big) &= \big|\zeta(v_\Lambda,\partial\Lambda)^2-\zeta(w_\Lambda,\partial\Lambda)^2\big| \\ &\leq \sum_{\nu\in\partial\Lambda}\big|\zeta_\nu(v_\Lambda)^2-\zeta_\nu(w_\Lambda)^2\big|,\end{aligned}$$

and the assertion follows by inserting the above estimate for $|\zeta_\nu(v_\Lambda)^2-\zeta_\nu(w_\Lambda)^2|$ and cancelling $\zeta(v_\Lambda,\partial\Lambda)+\zeta(w_\Lambda,\partial\Lambda)$ since $\sum_{\mu\in\Lambda}\|e_{\Lambda,\mu}\|_V^2 = \|e_\Lambda\|^2_{L^2_\pi(\Gamma;V)}$. □

5. A spatial error indicator

5.1. Residual-based estimation of the spatial error

For all $w_N\in\mathcal{V}_p(\Lambda,\mathcal{T})$, $T\in\mathcal{T}$ and $\mu\in\Lambda$, let

$$\eta_{\mu,T}(w_N) := \big(h_T^2\|\bar a^{-1/2}(f\delta_{\mu 0}+\nabla\cdot\sigma_\mu(w_N))\|^2_{L^2(T)} + h_T\|\bar a^{-1/2}[\![\sigma_\mu(w_N)]\!]\|^2_{L^2(\partial T\cap D)}\big)^{1/2}, \tag{5.1}$$

where $[\![\cdot]\!]$ denotes the normal jump over $S\in\mathcal{S}(\mathcal{T})$, *i.e.* if $\bar S=\bar T_1\cap\bar T_2$ and n_i is the exterior unit normal to T_i, then

$$[\![\sigma]\!] := \sigma|_{T_1}\cdot n_1+\sigma|_{T_2}\cdot n_2. \tag{5.2}$$

Summing over $\mu\in\Lambda$, we define the error indicator for the cell T as

$$\eta_T(w_N,\Lambda) := \left(\sum_{\mu\in\Lambda}\eta_{\mu,T}(w_N)^2\right)^{1/2}, \tag{5.3}$$

and for any subset $\mathcal{M}\subset\mathcal{T}$, these terms combine to

$$\eta(w_N,\Lambda,\mathcal{M}) := \left(\sum_{T\in\mathcal{M}}\eta_T(w_N,\Lambda)^2\right)^{1/2}. \tag{5.4}$$

Similarly, we define the oscillation of $w_N\in\mathcal{V}_p(\Lambda,\mathcal{T})$ as

$$\begin{aligned}\operatorname{osc}_{\mu,T}(w_N) := \big(&h_T^2\|\bar a^{-1/2}(\mathrm{id}-\Pi_{2p-2})(f\delta_{\mu 0}+\nabla\cdot\sigma_\mu(w_N))\|^2_{L^2(T)} \\ &+ h_T\|\bar a^{-1/2}(\mathrm{id}-\Pi_{2p-1})[\![\sigma_\mu(w_N)]\!]\|^2_{L^2(\partial T\cap D)}\big)^{1/2},\end{aligned} \tag{5.5}$$

where p is the local polynomial degree of the finite element space $V_p(\mathcal{T})$ and Π_n denotes the orthogonal projection in $L^2(T)$ with respect to the weight $\bar a^{-1}$ onto polynomials of degree n. Summing over $\mu\in\Lambda$ and $T\in\mathcal{M}\subset\mathcal{T}$ gives the total oscillations

$$\operatorname{osc}_T(w_N,\Lambda) := \left(\sum_{\mu\in\Lambda}\operatorname{osc}_{\mu,T}(w_N)^2\right)^{1/2}, \tag{5.6}$$

$$\operatorname{osc}(w_N,\Lambda,\mathcal{M}) := \left(\sum_{T\in\mathcal{M}}\operatorname{osc}_T(w_N,\Lambda)^2\right)^{1/2}, \tag{5.7}$$

where $\mathcal{M}$ is any nonempty subset of $\mathcal{T}$. These terms are used only in our analysis, and do not need to be computed in our adaptive algorithm. We note that the error indicator dominates the oscillation,

$$\operatorname{osc}_T(w_N,\Lambda)\leq\eta_T(w_N,\Lambda) \tag{5.8}$$

for all $T\in\mathcal{T}$, see ([2], Rem. 2.1).

5.2. Equivalence to the interior residual

Up to a term involving the oscillation in the lower bound, the spatial error indicator is equivalent to the residual of the Galerkin projection in $\mathcal{V}_p(\Lambda,\mathcal{T})$. The constants c_η and C_η appearing in Theorem 5.1 are independent of the set Λ of active indices since, as described in the proof, bounds for each coefficient of the residual hold with uniform constants.

Theorem 5.1. *The Galerkin projection u_N of u onto $\mathcal{V}_p(\Lambda,\mathcal{T})$ satisfies*

$$c_\eta\big(\eta(u_N,\Lambda,\mathcal{T})^2-\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2\big)\le\|\mathcal{R}_\Lambda(u_N)\|^2_{L^2_\pi(\Gamma;V^*)}\le C_\eta\eta(u_N,\Lambda,\mathcal{T})^2 \tag{5.9}$$

with constants $c_\eta,C_\eta>0$ depending only on $\bar a$, p and the shape regularity of $\mathbb{T}$, but not on Λ.

Proof. For any $\mu\in\Lambda$, the proof of ([7], Thm. 6.1) extends verbatim to arbitrary polynomial degrees p to show

$$|\langle r_\mu(u_N),v-\mathcal{I}_Nv\rangle|^2\le C_\eta\|v\|_V^2\sum_{T\in\mathcal{T}}\eta_{\mu,T}(u_N)^2$$

for all $v\in V$, where $\mathcal{I}_N$ denotes the Clément quasi-interpolation operator onto $V_p(\mathcal{T})$. By Galerkin orthogonality, $\langle r_\mu(u_N),v\rangle=\langle r_\mu(u_N),v-\mathcal{I}_Nv\rangle$, and thus

$$\|r_\mu(u_N)\|^2_{V^*}\le C_\eta\sum_{T\in\mathcal{T}}\eta_{\mu,T}(u_N)^2.$$

Similarly, the standard estimates from [18, 21] based on cell and facet bubble functions lead to the lower bound

$$\left(\sum_{T\in\mathcal{T}}\eta_{\mu,T}(u_N)^2\right)^{1/2}\le c\left[\|r_\mu(u_N)\|_{V^*}+\left(\sum_{T\in\mathcal{T}}\operatorname{osc}_{\mu,T}(u_N)^2\right)^{1/2}\right]$$

for all $\mu\in\Lambda$. Consequently,

$$c_\eta\left[\sum_{T\in\mathcal{T}}\eta_{\mu,T}(u_N)^2-\sum_{T\in\mathcal{T}}\operatorname{osc}_{\mu,T}(u_N)^2\right]\le\|r_\mu(u_N)\|^2_{V^*}$$

for $c_\eta=1/2c^2$, and the assertion follows by summing over $\mu\in\Lambda$. □

Theorem 5.1 and Lemma 4.2 provide the following bounds for the spatial error of $u_N\in\mathcal{V}_p(\Lambda,\mathcal{T})$, *i.e.* the energy norm of the difference between u_N and the semidiscrete approximation u_Λ.

Corollary 5.2. *The Galerkin projection u_N in $\mathcal{V}_p(\Lambda,\mathcal{T})$ satisfies*

$$\frac{c_\eta}{1+\gamma}\big(\eta(u_N,\Lambda,\mathcal{T})^2-\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2\big)\le\|u_N-u_\Lambda\|^2_{\mathcal{A}}\le\frac{C_\eta}{1-\gamma}\eta(u_N,\Lambda,\mathcal{T})^2. \tag{5.10}$$

Similarly, Lemmas 4.1, 4.4 and Theorem 5.1 lead to the following upper and lower bounds for the full error of u_N in the energy norm.

Corollary 5.3. *The energy norm error of the Galerkin projection u_N in $\mathcal{V}_p(\Lambda,\mathcal{T})$ satisfies*

$$\|u_N-u\|^2_{\mathcal{A}}\ge\frac{c_\eta}{1+\gamma}\big(\eta(u_N,\Lambda,\mathcal{T})^2-\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2\big), \tag{5.11}$$

$$\|u_N-u\|^2_{\mathcal{A}}\le\frac{C_\eta}{1-\gamma}\big(\eta(u_N,\Lambda,\mathcal{T})^2+\zeta(u_N,\partial\Lambda)^2\big). \tag{5.12}$$

The upper bound from Corollary 5.2 can be refined to estimate the difference of two discrete solutions with different spatial meshes. In this case, the error indicator is restricted to just the refined elements, and the estimate can thus be viewed as a local upper bound. We refer to ([2], Lem. 3.6) for a proof.

Lemma 5.4. *Let $\mathcal{T}, \mathcal{T}^* \in \mathbb{T}$ such that $\mathcal{T}^*$ is a refinement of $\mathcal{T}$, and let $u_N \in \mathcal{V}_p(\Lambda, \mathcal{T})$ and $u_N^* \in \mathcal{V}_p(\Lambda, \mathcal{T}^*)$ be the respective Galerkin projections. Then*

$$\|u_N - u_N^*\|_{\mathcal{A}}^2 \leq \bar{C}_\eta \eta(u_N, \Lambda, \mathcal{M})^2 \tag{5.13}$$

where $\mathcal{M} = \mathcal{T} \setminus (\mathcal{T}^* \cap \mathcal{T})$ is the set of refined cells and $\bar{C}_\eta$ is a uniform constant on $\mathbb{T}$ independent of Λ.

5.3. Lipschitz continuity of the spatial error indicator

Similarly to the error indicator $\zeta(w_N, \partial\Lambda)$, the spatial error indicator $\eta_T(w_N, \Lambda)$ depends Lipschitz-continuously on the argument w_N in $\mathcal{V}_p(\Lambda, \mathcal{T})$.

For any finite set $\Lambda \subset \mathcal{F}$ and any $\mathcal{T} \in \mathbb{T}$, we introduce the constant

$$c_{a,\delta}(\Lambda, \mathcal{T}) := \max\left\{ \left\| \frac{h_{\mathcal{T}} \nabla \varphi}{\bar{a}} \right\|_{L^\infty(D)} \Big/ \left\| \frac{\varphi}{\bar{a}} \right\|_{L^\infty(D)} ; \, \varphi \in \{\bar{a}\} \cup \{a_m \,;\, m \in \operatorname{supp} \Lambda\} \right\}, \tag{5.14}$$

i.e. the gradients of all a_m with $m \in \operatorname{supp}(\Lambda)$ satisfy

$$\left\| \frac{h_{\mathcal{T}} \nabla a_m}{\bar{a}} \right\|_{L^\infty(D)} \leq c_{a,\delta}(\Lambda, \mathcal{T}) \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(D)} \tag{5.15}$$

and the same estimate holds for $\bar{a}$ in place of a_m. This constant is always finite since $\operatorname{supp}(\Lambda)$ is a finite set, but $c_{a,\delta}(\Lambda, \mathcal{T})$ may degenerate if Λ is enlarged without appropriate refinements of $\mathcal{T}$.

The proof of the following statement mirrors that of Lemma 4.5. The seminorm $|\cdot|_{L^2_\pi(\Gamma; V|T)}$ refers to the restriction of the Bochner norm in $L^2_\pi(\Gamma; V)$ to any subdomain $T \subset D$, which in the following will be a triangular or tetrahedral element $T \in \mathcal{T}$.

Lemma 5.5. *For all $v_N, w_N \in \mathcal{V}_p(\Lambda, \mathcal{T})$ and all $T \in \mathcal{T}$,*

$$|\eta_T(v_N, \Lambda) - \eta_T(w_N, \Lambda)| \leq \big(c_{a,\delta}(\Lambda, \mathcal{T}) + \hat{c}_\eta\big)(1+\gamma) |v_N - w_N|_{L^2_\pi(\Gamma; V|T)} \tag{5.16}$$

with a uniform constant $\hat{c}_\eta$ (depending only on $\mathbb{T}$).

Proof. Let $\mu \in \Lambda$ and $e_N := v_N - w_N$. We split $\eta_{\mu,T}(w_N)$ into $\eta^0_{\mu,T}(w_N) := h_T \|\bar{a}^{-1/2}(f\delta_{\mu 0} + \nabla \cdot \sigma_\mu(w_N))\|_{L^2(T)}$ and $\eta^1_{\mu,T}(w_N) := h_T^{1/2} \|\bar{a}^{-1/2} [\![\sigma_\mu(w_N)]\!]\|_{L^2(\partial T \cap D)}$.

Let $c_{\mathrm{inv}} > 0$ such that, uniformly for all $\mathcal{T} \in \mathbb{T}$ and all $T \in \mathcal{T}$, $\|\bar{a}^{1/2} \Delta v_N\|_{L^2(T)} \leq c_{\mathrm{inv}} h_T^{-1} |v_N|_{V,T}$ and $\|\bar{a}^{1/2} \nabla v_N \cdot n_T\|_{L^2(\partial T \cap D)} \leq c_{\mathrm{inv}} h_T^{-1/2} |v_N|_{V,T}$ for all $v_N \in V_p(\mathcal{T})$.

The first of the above inverse inequalities $\|\bar{a}^{1/2} \Delta v_N\|_{L^2(T)} \leq c_{\mathrm{inv}} h_T^{-1} |v_N|_{V,T}$ for $v_N \in V_p(\mathcal{T})$ implies

$$\begin{aligned} |\eta^0_{\mu,T}(v_N) - \eta^0_{\mu,T}(w_N)| &\leq h_T \|\bar{a}^{-1/2} \nabla \cdot \sigma_\mu(e_N)\|_{L^2(T)} \\ &\leq \alpha_0^0 |e_{N,\mu}|_{V,T} + \sum_{m=1}^\infty \alpha_m^0 \big(\beta^m_{\mu_m+1} |e_{N,\mu+\epsilon_m}|_{V,T} + \beta^m_{\mu_m} |e_{N,\mu-\epsilon_m}|_{V,T}\big) \end{aligned}$$

for $\alpha_0^0 := c_{a,\delta}(\Lambda, \mathcal{T}) + c_{\mathrm{inv}}$ and $\alpha_m^0 := (c_{a,\delta}(\Lambda, \mathcal{T}) + c_{\mathrm{inv}}) \|a_m/\bar{a}\|_{L^\infty(D)}$. Furthermore, using that $\|\bar{a}^{1/2} \nabla v_N \cdot n_T\|_{L^2(\partial T \cap D)} \leq c_{\mathrm{inv}} h_T^{-1/2} |v_N|_{V,T}$ for all $v_N \in V_p(\mathcal{T})$,

$$\begin{aligned} |\eta^1_{\mu,T}(v_N) - \eta^1_{\mu,T}(w_N)| &\leq h_T^{1/2} \|\bar{a}^{-1/2} [\![\sigma_\mu(e_N)]\!]\|_{L^2(\partial T \cap D)} \\ &\leq \alpha_0^1 |e_{N,\mu}|_{V,T} + \sum_{m=1}^\infty \alpha_m^1 \big(\beta^m_{\mu_m+1} |e_{N,\mu+\epsilon_m}|_{V,T} + \beta^m_{\mu_m} |e_{N,\mu-\epsilon_m}|_{V,T}\big) \end{aligned}$$

with $\alpha_0^1 := 2c_{\mathrm{inv}}$ and $\alpha_m^1 := 2c_{\mathrm{inv}} \|a_m/\bar{a}\|_{L^\infty(D)}$.

Noting that

$$\big|\eta_{\mu,T}(v_N)^2-\eta_{\mu,T}(w_N)^2\big| = \big|\eta^0_{\mu,T}(v_N)-\eta^0_{\mu,T}(w_N)\big|s^0_\mu + \big|\eta^1_{\mu,T}(v_N)-\eta^1_{\mu,T}(w_N)\big|s^1_\mu$$

for $s^i_\mu := \eta^i_{\mu,T}(v_N)+\eta^i_{\mu,T}(w_N)$, the above estimates combine to

$$\begin{aligned}|\eta_T(v_N,\Lambda)^2-\eta_T(w_N,\Lambda)^2| &\le \sum_{\mu\in\Lambda}\big|\eta_{\mu,T}(v_N)^2-\eta_{\mu,T}(w_N)^2\big|\\ &\le \sum_{\mu\in\Lambda}|e_{N,\mu}|_{V,T}S_\mu \le \left(\sum_{\mu\in\Lambda}|e_{N,\mu}|^2_{V,T}\right)^{1/2}\left(\sum_{\mu\in\Lambda}S_\mu^2\right)^{1/2}\end{aligned}$$

with

$$\begin{aligned}S_\mu = {}& \alpha^0_0 s^0_\mu + \sum_{m=1}^\infty \alpha^0_m\big(\beta^m_{\mu_m+1}s^0_{\mu+\epsilon_m}+\beta^m_{\mu_m}s^0_{\mu-\epsilon_m}\big)\\ &+\alpha^1_0 s^1_\mu + \sum_{m=1}^\infty \alpha^1_m\big(\beta^m_{\mu_m+1}s^1_{\mu+\epsilon_m}+\beta^m_{\mu_m}s^1_{\mu-\epsilon_m}\big)\end{aligned}$$

and, due to Lemma 3.1,

$$\begin{aligned}\left(\sum_{\mu\in\Lambda}S_\mu^2\right)^{1/2} &\le \left(\alpha^0_0+\sum_{m=1}^\infty\alpha^0_m\right)\left(\sum_{\mu\in\Lambda}(s^0_\mu)^2\right)^{1/2} + \left(\alpha^1_0+\sum_{m=1}^\infty\alpha^1_m\right)\left(\sum_{\mu\in\Lambda}(s^1_\mu)^2\right)^{1/2}\\ &\le \left(\alpha^0_0+\alpha^1_0+\sum_{m=1}^\infty\alpha^0_m+\alpha^1_m\right)\big(\eta_T(v_N,\Lambda)+\eta_T(w_N,\Lambda)\big).\end{aligned}$$

The assertion follows with $\hat c_\eta = 3c_{\mathrm{inv}}$ using

$$|\eta_T(v_N,\Lambda)^2-\eta_T(w_N,\Lambda)^2| = \big|\eta_T(v_N,\Lambda)-\eta_T(w_N,\Lambda)\big|\big(\eta_T(v_N,\Lambda)+\eta_T(w_N,\Lambda)\big).$$ □

The spatial error indicators are also continuous in their second argument, as described in the following statement.

Lemma 5.6. *Let $0\in\Lambda\subset\Lambda^*\subset\mathcal{F}$, $\mathcal{T}\in\mathbb{T}$ and $w_N\in\mathcal{V}_p(\Lambda,\mathcal{T})$. Then*

$$\eta(w_N,\Lambda^*\setminus\Lambda,\mathcal{T}) \le \big(2c_{a,\delta}(\Lambda^*,\mathcal{T})+\hat c_{\eta,\zeta}\big)\zeta(w_N,\partial\Lambda\cap\Lambda^*) \tag{5.17}$$

with a uniform constant $\hat c_{\eta,\zeta}$ on $\mathbb{T}$.

Proof. By definition, using $\eta_{\nu,T}(w_N)=0$ for $\nu\in\Lambda^*\setminus(\Lambda\cup\partial\Lambda)$,

$$\eta(w_N,\Lambda^*\setminus\Lambda,\mathcal{T})^2 = \sum_{T\in\mathcal{T}}\sum_{\nu\in\partial\Lambda\cap\Lambda^*}\eta_{\nu,T}(w_N)^2$$

As in the proof of Lemma 5.5, we split $\eta_{\nu,T}(w_N)$ into $\eta^0_{\nu,T}(w_N) := h_T\|\bar a^{-1/2}(f\delta_{\nu 0}+\nabla\cdot\sigma_\nu(w_N))\|_{L^2(T)}$ and $\eta^1_{\nu,T}(w_N) := h_T^{1/2}\|\bar a^{-1/2}[\![\sigma_\nu(w_N)]\!]\|_{L^2(\partial T\cap D)}$ for any $\nu\in\partial\Lambda\cap\Lambda^*$ and $T\in\mathcal{T}$.

Let $c_{\mathrm{inv}}>0$ such that the inverse inequalities $\|\bar a^{1/2}h_{\mathcal{T}}\Delta v_N\|_{L^2(D)}\le c_{\mathrm{inv}}\|v_N\|_V$ and $\sum_{T\in\mathcal{T}}h_T\|\bar a^{1/2}\nabla v_N\cdot n_T\|_{L^2(\partial T\cap D)}\le c^2_{\mathrm{inv}}\|v_N\|_V$ hold for all $v_N\in V_p(\mathcal{T})$ uniformly on $\mathbb{T}$.

The former inverse inequality and $w_{N,\nu} = 0$ imply

$$\begin{aligned}\left(\sum_{T\in\mathcal{T}} \eta^0_{\nu,T}(w_N)^2\right)^{1/2} &= \left\| \bar{a}^{-1/2} h_{\mathcal{T}} \sum_{m=1}^{\infty} \nabla\cdot\left(a_m(\beta^m_{\nu_m+1}\nabla w_{N,\nu+\epsilon_m} + \beta^m_{\nu_m}\nabla w_{N,\nu-\epsilon_m})\right)\right\|_{L^2(D)} \\ &\le \sum_{m=1}^{\infty} \left\| \frac{h_{\mathcal{T}}\nabla a_m}{\bar{a}}\right\|_{L^\infty(D)} \left(\beta^m_{\nu_m+1}\|w_{N,\nu+\epsilon_m}\|_V + \beta^m_{\nu_m}\|w_{N,\nu-\epsilon_m}\|_V\right) \\ &\quad + c_{\mathrm{inv}} \sum_{m=1}^{\infty} \left\| \frac{a_m}{\bar{a}}\right\|_{L^\infty(D)} \left(\beta^m_{\nu_m+1}\|w_{N,\nu+\epsilon_m}\|_V + \beta^m_{\nu_m}\|w_{N,\nu-\epsilon_m}\|_V\right).\end{aligned}$$

With (5.15), the last term is bounded by $(c_{a,\delta}(\Lambda^*,\mathcal{T}) + c_{\mathrm{inv}})\zeta_\nu(w_N)$. Similarly, the triangle inequality on the skeleton $\mathcal{S}$ of $\mathcal{T}$ leads to

$$\begin{aligned}\left(\sum_{T\in\mathcal{T}} \eta^1_{\nu,T}(w_N)^2\right)^{1/2} &\le \sum_{m=1}^{\infty} \left\| \frac{a_m}{\bar{a}}\right\|_{L^\infty(D)} \beta^m_{\nu_m+1} \left(\sum_{T\in\mathcal{T}} h_T \|\bar{a}^{1/2}[\![\nabla w_{N,\nu+\epsilon_m}]\!]\|^2_{L^2(\partial T\cap D)}\right)^{1/2} \\ &\quad + \sum_{m=1}^{\infty} \left\| \frac{a_m}{\bar{a}}\right\|_{L^\infty(D)} \beta^m_{\nu_m} \left(\sum_{T\in\mathcal{T}} h_T \|\bar{a}^{1/2}[\![\nabla w_{N,\nu-\epsilon_m}]\!]\|^2_{L^2(\partial T\cap D)}\right)^{1/2}\end{aligned}$$

and the inverse inequality $\sum_{T\in\mathcal{T}} h_T\|\bar{a}^{1/2}\nabla v_N\cdot n_T\|_{L^2(\partial T\cap D)} \le c^2_{\mathrm{inv}}\|v_N\|_V$ for $v_N\in V_p(\mathcal{T})$ implies

$$\left(\sum_{T\in\mathcal{T}} \eta^1_{\nu,T}(w_N)^2\right)^{1/2} \le 2c_{\mathrm{inv}}\zeta_\nu(w_N).$$

Combining these bounds, we have

$$\left(\sum_{T\in\mathcal{T}} \eta_{\nu,T}(w_N)^2\right)^{1/2} \le \left((c_{a,\delta}(\Lambda^*,\mathcal{T}) + c_{\mathrm{inv}})^2 + 4c^2_{\mathrm{inv}}\right)^{1/2}\zeta_\nu(w_N),$$

and the assertion follows by summing over $\nu\in\partial\Lambda\cap\Lambda^*$. □

A continuity property similar to that in Lemma 5.5 holds for the oscillation $\mathrm{osc}_T(w_N,\Lambda)$. The proof of the following lemma is analogous to the above argument (see also [2], Lem. 3.3).

Lemma 5.7. *For all $v_N, w_N\in\mathcal{V}_p(\Lambda,\mathcal{T})$ and all $T\in\mathcal{T}$,*

$$|\mathrm{osc}_T(v_N,\Lambda) - \mathrm{osc}_T(w_N,\Lambda)| \le \left(c_{a,\delta}(\Lambda,\mathcal{T}) + \hat{c}_{\mathrm{osc}}\right)(1+\gamma)|v_N - w_N|_{L^2_\pi(\Gamma;V|T)} \tag{5.18}$$

with a uniform constant $\hat{c}_{\mathrm{osc}}$ on $\mathbb{T}$.

6. The adaptive algorithm

6.1. Modules

Given a mesh $\mathcal{T}\in\mathbb{T}$ and a finite set $\Lambda\subset\mathcal{F}$ containing 0, we assume that a routine

$$u_N \leftarrow \texttt{Solve}[\Lambda,\mathcal{T}] \tag{6.1}$$

is available which returns the exact Galerkin projection u_N determined by (3.14) in the space $\mathcal{V}_p(\Lambda,\mathcal{T})$ from (3.13), for a fixed local polynomial degree p.

The error indicators from Sections 4.2 and 5.1 are computed by the modules

$$(\eta_T(u_N,\Lambda))_{T\in\mathcal{T}},\eta(u_N,\Lambda,\mathcal{T})\leftarrow \texttt{Estimate}_x[u_N,\Lambda,\mathcal{T}], \tag{6.2}$$

$$(\zeta_\nu(u_N))_{\nu\in\partial^\circ\Lambda},\zeta(u_N,\partial\Lambda),(\|u_N\|_V)_{\mu\in\Lambda}\leftarrow \texttt{Estimate}_y[u_N,\Lambda], \tag{6.3}$$

where (4.16) is used to compute $\zeta(u_N,\partial\Lambda)$ as a finite sum. These error indicators are subsequently used to mark cells of the spatial mesh $\mathcal{T}$ for refinement, and to activate indices in $\partial\Lambda$.

We consider separate marking and refinement procedures for $\mathcal{T}$ and Λ. For a parameter $0<\vartheta_x<1$, let the routine

$$\mathcal{M}\leftarrow \texttt{Mark}_x[\vartheta_x,(\eta_T(u_N,\Lambda))_{T\in\mathcal{T}},\eta(u_N,\Lambda,\mathcal{T})] \tag{6.4}$$

return a subset $\mathcal{M}\subset\mathcal{T}$ satisfying the Dörfler property

$$\eta(u_N,\Lambda,\mathcal{M})\geq\vartheta_x\eta(u_N,\Lambda,\mathcal{T}), \tag{6.5}$$

and let the module

$$\mathcal{T}^*\leftarrow \texttt{Refine}_x[\mathcal{T},\mathcal{M}] \tag{6.6}$$

construct a conforming mesh $\mathcal{T}^*\in\mathbb{T}$ in which at least all elements of $\mathcal{M}$ have been bisected at least once compared to $\mathcal{T}$. These methods are standard to adaptive finite element algorithms, and do not depend on $\Lambda\subset\mathcal{F}$.

A similar routine that constructs a finite set $\Delta\subset\partial\Lambda$ with

$$\zeta(u_N,\Delta)\geq\vartheta_y\zeta(u_N,\partial\Lambda) \tag{6.7}$$

for a parameter $0<\vartheta_y<1$ is discussed in the next section. Let

$$\Lambda^*\leftarrow \texttt{Refine}_y[\Lambda,\Delta] \tag{6.8}$$

return a set $\Lambda\cup\Delta\subset\Lambda^*\subset\Lambda\cup\partial\Lambda$. A simple choice is $\Lambda^*:=\Lambda\cup\Delta$, but we do not assume this particular definition, and indeed a larger set may be chosen in order to ensure favorable properties of Λ^*, such as monotonicity, if preserving such properties is deemed worth the potentially significant additional computational cost.

Finally, in order to control the constant $c_{a,\delta}(\Lambda,\mathcal{T})$ from (5.14), we select an arbitrary $\bar{c}_{a,\delta}>0$ and, for each $m\in\mathbb{N}$, presume that a mesh $\mathcal{T}_{a,m}\in\mathbb{T}$ is given such that $\|h_{\mathcal{T}_{a,m}}\nabla a_m/\bar{a}\|_{L^\infty(D)}\leq\bar{c}_{a,\delta}\|a_m/\bar{a}\|_{L^\infty(D)}$. Similarly, let $\mathcal{T}_{\bar{a}}\in\mathbb{T}$ such that $\|h_{\mathcal{T}_{\bar{a}}}\nabla\bar{a}/\bar{a}\|_{L^\infty(D)}\leq\bar{c}_{a,\delta}$. For any subset $S\subset\mathbb{N}$, let

$$\mathcal{T}_{a,S}:=\mathcal{T}_{\bar{a}}\oplus\bigoplus_{m\in S}\mathcal{T}_{a,m} \tag{6.9}$$

be the overlay of the meshes corresponding to $m\in S$. Then $c_{a,\delta}(\Lambda,\mathcal{T}_{a,\operatorname{supp}\Lambda})\leq\bar{c}_{a,\delta}$ for any finite $\Lambda\subset\mathcal{F}$.

6.2. Marking of parametric modes

A typical way to ensure the Dörfler property (6.7) while minimizing the size of Δ is to sort $\nu\in\partial\Lambda$ according to $\zeta_\nu(u_N)$ and construct Δ by successively selecting those ν with maximal $\zeta_\nu(u_N)$ until (6.7) is fulfilled. However, this is infeasible due to the infinite cardinality of $\partial\Lambda$.

The routine

$$\Delta\leftarrow \texttt{Mark}_y[\vartheta_y,(\zeta_\nu(u_N))_{\nu\in\partial^\circ\Lambda},\zeta(u_N,\partial\Lambda),(\|u_{N,\mu}\|_V)_{\mu\in\Lambda}] \tag{6.10}$$

functions by a slight extension of the above algorithm. Initially, only indices ν in the finite set $\partial^\circ\Lambda$ are considered for inclusion in Δ. Whenever an index of the form $\nu=\mu+\epsilon_m$ with $\mu\in\Lambda$ and $m=\max(\operatorname{supp}\Lambda)$ is added to Δ, the error indicator $\zeta_{\nu'}(u_N)=\|a_m/\bar{a}\|_{L^\infty(D)}\beta_1^m\|u_{N,\mu}\|_V$ for $\nu'=\mu+\epsilon_{m'}$ with $m'=\min(\mathbb{N}\setminus\operatorname{supp}\Lambda)$ is constructed and inserted into the sorted list of error indicators. Similarly, whenever such a ν' is added to Δ,

the index $\nu'' = \mu + \epsilon_{m''}$ is subsequently considered for the next larger m'' in $\mathbb{N} \setminus \operatorname{supp}(\Lambda)$. Thus, at every step, only a finite subset of $\partial\Lambda$ is considered for addition to Δ. The dynamic computation of $\zeta_\nu(u_N)$ for $\nu \in \partial\Lambda \setminus \partial^\circ\Lambda$ is inexpensive due to the simple structure (4.14). This process is continued until the Dörfler property (6.7) is satisfied.

Remark 6.1. If $\|a_m/\bar{a}\|_{L^\infty(D)}\beta_1^m$ are arranged in decreasing order and $\operatorname{supp}(\Lambda) = \{1, \dots, M\}$ for an $M \in \mathbb{N}$, then $\texttt{Mark}_y$ constructs a set Δ of minimal cardinality subject to the Dörfler property (6.7) since indices $\nu \in \partial\Lambda \setminus \partial^\circ\Lambda$ are considered in decreasing order of $\zeta_\nu(u_N)$, and these error indicators are bounded by $\zeta_\nu(u_N)$ with $\nu \in \partial^\circ\Lambda$. Furthermore, $\operatorname{supp}(\Lambda \cup \Delta) = \{1, \dots, M'\}$ for an $M' \in \mathbb{N}$, ensuring the optimality of a subsequent marking, after the refinement to $\Lambda^* := \Lambda \cup \Delta$, or after applying some other reasonable refinement strategy.

6.3. Adaptive algorithm

The above modules combine to form the adaptive stochastic Galerkin finite element algorithm `ASGFEM`. In each iteration, either a spatial refinement is performed or the set of active indices is enlarged, depending on which error indicator is larger.

The following statement is a direct consequence of Corollary 5.3 and the termination criterion of the algorithm.

$u_\epsilon \leftarrow \texttt{ASGFEM}[\epsilon, \Lambda_0, \mathcal{T}_0, \varrho, \vartheta_x, \vartheta_y]$

for $j = 0, 1, 2, \dots$ **do**
- $u_j \leftarrow \texttt{Solve}[\Lambda_j, \mathcal{T}_j]$
- $(\zeta_{j,\nu})_{\nu\in\partial^\circ\Lambda_j}, \zeta_j, (\|u_{j,\mu}\|_V)_{\mu\in\Lambda_j} \leftarrow \texttt{Estimate}_y[u_j, \Lambda_j]$
- $(\eta_{j,T})_{T\in\mathcal{T}_j}, \eta_j \leftarrow \texttt{Estimate}_x[u_j, \Lambda_j, \mathcal{T}_j]$
- **if** $\eta_j^2 + \zeta_j^2 \le \epsilon^2$ **then**
  - **return** $u_\epsilon \leftarrow u_j$
- **if** $\eta_j \ge \varrho\zeta_j$ **then**
  - $\Lambda_{j+1} \leftarrow \Lambda_j$
  - $\mathcal{M}_{j+1} \leftarrow \texttt{Mark}_x[\vartheta_x, (\eta_{j,T})_{T\in\mathcal{T}_j}, \eta_j]$
  - $\mathcal{T}_{j+1} \leftarrow \texttt{Refine}_x[\mathcal{T}_j, \mathcal{M}_{j+1}]$
- **else**
  - $\Delta_j \leftarrow \texttt{Mark}_y[\vartheta_y, (\zeta_{j,\nu})_{\nu\in\partial\Lambda_j}, \zeta_j, (\|u_{j,\mu}\|_V)_{\mu\in\Lambda_j}]$
  - $\Lambda_{j+1} \leftarrow \texttt{Refine}_y[\Lambda_j, \Delta_j]$
  - $\mathcal{T}_{j+1} \leftarrow \mathcal{T}_j \oplus \mathcal{T}_{a,\operatorname{supp}\Lambda_{j+1}}$

Theorem 6.2. *Let $\epsilon > 0$, $\Lambda_0 \subset \mathcal{F}$ be finite and contain 0, $\mathcal{T}_0 \in \mathbb{T}$ with $\mathcal{T}_{a,\operatorname{supp}\Lambda_0} \preceq \mathcal{T}_0$, $\varrho > 0$ and $0 < \vartheta_x, \vartheta_y < 1$. If `ASGFEM`$[\epsilon, \Lambda_0, \mathcal{T}_0, \varrho, \vartheta_x, \vartheta_y]$ terminates, it returns an approximate solution u_ϵ with*

$$\|u_\epsilon - u\|_{\mathcal{A}}^2 \le \frac{C_\eta}{1-\gamma}\epsilon^2. \tag{6.11}$$

We tacitly assume that the assumptions of Theorem 6.2 hold in the following. In particular, $\Lambda_0 \subset \mathcal{F}$ is any finite set containing 0, and $\mathcal{T}_0 \in \mathbb{T}$ is adapted to $\bar{a}$ in the sense that $\mathcal{T}_{a,\operatorname{supp}\Lambda_0} \preceq \mathcal{T}_0$.

7. Contraction property

7.1. A preliminary estimate

Our analysis is adapted from [2]. The following statement is an analogue to ([2], Cor. 3.4).

Lemma 7.1. *For any nonempty finite sets $\Lambda \subset \Lambda^* \subset \mathcal{F}$ and any meshes $\mathcal{T} \preceq \mathcal{T}^* \in \mathbb{T}$, let $\mathcal{M} := \mathcal{T} \setminus (\mathcal{T}^* \cap \mathcal{T})$ denote the set of refined cells in $\mathcal{T}^*$ compared to $\mathcal{T}$, and let $\Delta := \partial\Lambda \cap \Lambda^*$. For any $v_N \in \mathcal{V}_p(\Lambda, \mathcal{T})$, $v_N^* \in \mathcal{V}_p(\Lambda^*, \mathcal{T}^*)$, $\chi, \tau > 0$ and $\kappa \geq 0$,*

$$\begin{aligned}\eta(v_N^*, \Lambda^*, \mathcal{T}^*)^2 + \kappa\zeta(v_N^*, \partial\Lambda^*)^2 \leq (1+\chi)\big[\eta(v_N, \Lambda, \mathcal{T})^2 - \lambda\eta(v_N, \Lambda, \mathcal{M})^2\big] \\ + (1+\tau)\kappa\zeta(v_N, \partial\Lambda)^2 - \big[(1+\tau)\kappa - \bar{c}_\zeta^2(1+\chi)\big]\zeta(v_N, \Delta)^2 \\ + [(1+\chi^{-1})\bar{c}_\eta^2 + (1+\tau^{-1})\kappa\gamma^2](1-\gamma)^{-1}\|v_N - v_N^*\|_{\mathcal{A}}^2\end{aligned}$$

with $\lambda = 1 - 2^{1/d}$, $\bar{c}_\zeta := 2c_{a,\delta}(\Lambda^*, \mathcal{T}^*) + \hat{c}_{\eta,\zeta}$ and $\bar{c}_\eta := [c_{a,\delta}(\Lambda^*, \mathcal{T}^*) + \hat{c}_\eta](1+\gamma)$.

Proof. Let $v_N \in \mathcal{V}_p(\Lambda, \mathcal{T})$ and $v_N^* \in \mathcal{V}_p(\Lambda^*, \mathcal{T}^*)$. Since $\mathcal{V}_p(\Lambda, \mathcal{T}) \subset \mathcal{V}_p(\Lambda^*, \mathcal{T}^*)$, Lemma 5.5 together with Young's inequality imply

$$\begin{aligned}\eta(v_N^*, \Lambda^*, \mathcal{T}^*)^2 &\leq \sum_{T^* \in \mathcal{T}^*} \big[\eta_{T^*}(v_N, \Lambda^*) + \bar{c}_\eta |v_N - v_N^*|_{L^2_\pi(\Gamma; V|T^*)}\big]^2 \\ &\leq (1+\chi)\eta(v_N, \Lambda^*, \mathcal{T}^*)^2 + (1+\chi^{-1})\bar{c}_\eta^2\|v_N - v_N^*\|^2_{L^2_\pi(\Gamma;V)}\end{aligned}$$

with $\bar{c}_\eta := [c_{a,\delta}(\Lambda^*, \mathcal{T}^*) + \hat{c}_\eta](1+\gamma)$. Due to Lemma 5.6, for $\bar{c}_\zeta := 2c_{a,\delta}(\Lambda^*, \mathcal{T}^*) + \hat{c}_{\eta,\zeta}$,

$$\eta(v_N, \Lambda^*, \mathcal{T}^*)^2 \leq \eta(v_N, \Lambda, \mathcal{T}^*)^2 + \bar{c}_\zeta^2\zeta(v_N, \Delta)^2.$$

Let $T \in \mathcal{M} \subset \mathcal{T}$ and let $\mathcal{T}^*(T) := \{T^* \in \mathcal{T}^*;\, T^* \subset T\}$. For any $\mu \in \Lambda$, $[\![\sigma_\mu(v_N)]\!] = 0$ on all facets of $\mathcal{T}^*$ in the interior of T since v_N is continuous on T. Furthermore, $h_{T^*} = |T^*|^{1/d} \leq (|T|/2)^{1/d} = 2^{-1/d}h_T$ for all $T^* \in \mathcal{T}^*(T)$. Thus

$$\begin{aligned}\eta(v_N, \Lambda, \mathcal{T}^*)^2 &\leq \eta(v_N, \Lambda, \mathcal{T} \setminus \mathcal{M})^2 + 2^{-1/d}\eta(v_N, \Lambda, \mathcal{M})^2 \\ &= \eta(v_N, \Lambda, \mathcal{T})^2 - \lambda\eta(v_N, \Lambda, \mathcal{M})^2\end{aligned}$$

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

Similarly, Lemma 4.5 and Young's inequality imply

$$\begin{aligned}\zeta(v_N^*, \Lambda^*)^2 &\leq \big(\zeta(v_N, \partial\Lambda^*) + \gamma\|v_N - v_N^*\|_{L^2_\pi(\Gamma;V)}\big)^2 \\ &\leq (1+\tau)\zeta(v_N, \partial\Lambda^*)^2 + (1+\tau^{-1})\gamma^2\|v_N - v_N^*\|^2_{L^2_\pi(\Gamma;V)}.\end{aligned}$$

Since $\zeta_\nu(v_N) = 0$ for $\nu \in \partial\Lambda^* \setminus \partial\Lambda$ and $\Delta = \partial\Lambda \cap \Lambda^* = \partial\Lambda \setminus \partial\Lambda^*$,

$$\zeta(v_N, \partial\Lambda^*)^2 = \zeta(v_N, \partial\Lambda)^2 - \zeta(v_N, \partial\Lambda \setminus \partial\Lambda^*)^2 = \zeta(v_N, \partial\Lambda)^2 - \zeta(v_N, \Delta)^2.$$

The assertion follows with $\|v_N - v_N^*\|^2_{L^2_\pi(\Gamma;V)} \leq (1-\gamma)^{-1}\|v_N - v_N^*\|_{\mathcal{A}}^2$. □

7.2. Convergence of the adaptive algorithm

We show in Theorem 7.2 that for certain $\omega_\eta, \omega_\zeta > 0$, the adaptive algorithm `ASGFEM` is a contraction for the quasi-error

$$\|u_N - u\|_{\mathcal{A}}^2 + \omega_\eta\eta(u_N, \Lambda, \mathcal{T})^2 + \omega_\zeta\zeta(u_N, \partial\Lambda)^2. \tag{7.1}$$

As is evident from the proof, it is vital that ω_η and ω_ζ may be distinct constants; indeed, ω_ζ may be larger than ω_η by a factor depending on $\bar{c}_{a,\delta}$.

Theorem 7.2. *Let $\varrho > 0$ and $0 < \vartheta_x, \vartheta_y < 1$, and let u_j, $\mathcal{T}_j$, $\mathcal{M}_j$, Δ_j, η_j and ζ_j denote the sequences of approximate solutions, finite element meshes, marked cells, marked indices and error indicators, respectively, generated in `ASGFEM`. There exist constants $0 < \delta < 1$, $\omega_\eta > 0$ and $\omega_\zeta > 0$ such that*

$$\|u_{j+1} - u\|_{\mathcal{A}}^2 + \omega_\eta\eta_{j+1}^2 + \omega_\zeta\zeta_{j+1}^2 \leq \delta\big(\|u_j - u\|_{\mathcal{A}}^2 + \omega_\eta\eta_j^2 + \omega_\zeta\zeta_j^2\big) \tag{7.2}$$

for all $j \in \mathbb{N}_0$.

Proof. We abbreviate $e_j := \|u_j - u\|_{\mathcal{A}}$ and $d_j := \|u_j - u_{j+1}\|_{\mathcal{A}}$. Lemma 7.1 implies

$$\begin{aligned}\eta_{j+1}^2 + \kappa\zeta_{j+1}^2 \leq{}& (1+\chi)[\eta_j^2 - \lambda\eta(u_j, \Lambda_j, \mathcal{M}_j)^2] \\ &+ (1+\tau)\kappa\zeta_j^2 - [(1+\tau) - (1+\chi)\bar{c}_\zeta^2\kappa^{-1}]\kappa\zeta(u_j, \Delta_j)^2 \\ &+ [(1+\chi^{-1})\bar{c}_\eta^2 + (1+\tau^{-1})\kappa\gamma^2](1-\gamma)^{-1}d_j^2\end{aligned}$$

with $\lambda = 1 - 2^{1/d}$, $\bar{c}_\zeta := 2\bar{c}_{a,\delta} + \hat{c}_{\eta,\zeta}$ and $\bar{c}_\eta := (\bar{c}_{a,\delta} + \hat{c}_\eta)(1+\gamma)$ provided that $(1+\tau) \geq (1+\chi)\bar{c}_\zeta^2\kappa^{-1}$. Using Galerkin orthogonality to expand $e_{j+1}^2 = e_j^2 - d_j^2$ leads to

$$\begin{aligned}e_{j+1}^2 + \omega(\eta_{j+1}^2 + \kappa\zeta_{j+1}^2) \leq{}& e_j^2 - \big[1 - \omega\big((1+\chi^{-1})\bar{c}_\eta^2 + (1+\tau^{-1})\kappa\gamma^2\big)(1-\gamma)^{-1}\big]d_j^2 \\ &+ \omega(1+\chi)[\eta_j^2 - \lambda\eta(u_j, \Lambda_j, \mathcal{M}_j)^2] \\ &+ \omega(1+\tau)\kappa\zeta_j^2 - \omega[(1+\tau) - (1+\chi)\bar{c}_\zeta^2\kappa^{-1}]\kappa\zeta(u_j, \Delta_j)^2.\end{aligned}$$

We set $\omega := \omega(\chi, \tau, \kappa) := (1-\gamma)/[(1+\chi^{-1})\bar{c}_\eta^2 + (1+\tau^{-1})\kappa\gamma^2]$ such that the term containing d_j drops from this estimate. We expand $e_j^2 = (1-\alpha)e_j^2 + \alpha e_j^2$ with $0 < \alpha < 1$ and apply the upper bound (5.12) to αe_j^2 to get

$$\begin{aligned}e_{j+1}^2 + \omega(\eta_{j+1}^2 + \kappa\zeta_{j+1}^2) \leq{}& (1-\alpha)e_j^2 + \alpha C_\eta(1-\gamma)^{-1}(\eta_j^2 + \zeta_j^2) \\ &+ \omega(1+\chi)[\eta_j^2 - \lambda\eta(u_j, \Lambda_j, \mathcal{M}_j)^2] \\ &+ \omega(1+\tau)\kappa\zeta_j^2 - \omega[(1+\tau) - (1+\chi)\bar{c}_\zeta^2\kappa^{-1}]\kappa\zeta(u_j, \Delta_j)^2.\end{aligned}$$

If $\eta_j \geq \varrho\zeta_j$, then $\Delta_j = \emptyset$, thus $\zeta(u_j, \Delta_j) = 0$, and by the Dörfler property (6.5), using $(1+\beta_x)\tau\kappa\zeta_j^2 \leq (1+\beta_x)\tau\kappa\varrho^{-2}\eta_j^2$ for any $\beta_x > 0$,

$$\begin{aligned}e_{j+1}^2 + \omega(\eta_{j+1}^2 + \kappa\zeta_{j+1}^2) \leq{}& (1-\alpha)e_j^2 \\ &+ \omega\big[(1+\chi)(1-\lambda\vartheta_x^2) + (1+\beta_x)\tau\kappa\varrho^{-2} + \alpha C_\eta(1-\gamma)^{-1}\omega^{-1}\big]\eta_j^2 \\ &+ \omega(1 - \beta_x\tau + \alpha C_\eta(1-\gamma)^{-1}\omega^{-1}\kappa^{-1})\kappa\zeta_j^2.\end{aligned}$$

Conversely, if $\eta_j < \varrho\zeta_j$, then $\mathcal{M}_j = \emptyset$ and consequently $\eta(u_j, \Lambda_j, \mathcal{M}_j) = 0$. The Dörfler property (6.7) along with $(1+\beta_y)\chi\eta_j^2 \leq (1+\beta_y)\chi\varrho^2\zeta_j^2$ for $\beta_y > 0$ imply

$$\begin{aligned}e_{j+1}^2 + \omega(\eta_{j+1}^2 + \kappa\zeta_{j+1}^2) \leq{}& (1-\alpha)e_j^2 + \omega(1 - \beta_y\chi + \alpha C_\eta(1-\gamma)^{-1}\omega^{-1})\eta_j^2 \\ &+ \omega\kappa\big[(1+\tau) - \vartheta_y^2\big((1+\tau) - (1+\chi)\bar{c}_\zeta^2\kappa^{-1}\big) + (1+\beta_y)\chi\varrho^2\kappa^{-1} \\ &+ \alpha C_\eta(1-\gamma)^{-1}\omega^{-1}\kappa^{-1}\big]\zeta_j^2.\end{aligned}$$

All of the factors in the above estimates must be made less than one while ensuring $(1+\tau) \geq (1+\chi)\bar{c}_\zeta^2\kappa^{-1}$. We select $\kappa > \bar{c}_\zeta^2$ and

$$0 < \tau < \min\big(\vartheta_y^2(1 - \bar{c}_\zeta^2\kappa^{-1})(1-\vartheta_y^2)^{-1}, \lambda\vartheta_x^2\varrho^2\kappa^{-1}\big)$$

such that $1+\tau-\vartheta_y^2(1+\tau-\bar{c}_\zeta^2\kappa^{-1}) < 1$ and $1-\lambda\vartheta_x^2+\tau\kappa\varrho^{-2} < 1$. Next, we choose $\chi > 0$ sufficiently small such that $\chi \leq (1+\tau)\kappa\bar{c}_\zeta^{-2} - 1$, which implies $(1+\tau) \geq (1+\chi)\bar{c}_\zeta^2\kappa^{-1}$, simultaneously with $1+\tau-\vartheta_y^2((1+\tau)-(1+\chi)\bar{c}_\zeta^2\kappa^{-1})+\chi\varrho^2\kappa^{-1} < 1$ and $(1+\chi)(1-\lambda\vartheta_x^2)+\tau\kappa\varrho^{-2} < 1$. This permits $\beta_x > 0$ with $(1+\chi)(1-\lambda\vartheta_x^2)+(1+\beta_x)\tau\kappa\varrho^{-2} < 1$ and $\beta_y > 0$ with $1+\tau-\vartheta_y^2((1+\tau)-(1+\chi)\bar{c}_\zeta^2\kappa^{-1})+(1+\beta_y)\chi\varrho^2\kappa^{-1} < 1$. Finally, we choose $\alpha > 0$ sufficiently small such that all the factors in the above estimates remain smaller than one. The assertion follows with δ equal to the maximum of these factors, $\omega_\eta := \omega$ and $\omega_\zeta := \kappa\omega$. □

7.3. Contraction of the spatial error

Theorem 7.2 achieves a contraction of the quasi-error (7.1) by balancing a potential increase in one error indicator with a decrease in the other. If the adaptive algorithm ASGFEM performs only spatial refinements within a succession of iterations, and the set Λ of active indices in $\mathcal{F}$ therefore remains fixed, then a similar contraction property holds for just the spatial error, with constants independent of Λ. This is elaborated in following theorem, which follows ([2], Thm. 4.1).

Theorem 7.3. *Let $\varrho > 0$ and $0 < \vartheta_x < 1$, and let u_j, $\mathcal{T}_j$, $\mathcal{M}_j$, Λ_j and η_j denote the sequences of approximate solutions, finite element meshes, marked cells, active indices and error indicators, respectively, generated in ASGFEM. There exist constants $0 < \delta_x < 1$ and $\omega_x > 0$ such that for any $j \in \mathbb{N}_0$ with $\Lambda_{j+1} = \Lambda_j =: \Lambda$,*

$$\|u_{j+1} - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_{j+1}^2 \le \delta_x \big(\|u_j - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_j^2\big). \tag{7.3}$$

Proof. We abbreviate $e_j := \|u_j - u_\Lambda\|_{\mathcal{A}}$ and $d_j := \|u_j - u_{j+1}\|_{\mathcal{A}}$. Lemma 7.1 with $\kappa = 0$ and $\Delta = \emptyset$ implies

$$\eta_{j+1}^2 \le (1+\chi)[\eta_j^2 - \lambda \eta(u_j, \Lambda_j, \mathcal{M}_j)^2] + (1+\chi^{-1})\bar{c}_\eta^2 (1-\gamma)^{-1} d_j^2,$$

with $\bar{c}_\zeta := 2\bar{c}_{a,\delta} + \hat{c}_{\eta,\zeta}$ for any $\chi > 0$. Since $e_{j+1}^2 = e_j^2 - d_j^2$ by Galerkin orthogonality, and using the Dörfler property (6.5), we have

$$e_{j+1}^2 + \omega_x \eta_{j+1}^2 \le e_j^2 - [1 - \omega_x(1+\chi^{-1})\bar{c}_\eta^2(1-\gamma)^{-1}] d_j^2 + \omega_x(1+\chi)(1-\lambda\vartheta_x^2)\eta_j^2$$

for any $\omega_x > 0$. We choose $\omega_x := (1-\gamma)/[(1+\chi^{-1})\bar{c}_\eta^2]$, depending on χ, such that the term involving d_j drops. Expanding e_j^2 as $(1-\alpha)e_j^2 + \alpha e_j^2$ with $0 < \alpha < 1$ and applying Corollary 5.2 to αe_j^2 leads to

$$e_{j+1}^2 + \omega_x \eta_{j+1}^2 \le (1-\alpha) e_j^2 + \omega_x [C_1(\chi) + C_2(\chi, \alpha)] \eta_j^2$$

with $C_1(\chi) = (1+\chi)(1-\lambda\vartheta_x^2)$ and $C_2(\chi,\alpha) = \alpha(1+\chi^{-1})C_\eta \bar{c}_\eta^2 (1-\gamma)^{-2}$. Estimate (7.3) follows with $\delta_x = \max(1-\alpha, C_1(\chi) + C_2(\chi,\alpha)) < 1$ by selecting $\chi > 0$ sufficiently small such that $C_1(\chi) < 1$, and then choosing $\alpha > 0$ sufficiently small such that $C_2(\chi,\alpha) < 1 - C_1(\chi)$. □

8. Quasi-optimality of the spatial discretization

8.1. The total spatial error

Let $w_N \in \mathcal{V}_p(\Lambda, \mathcal{T})$ be any approximation of u for a finite set $\Lambda \in \mathcal{F}$ and a mesh $\mathcal{T} \in \mathbb{T}$. The total spatial error

$$\left(\|w_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(w_N, \Lambda, \mathcal{T})^2 \right)^{1/2} \tag{8.1}$$

combines the energy-norm error with the oscillation. Due to Corollary 5.2 and (5.8), for the Galerkin projection $u_N \in \mathcal{V}_p(\Lambda, \mathcal{T})$,

$$\begin{aligned} \frac{c_\eta}{1+\gamma} \eta(u_N, \Lambda, \mathcal{T})^2 &\le \|u_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(u_N, \Lambda, \mathcal{T})^2 \\ &\le \left(\frac{c_\eta}{1+\gamma} + \frac{C_\eta}{1-\gamma} \right) \eta(u_N, \Lambda, \mathcal{T})^2, \end{aligned} \tag{8.2}$$

i.e. the total spatial error is equivalent to the spatial error indicator. Furthermore, u_N is a quasi-optimal approximation of u_Λ in $\mathcal{V}_p(\Lambda, \mathcal{T})$ with respect to the total spatial error.

Lemma 8.1. *If $c_{a,\delta}(\Lambda,\mathcal{T}) \leq \bar{c}_{a,\delta}$, then the Galerkin projection $u_N \in \mathcal{V}_p(\Lambda,\mathcal{T})$ satisfies*

$$\|u_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2 \leq \hat{C} \inf_{w_N \in \mathcal{V}_p(\Lambda,\mathcal{T})} \left(\|w_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N,\Lambda,\mathcal{T})^2 \right) \tag{8.3}$$

with a constant $\hat{C} := 2\max(1, c_\eta(\bar{c}_{a,\delta} + \hat{c}_{\mathrm{osc}})^2(1+\gamma)(1-\gamma)^{-1})$ independent of $\mathcal{T}$ and Λ.

Proof. Let $w_N \in \mathcal{V}_p(\Lambda,\mathcal{T})$. Due to Lemma 5.7,

$$\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2 \leq 2\operatorname{osc}(w_N,\Lambda,\mathcal{T})^2 + \frac{2(\bar{c}_{a,\delta}+\hat{c}_{\mathrm{osc}})^2(1+\gamma)^2}{1-\gamma}\|w_N - u_N\|_{\mathcal{A}}^2.$$

By Galerkin orthogonality, $\|w_N - u_N\|_{\mathcal{A}}^2 \leq \|w_N - u_\Lambda\|_{\mathcal{A}}^2$ and $\|u_N - u_\Lambda\|_{\mathcal{A}}^2 \leq \|w_N - u_\Lambda\|_{\mathcal{A}}^2$. Consequently,

$$\|u_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2 \leq \hat{C}\left(\|w_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N,\Lambda,\mathcal{T})^2 \right)$$

with $\hat{C}$ as in the statement of the lemma, and the assertion follows by taking the infimum over $w_N \in \mathcal{V}_p(\Lambda,\mathcal{T})$. □

Similar to ([2], Lem. 5.9), there is an intimate connection between a reduction of the total spatial error and the Dörfler property (6.5).

Lemma 8.2. *Let u_N, u_N^* denote the Galerkin solutions in $\mathcal{V}_p(\Lambda,\mathcal{T})$ and $\mathcal{V}_p(\Lambda,\mathcal{T}^*)$, respectively, for meshes $\mathcal{T},\mathcal{T}^*$ with $\mathcal{T} \preceq \mathcal{T}^*$ and $c_{a,\delta}(\Lambda,\mathcal{T}^*) \leq \bar{c}_{a,\delta}$, and let*

$$\|u_N^* - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N^*,\Lambda,\mathcal{T}^*)^2 \leq c_{\mathrm{red}}\left(\|u_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2 \right) \tag{8.4}$$

with $c_{\mathrm{red}} < 1/2$. Then

$$\eta(u_N,\Lambda,\mathcal{M}) \geq \vartheta_x \eta(u_N,\Lambda,\mathcal{T}) \tag{8.5}$$

for the set $\mathcal{M} := \mathcal{T} \setminus (\mathcal{T}^* \cap \mathcal{T})$ of refined cells and $\vartheta_x^2 = (1-2c_{\mathrm{red}})\hat{\vartheta}_x^2$, where

$$\hat{\vartheta}_x := \left(1 + \bar{C}_\eta\left(\frac{1+\gamma}{c_\eta} + 2(\bar{c}_{a,\delta}+\hat{c}_{\mathrm{osc}})\frac{1+\gamma}{1-\gamma}\right)\right)^{-1/2}. \tag{8.6}$$

Proof. Due to the lower bound in Corollary 5.2,

$$\frac{c_\eta}{1+\gamma}\eta(u_N,\Lambda,\mathcal{T})^2 \leq \|u_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2.$$

Inserting the estimate (8.4), we have

$$\begin{aligned}(1-2c_{\mathrm{red}})\frac{c_\eta}{1+\gamma}\eta(u_N,\Lambda,\mathcal{T})^2 \leq\ & \|u_N - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N,\Lambda,\mathcal{T})^2 \\ & - 2\|u_N^* - u_\Lambda\|_{\mathcal{A}}^2 - 2\frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N^*,\Lambda,\mathcal{T}^*)^2.\end{aligned}$$

By Galerkin orthogonality and Lemma 5.4,

$$\|u_N - u_\Lambda\|_{\mathcal{A}}^2 - 2\|u_N^* - u_\Lambda\|_{\mathcal{A}}^2 \leq \|u_N - u_N^*\|_{\mathcal{A}}^2 \leq \bar{C}_\eta \eta(u_N,\Lambda,\mathcal{M})^2.$$

Furthermore, since $\operatorname{osc}_T(u_N,\Lambda) \leq \eta_T(u_N,\Lambda)$ for all $T \in \mathcal{M}$ by (5.8) and

$$\operatorname{osc}_T(u_N,\Lambda)^2 \leq 2\operatorname{osc}_T(u_N^*,\Lambda)^2 + 2(\bar{c}_{a,\delta}+\hat{c}_{\mathrm{osc}})(1+\gamma)|u_N - u_N^*|_{L^2_\pi(\Gamma;V|T)}$$

by Lemma 5.7 for $T \in \mathcal{T} \setminus \mathcal{M}$, employing the local upper bound Lemma 5.4 again, we have

$$\begin{aligned}\operatorname{osc}(u_N, \Lambda, \mathcal{T})^2 - 2\operatorname{osc}(u_N^*, \Lambda, \mathcal{T}^*)^2 &\leq \eta(u_N, \Lambda, \mathcal{M})^2 + 2(\bar{c}_{a,\delta} + \hat{c}_{\mathrm{osc}})\frac{1+\gamma}{1-\gamma}\|u_N - u_N^*\|_{\mathcal{A}}^2 \\ &\leq \Big(1 + 2\bar{C}_\eta(\bar{c}_{a,\delta} + \hat{c}_{\mathrm{osc}})\frac{1+\gamma}{1-\gamma}\Big)\eta(u_N, \Lambda, \mathcal{M})^2.\end{aligned}$$

Thus

$$(1 - 2c_{\mathrm{red}})\frac{c_\eta}{1+\gamma}\eta(u_N, \Lambda, \mathcal{T})^2 \leq \left(\bar{C}_\eta + \frac{c_\eta}{1+\gamma}\Big(1 + 2\bar{C}_\eta(\bar{c}_{a,\delta} + \hat{c}_{\mathrm{osc}})\frac{1+\gamma}{1-\gamma}\Big)\right)\eta(u_N, \Lambda, \mathcal{M})^2,$$

which is (8.5). □

8.2. An approximation class

For any finite set $\Lambda \subset \mathcal{F}$ and any $N \in \mathbb{N}$, let

$$\Sigma_N(u, \Lambda) := \inf\Big(\|w_N^* - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N^*, \Lambda, \mathcal{T}^*)^2\Big)^{1/2} \tag{8.7}$$

where the infimum is taken over all meshes $\mathcal{T}^* \in \mathbb{T}$ with $\#\mathcal{T}^* - \#\mathcal{T}_{\mathrm{init}} \leq N$ and $c_{a,\delta}(\Lambda, \mathcal{T}^*) \leq \bar{c}_{a,\delta}$, and all $w_N^* \in \mathcal{V}_p(\Lambda, \mathcal{T}^*)$. Furthermore, for any $s > 0$, let

$$|u|_{s,\Lambda} := \sup\big\{\epsilon\big(\min\{N \in \mathbb{N}_0\,;\ \Sigma_N(u, \Lambda) < \epsilon\}\big)^s\,;\ \epsilon \geq \check{c}\|u_\Lambda - u\|_{\mathcal{A}}\big\} \tag{8.8}$$

for a constant $\check{c} > 0$ specified in (8.14) below. We consider u to be in the approximation class $\mathbb{A}_s$ if

$$|u|_{\mathbb{A}_s} := \sup\{|u|_{s,\Lambda}\,;\ \Lambda \subset \mathcal{F} \text{ finite}, 0 \in \Lambda\} < \infty. \tag{8.9}$$

In this case, for any finite set $\Lambda \subset \mathcal{F}$ containing 0 and any error tolerance $\epsilon \geq \check{c}\|u_\Lambda - u\|_{\mathcal{A}}$, *i.e.* no smaller than the error effected by the restriction to the set Λ, up to a constant factor, there is an approximation $w_N^* \in \mathcal{V}_p(\Lambda, \mathcal{T}^*)$ with total spatial error

$$\|w_N^* - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N^*, \Lambda, \mathcal{T}^*)^2 \leq \epsilon^2 \tag{8.10}$$

for a mesh $\mathcal{T}^* \in \mathbb{T}$ of size

$$\#\mathcal{T}^* - \#\mathcal{T}_{\mathrm{init}} \leq \epsilon^{-1/s}|u|_{\mathbb{A}_s}^{1/s} \tag{8.11}$$

satisfying $c_{a,\delta}(\Lambda, \mathcal{T}^*) \leq \bar{c}_{a,\delta}$, *i.e.* the total spatial error decays as

$$\Big(\|w_N^* - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N^*, \Lambda, \mathcal{T}^*)^2\Big)^{1/2} \leq |u|_{\mathbb{A}_s}(\#\mathcal{T}^* - \#\mathcal{T}_{\mathrm{init}})^{-s}. \tag{8.12}$$

The full error of this approximation is bounded by $\|w_N^* - u\|_{\mathcal{A}} \leq (1 + \check{c}^{-2})^{1/2}\epsilon$ and decays at the same rate s with respect to the size of the mesh $\mathcal{T}^*$ as Λ is suitably enlarged to maintain $\|u_\Lambda - u\|_{\mathcal{A}} \leq \check{c}^{-1}\epsilon$.

8.3. Quasi-optimal convergence

We make the following assumptions:

(1) The routine $\mathcal{M} \leftarrow \mathtt{Mark}_x[\vartheta_x, (\eta_T(u_N, \Lambda))_{T\in\mathcal{T}}, \eta(u_N, \Lambda, \mathcal{T})]$ constructs a set $\mathcal{M} \subset \mathcal{T}$ of *minimal cardinality* satisfying the Dörfler property (6.5).
(2) The Dörfler constant ϑ_x from (6.5) satisfies $0 < \vartheta_x < \hat{\vartheta}_x$ for $\hat{\vartheta}_x$ from (8.6).
(3) The distribution of refinement facets in $\mathcal{T}_{\mathrm{init}}$ satisfies (b) of ([20], Sect. 4).

Lemma 8.2 and the assumed optimal marking lead to a bound on the cardinality of the sets $\mathcal{M}_j$ of marked cells in ASGFEM, following ([2], Lem. 5.10). We abbreviate

$$c_{\mathrm{red}} := \frac{1}{2}\left(1 - \frac{\vartheta_x^2}{\hat{\vartheta}_x^2}\right) > 0 \tag{8.13}$$

and specify the constant $\check{c}$ left arbitrary in Section 8.2 as

$$\check{c} := \left(\frac{c_{\mathrm{red}} c_\eta (1-\gamma)}{(1+\varrho^{-2})\hat{C} C_\eta (1+\gamma)}\right)^{1/2}. \tag{8.14}$$

Lemma 8.3. *If $u \in \mathbb{A}_s$, then*

$$\#\mathcal{M}_j \le |u|_{\mathbb{A}_s}^{1/s} c_{\mathrm{red}}^{-1/2s} \hat{C}^{1/2s} \Big(\|u_j - u_{\Lambda_j}\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_j, \Lambda_j, \mathcal{T}_j)^2\Big)^{-1/2s} \tag{8.15}$$

for all $j \in \mathbb{N}_0$ with $\eta_j \ge \varrho\zeta_j$.

Proof. Let $j \in \mathbb{N}_0$ with $\eta_j \ge \varrho\zeta_j$, such that a spatial refinement is performed and thus $\mathcal{M}_j$ is defined in ASGFEM. Let $\epsilon^2 = c_{\mathrm{red}}\hat{C}^{-1}[\|u_j - u_{\Lambda_j}\|_{\mathcal{A}}^2 + c_\eta(1+\gamma)^{-1}\operatorname{osc}(u_j, \Lambda_j, \mathcal{T}_j)^2]$, which satisfies

$$\begin{aligned}\epsilon^2 &\ge \frac{c_{\mathrm{red}} c_\eta}{\hat{C}(1+\gamma)}\eta_j^2 \ge \frac{c_{\mathrm{red}} c_\eta}{\hat{C}(1+\gamma)(1+\varrho^{-2})}(\eta_j^2 + \zeta_j^2)\\ &\ge \frac{c_{\mathrm{red}} c_\eta (1-\gamma)}{\hat{C}(1+\gamma)(1+\varrho^{-2})C_\eta}\|u_j - u\|_{\mathcal{A}}^2 \ge \check{c}^2\|u_{\Lambda_j} - u\|_{\mathcal{A}}^2\end{aligned}$$

due to (8.2), (5.12) and Galerkin orthogonality. Thus the assumption $u \in \mathbb{A}_s$ implies that there exist $\mathcal{T}^\epsilon \in \mathbb{T}$ and $w_N^\epsilon \in \mathcal{V}_p(\Lambda_j, \mathcal{T}^\epsilon)$ such that $c_{a,\delta}(\Lambda_j, \mathcal{T}^\epsilon) \le \bar{c}_{a,\delta}$, $\#\mathcal{T}^\epsilon - \#\mathcal{T}_{\mathrm{init}} \le \epsilon^{-1/s}|u|_{\mathbb{A}_s}^{1/s}$ and

$$\|w_N^\epsilon - u_{\Lambda_j}\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N^\epsilon, \Lambda_j, \mathcal{T}^\epsilon)^2 \le \epsilon^2.$$

Let u_N^* be the Galerkin solution in $\mathcal{V}_p(\Lambda_j, \mathcal{T}^*)$ for the overlay $\mathcal{T}^* := \mathcal{T}^\epsilon \oplus \mathcal{T}_j$. Since $\mathcal{T}^\epsilon \preceq \mathcal{T}^*$, Lemma 8.1 implies

$$\begin{aligned}\|u_N^* - u_{\Lambda_j}\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_N^*, \Lambda_j, \mathcal{T}^*)^2 &\le \hat{C}\Big(\|w_N^\epsilon - u_{\Lambda_j}\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(w_N^\epsilon, \Lambda_j, \mathcal{T}^*)^2\Big)\\ &\le \hat{C}\epsilon^2 = c_{\mathrm{red}}\Big(\|u_j - u_{\Lambda_j}\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_j, \Lambda_j, \mathcal{T}_j)^2\Big),\end{aligned}$$

where we used the monotonicity of the oscillation with respect to the mesh $\mathcal{T} \in \mathbb{T}$ in the second estimate. Consequently, Lemma 8.2 implies that the set $\mathcal{M}^* := \mathcal{T}\setminus(\mathcal{T}^* \cap \mathcal{T})$ satisfies the Dörfler property $\eta(u_j, \Lambda_j, \mathcal{M}^*) \ge \vartheta_x \eta(u_j, \Lambda_j, \mathcal{T}_j)$. Due to the minimality of $\#\mathcal{M}_j$ and using (3.16) in the last step,

$$\#\mathcal{M}_j \le \#\mathcal{M}^* \le \#\mathcal{T}^* - \#\mathcal{T}_j \le \#\mathcal{T}^\epsilon - \#\mathcal{T}_{\mathrm{init}}.$$

The assertion follows by applying the bound $\#\mathcal{T}^\epsilon - \#\mathcal{T}_{\mathrm{init}} \le \epsilon^{-1/s}|u|_{\mathbb{A}_s}^{1/s}$ and inserting the definition of ϵ. □

Using the above tools, we derive the following optimality statement by an argument similar to ([2], Thm. 5.11). As illustrated by a comparison with (8.12), within any succession of spatial refinements in ASGFEM, the convergence of the total spatial error achieves the maximal rate s afforded by the approximation class $\mathbb{A}_s$.

Theorem 8.4. *If $u \in \mathbb{A}_s$, then for any $j_0 \in \mathbb{N}_0$ and any $j \ge j_0$ with $\Lambda_j = \Lambda_{j_0} =: \Lambda$,*

$$\Big(\|u_j - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma}\operatorname{osc}(u_j, \Lambda, \mathcal{T}_j)^2\Big)^{1/2} \le C|u|_{\mathbb{A}_s}\big(\#\mathcal{T}_j - \#\mathcal{T}_{j_0}\big)^{-s} \tag{8.16}$$

with a constant C depending only on $\mathbb{T}$, $\vartheta_x/\hat{\vartheta}_x$, c_η, C_η, $\bar{c}_{a,\delta}$, γ, ω_x, δ_x and ϱ.

Proof. Let $j \geq j_0$ with $\Lambda_j = \Lambda_{j_0}$. Due to ([1], Thm. 2.4, [20], Thm. 6.1), and Lemma 8.3,

$$\#\mathcal{T}_j - \#\mathcal{T}_{j_0} \leq c_{\mathbb{T}} \sum_{k=0}^{j-1} \#\mathcal{M}_k \leq M \sum_{k=0}^{j-1} \Big(\|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(u_k, \Lambda, \mathcal{T}_k)^2 \Big)^{-1/2s}$$

with $M = |u|_{\mathbb{A}_s}^{1/s} c_{\mathbb{T}} c_{\mathrm{red}}^{-1/2s} \hat{C}^{1/2s}$ and a constant $c_{\mathbb{T}}$ depending only on $\mathbb{T}$. For any $j_0 \leq k \leq j-1$, the lower bound in Corollary 5.2 implies

$$\begin{aligned} \|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_k^2 &\leq \Big(1 + \omega_x \frac{1+\gamma}{c_\eta}\Big) \|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \operatorname{osc}(u_k, \Lambda, \mathcal{T}_k)^2 \\ &\leq \Big(1 + \omega_x \frac{1+\gamma}{c_\eta}\Big) \Big(\|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(u_k, \Lambda, \mathcal{T}_k)^2 \Big). \end{aligned}$$

Furthermore, the contraction property from Theorem 7.3 implies

$$\|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_k^2 \geq \delta_x^{k-j} \big(\|u_j - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_j^2 \big).$$

Consequently,

$$\#\mathcal{T}_j - \#\mathcal{T}_{j_0} \leq M \Big(1 + \omega_x \frac{1+\gamma}{c_\eta}\Big)^{1/2s} \big(\|u_j - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_j^2 \big)^{-1/2s} \sum_{k=0}^{j-1} \delta_x^{(j-k)/2s}$$

and since $0 < \delta_x < 1$, the remaining sum is

$$\sum_{k=0}^{j-1} \delta_x^{(j-k)/2s} \leq \sum_{i=1}^{\infty} \delta_x^{i/2s} = \frac{\delta_x^{1/2s}}{1 - \delta_x^{1/2s}} =: D.$$

The assertion follows with the estimate

$$\|u_j - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(u_j, \Lambda, \mathcal{T}_j)^2 \leq \max\Big(1, \frac{c_\eta}{\omega_x(1+\gamma)}\Big) \big(\|u_j - u_\Lambda\|_{\mathcal{A}}^2 + \omega_x \eta_j^2 \big)$$

from (5.8). □

By a similar argument as in Theorem 8.4 leveraging the contraction property in Theorem 7.2 of the full error, we derive in Theorem 8.6 a statement concerning the convergence behavior of `ASGFEM` across both types of refinements.

Lemma 8.5. *For all $j \in \mathbb{N}$,*

$$\#\mathcal{T}_j \leq \#\mathcal{T}_0 + \#\mathcal{T}_{a,\operatorname{supp}\Lambda_j} + c_{\mathbb{T}} \sum_{k=0}^{j-1} \#\mathcal{M}_k \tag{8.17}$$

with a constant $c_{\mathbb{T}}$ depending only on $\mathbb{T}$, where we define $\mathcal{M}_k := \emptyset$ if $\eta_k < \varrho\zeta_k$.

Proof. If $\eta_k \geq \zeta_k$, then ([1], Thm. 2.4) and ([20], Thm. 6.1) imply

$$\#\mathcal{T}_{k+1} - \#\mathcal{T}_k \leq c_{\mathbb{T}} \#\mathcal{M}_k.$$

Conversely, if $\eta_k < \varrho\zeta_k$, then $\mathcal{T}_{k+1} = \mathcal{T}_k \oplus \mathcal{T}_{a,\operatorname{supp}\Lambda_{k+1}}$, and thus (3.16) implies

$$\#\mathcal{T}_{k+1} - \#\mathcal{T}_k \leq \#\mathcal{T}_{a,\operatorname{supp}\Lambda_{k+1}} - \#\mathcal{T}_{a,\operatorname{supp}\Lambda_k}$$

since $\mathcal{T}_{a,\operatorname{supp}\Lambda_k} \preceq \mathcal{T}_k$ and $\mathcal{T}_{a,\operatorname{supp}\Lambda_k} \preceq \mathcal{T}_{a,\operatorname{supp}\Lambda_{k+1}}$. The assertion follows by summing over $k = 0, \ldots, j-1$. □

Theorem 8.6. *If $u \in \mathbb{A}_s$, then for all $j \in \mathbb{N}_0$,*

$$\big(\|u_j - u\|_{\mathcal{A}}^2 + \omega_\eta \eta_j^2 + \omega_\zeta \zeta_j^2\big)^{1/2} \leq C|u|_{\mathbb{A}_s}\big(\#\mathcal{T}_j - \#\mathcal{T}_0 - \#\mathcal{T}_{a,\operatorname{supp}\Lambda_j}\big)^{-s} \tag{8.18}$$

with a constant C depending only on $\mathbb{T}$, $\vartheta_x/\hat{\vartheta}_x$, c_η, C_η, $\bar{c}_{a,\delta}$, γ, ω_η, ω_ζ, δ and ϱ.

Proof. Lemmas 8.5 and 8.3 imply

$$\#\mathcal{T}_j - \#\mathcal{T}_0 - \#\mathcal{T}_{a,\operatorname{supp}\Lambda_j} \leq c_{\mathbb{T}} \sum_{k=0}^{j-1} \#\mathcal{M}_k$$

with $\#\mathcal{M}_k = 0$ if $\eta_k < \varrho\zeta_k$ and

$$\#\mathcal{M}_k \leq |u|_{\mathbb{A}_s}^{1/s} c_{\mathrm{red}}^{-1/2s} \hat{C}^{1/2s} \Big(\|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(u_k, \Lambda, \mathcal{T}_k)^2\Big)^{-1/2s}$$

if $\eta_k \geq \varrho\zeta_k$. In this latter case, we use the upper bound in Corollary 5.3 and the lower bound in Corollary 5.2 to estimate

$$\begin{aligned}\|u_k - u\|_{\mathcal{A}}^2 + \omega_\eta \eta_k^2 + \omega_\zeta \zeta_k^2 &\leq \Big(\frac{C_\eta(1+\varrho^{-2})}{1-\gamma} + \omega_\eta + \omega_\zeta \varrho^{-2}\Big)\eta_k^2 \\ &\leq E\Big(\|u_k - u_\Lambda\|_{\mathcal{A}}^2 + \frac{c_\eta}{1+\gamma} \operatorname{osc}(u_k, \Lambda, \mathcal{T}_k)^2\Big)\end{aligned}$$

with $E := c_\eta(1+\gamma)^{-1}[C_\eta(1+\varrho^{-2})(1-\gamma)^{-1} + \omega_\eta + \omega_\zeta\varrho^{-2}]$. Theorem 7.2 provides the bound

$$\|u_k - u\|_{\mathcal{A}}^2 + \omega_\eta \eta_k^2 + \omega_\zeta \zeta_k^2 \geq \delta^{j-k}\big(\|u_j - u\|_{\mathcal{A}}^2 + \omega_\eta \eta_j^2 + \omega_\zeta \zeta_j^2\big),$$

and thus

$$\#\mathcal{T}_j - \#\mathcal{T}_0 - \#\mathcal{T}_{a,\operatorname{supp}\Lambda_j} \leq |u|_{\mathbb{A}_s}^{1/s} c_{\mathbb{T}} c_{\mathrm{red}}^{-1/2s} \hat{C}^{1/2s} E^{1/2s} D\big(\|u_j - u\|_{\mathcal{A}}^2 + \omega_\eta \eta_j^2 + \omega_\zeta \zeta_j^2\big)^{-1/2s}$$

with $D = \delta^{1/2s}(1-\delta^{1/2s})^{-1}$. □

Since the error indicator η_j alone is equivalent to the total spatial error by (8.2), the estimate in Theorem 8.6 carries over to the total spatial error with a different constant, thereby extending Theorem 8.4 to the full set of approximations generated in `ASGFEM`.

Remark 8.7. Theorem 8.6 can be interpreted as a bound on the number of cells in the mesh $\mathcal{T}_j$,

$$\#\mathcal{T}_j \leq \#\mathcal{T}_0 + \#\mathcal{T}_{a,\operatorname{supp}_{\Lambda_j}} + C^{1/s}|u|_{\mathbb{A}_s}^{1/s}\big(\|u_j - u\|_{\mathcal{A}}^2 + \omega_\eta \eta_j^2 + \omega_\zeta \zeta_j^2\big)^{1/2s}. \tag{8.19}$$

If the meshes $\mathcal{T}_{\bar{a}}$ and $\mathcal{T}_{a,m}$ are minimal in $\mathbb{T}$ with respect to the partial order $\preceq$ subject to the conditions $\|h_{\mathcal{T}_{\bar{a}}}\nabla\bar{a}/\bar{a}\|_{L^\infty(D)} \leq \bar{c}_{a,\delta}$ and $\|h_{\mathcal{T}_{a,m}}\nabla a_m/\bar{a}\|_{L^\infty(D)} \leq \bar{c}_{a,\delta}\|a_m/\bar{a}\|_{L^\infty(D)}$, then $\mathcal{T}_{a,\operatorname{supp}\Lambda_j}$ is minimal in $\mathbb{T}$ subject to $c_{a,\delta}(\Lambda_j, \mathcal{T}_{a,\operatorname{supp}\Lambda_j}) \leq \bar{c}_{a,\delta}$, *i.e.* for any mesh $\mathcal{T} \in \mathbb{T}$, $c_{a,\delta}(\Lambda_j, \mathcal{T}) \leq \bar{c}_{a,\delta}$ implies $\mathcal{T}_{a,\operatorname{supp}\Lambda} \preceq \mathcal{T}$. In particular, the term $\#\mathcal{T}_{a,\operatorname{supp}_{\Lambda_j}}$ in (8.19) is minimal subject to $c_{a,\delta}(\Lambda_j, \mathcal{T}_j) \leq \bar{c}_{a,\delta}$, and the spatial refinement performed in `ASGFEM` in the case $\eta_{j-1} < \varrho\zeta_{j-1}$ is the minimal refinement required to ensure this property.

9. Numerical examples

The implementation of the proposed adaptive algorithm of Section 6 uses the open source framework `ALEA` [8] which was already the basis for the ASGFEM in [7]. In comparison to that paper, the main difference here is the use of a single adaptively refined mesh for all gpc modes. Moreover, higher order conforming finite element spaces are employed. By the restriction to a single mesh, the projection of solutions between different meshes is no longer required which was one of the main computational tasks of the first adaptive algorithm. Hence, this approach represents a substantial simplification for the actual implementation and evaluation of the numerical solution. In order to distinguish the two approaches, we denote by ASGFEM2 the algorithm presented in this paper and the preceding algorithm by ASGFEM1. The implementation of ASGFEM2 is based on the code of ASGFEM1 and follows to a large extend the description given in [7]. There, the construction of the operator and the treatment of inhomogeneous Dirichlet boundary conditions in the given setting was discussed. For the adaptive algorithm of Section 6, a different bound for the tail estimation and a modified marking strategy had to be implemented. Apart from these extensions, only minor adjustments of the existing code were required.

The evaluation of the energy error of the numerical solution with regard to some reference solution is described in Section 9.1. The performance of the new algorithm applied to some of the benchmark problems from [7] is assessed in Section 9.2.

Since the construction of different adapted meshes with ASGFEM1 results in an optimised sparse representation of the problem, it is interesting to compare the adaptive approaches for multi (sparse) and single mesh adaptivity. This is done in Section 9.3. A central observation in [14] is that higher order approximations can (under certain conditions) compensate for sparsity which is illustrated by the results, for sufficiently regular solutions.

9.1. Evaluation of the error

For experimental verification of the reliability of the error estimator, a reference error is computed by Monte Carlo simulations. For this, a set of M independent realizations $\{y^{(i)}\}_{i=1}^{M}$ of the stochastic parameters is computed. The $y_m^{(i)}$ are sampled according to the probability measure π_m of the random variable y_m. The mean-square error e of the parametric SGFEM solution $u_N \in \mathcal{V}_N$ is approximated by a Monte Carlo sample average

$$\|e\|_V^2 = \int_\Gamma \|u(y) - u_N(y)\|_V^2 \, d\pi(y) \approx \frac{1}{M} \sum_{i=1}^{M} \|\tilde{u}(y^{(i)}) - u_N(y^{(i)})\|_V^2 \, . \tag{9.1}$$

Here, the samples $y^{(i)} \in \Gamma$ of parameter sequences are assumed to be statistically independent and identically distributed with law π. Note that the sampled solutions $\tilde{u}(y^{(i)})$ are approximations of the exact $u(y^{(i)}) = A^{-1}(y^{(i)})f$ since the operator is discretized on a reference mesh. This mesh is determined as the union of the finest meshes, *i.e.*, the meshes of the respective last iteration of all polynomial degrees in each experiment, and a subsequent uniform refinement. Moreover, the expansion (2.1) of the random field $a(y,x)$ is truncated to the maximal length occuring in the approximate parametric solutions with another 20 gpc modes added to the tail. We choose $M = 150$ for the Monte Carlo approximation of the reference error (9.1) which proved to be sufficient to assess the reliability of the error estimator.

9.2. The stochastic diffusion problem

We examine numerical simulations for the stationary diffusion problem (2.2) in a plane, polygonal domain $D \subset \mathbb{R}^2$. Recall from Section 2 that $x = (x_1, x_2) \in D$ denotes points in D and $y = (y_1, y_2, \dots) \in \Gamma$ denotes the parameter sequence in the coefficient (2.1).

As in [7], the expansion coefficients of the stochastic field (2.1) are chosen to be

$$a_m(x) := \alpha_m \cos(2\pi\beta_1(m)x_1) \cos(2\pi\beta_2(m)x_2) \tag{9.2}$$

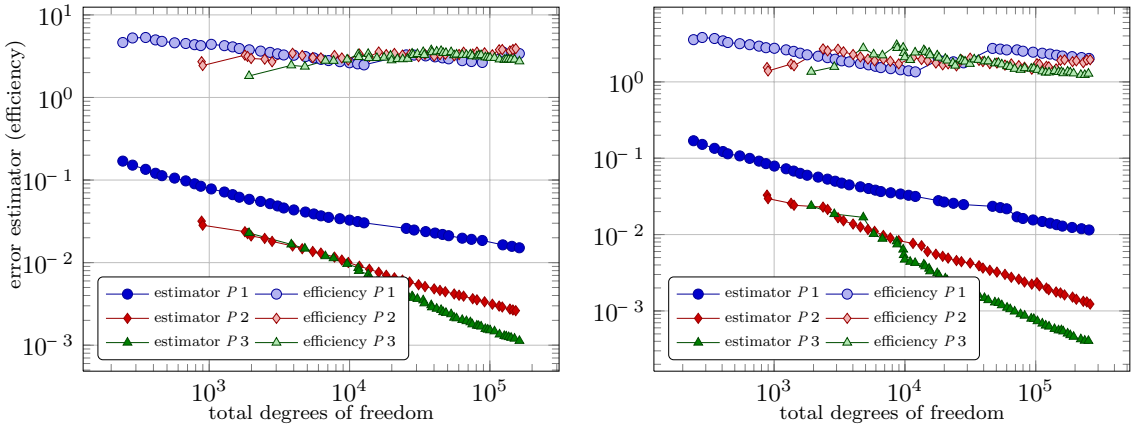

FIGURE 1. Convergence of the error estimator in the energy norm with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the square with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay. Total number of degrees of freedom and efficiency of the error estimator with respect to the MC reference error.

where α_m is of the form $\bar{\alpha} m^{-\tilde{\sigma}}$ with $\tilde{\sigma} > 1$ and some $0 < \bar{\alpha} < 1/\zeta(\tilde{\sigma})$ with the Riemann zeta function ζ. Then, (2.3) holds with $\gamma = \bar{\alpha}\zeta(\tilde{\sigma})$. Moreover,

$$\beta_1(m) = m - k(m)(k(m)+1)/2 \quad \text{and} \quad \beta_2(m) = k(m) - \beta_1(m) \tag{9.3}$$

with $k(m) := \lfloor -1/2 + \sqrt{1/4 + 2m} \rfloor$, *i.e.*, the coefficient functions a_m enumerate all planar Fourier sine modes in increasing total order. To illustrate the influence which the stochastic coefficient plays in the adaptive algorithm, we examine the expansion with slow and fast decay of α_m, setting $\tilde{\sigma}$ in (9.2) to either 2 or 4. The computations are carried out with conforming FEM spaces of polynomial degree 1, 2 and 3.

For the adaptive algorithm of Section 6.3 the parameters are chosen as

$$\vartheta_x = 2/5, \quad \vartheta_y = 10 \quad \text{and} \quad \epsilon = 10^{-8}\,.$$

The employed quadrature is exact for polynomials up to degree 20.

9.2.1. Square domain

The first example is the stationary diffusion equation (2.2) on the unit square $D = (0,1)^2$ with homogeneous Dirichlet boundary conditions and with right-hand side $f = 1$. The results of the adaptive algorithm of Section 6.3 for a slow decay of the coefficients with $\tilde{\sigma} = 2$ and a fast decay with $\tilde{\sigma} = 4$ are shown in Figures 1 and 2. The amplitude $\bar{\alpha}$ in (9.2) was chosen as $\gamma/\zeta(\tilde{\sigma})$ with $\gamma = 0.9$, resulting in $\bar{\alpha} \approx 0.547$ for $\tilde{\sigma} = 2$ and $\bar{\alpha} \approx 0.832$ for $\tilde{\sigma} = 4$. Depicted is the residual estimator, the reference error obtained by Monte Carlo sampling, the efficiency of the estimator and the number of active multi-indices. The observed convergence rate of 1/2 for P1 FEM with respect to the total number of degrees of freedom, which is the convergence rate for a single non-parametric problem, coincides with the approximation rates predicted by [5, 13]. Both $\tilde{\sigma} = 2$ and $\tilde{\sigma} = 4$ afford sufficient summability of the coefficients of the solution to attain the convergence rate of the spatial discretization for a single non-parametric problem, as elaborated in [5, 13]. For quadratic and cubic FEM spaces, the convergence rate increases, also see Figure 9. However, the rate achieved with P3 is not consistently better than that of a P2 discretisation as the error estimator in Figure 1 might suggest.

The efficiency indices for the different polynomial degrees are similar and lie between 1 and 10. Since the reliability bound of the error estimator contains unknown constants, the purpose of the efficiency graphs in this and the next subsection is mainly to illustrate the progression of the estimator/error ratio for polynomial

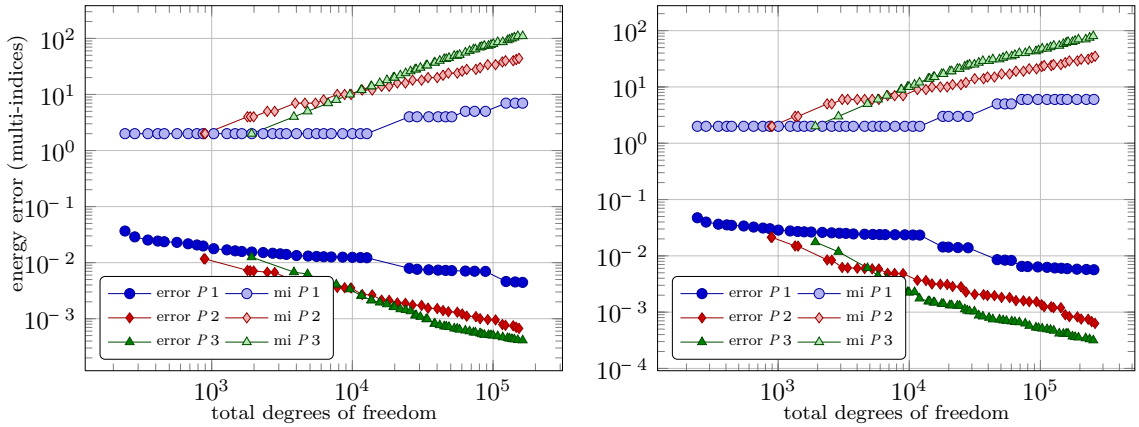

FIGURE 2. Convergence of the error in the energy norm with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the square with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay. Total number of degrees of freedom and active multi-indices.

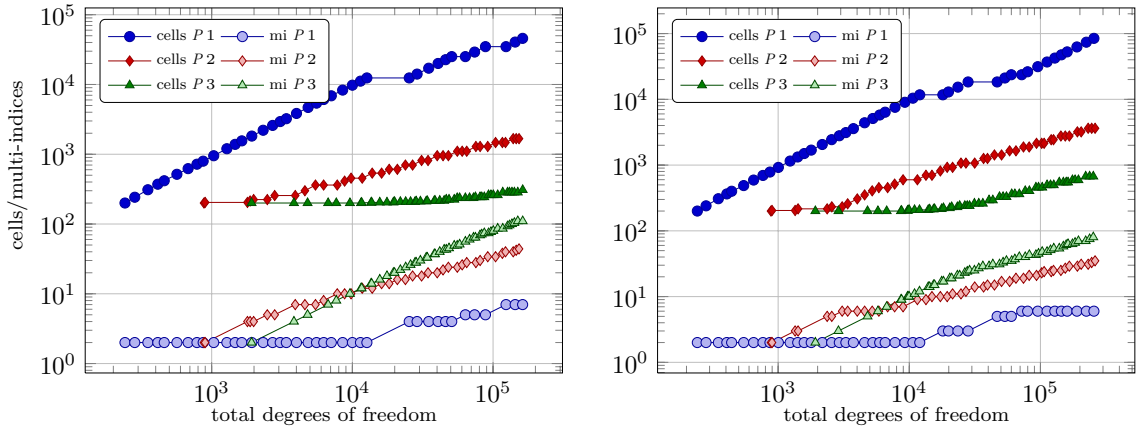

FIGURE 3. Number of mesh cells and active multi-indices with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the square with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay with respect to total number of degrees of freedom.

FE degrees 1−3 and not to show the accuracy of the error estimator. We further observe that the number of activated gpc modes increases substantially with the polynomial degree of the FE approximation. At the same time, the grids remain relatively coarse in comparison to the P1 FEM. This feature is illustrated in Figure 3 which depicts the number of mesh cells and active multi-indices in the course of the adaptive algorithm. On the one hand, higher order FEM activate significantly more multi-indices (more than 100) while the mesh is kept relatively coarse at the same time. On the other hand, P1 FEM leads to a strongly refined mesh and only few activated multi-indices (less than 10). Of course, higher order finite elements methods compensate for the coarser mesh through the higher local polynomial degree. The relation of active multi-indices to total energy error is depicted in Figure 4. This illustrates the independence of the multi-index activation with regard to the polynomial degree of the spatial approximation.

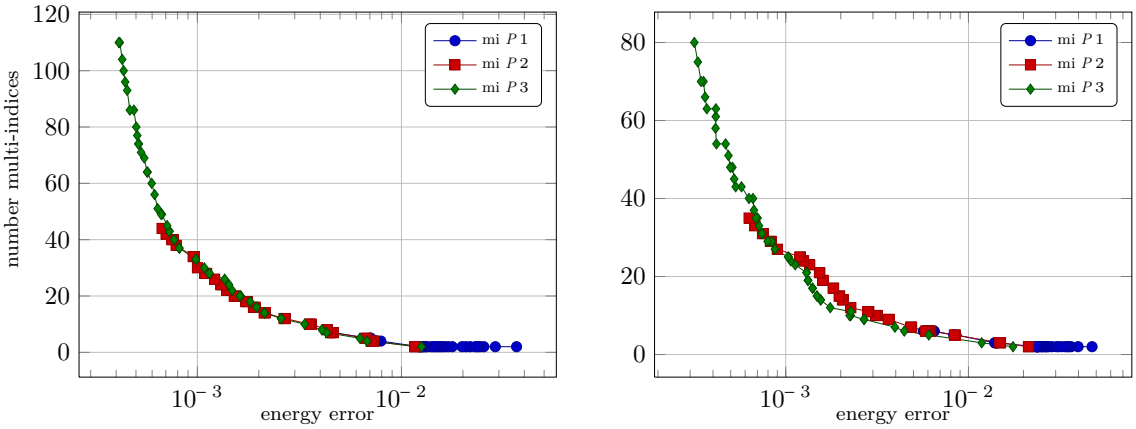

FIGURE 4. Number of active multi-indices with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the square domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay with respect to the energy error.

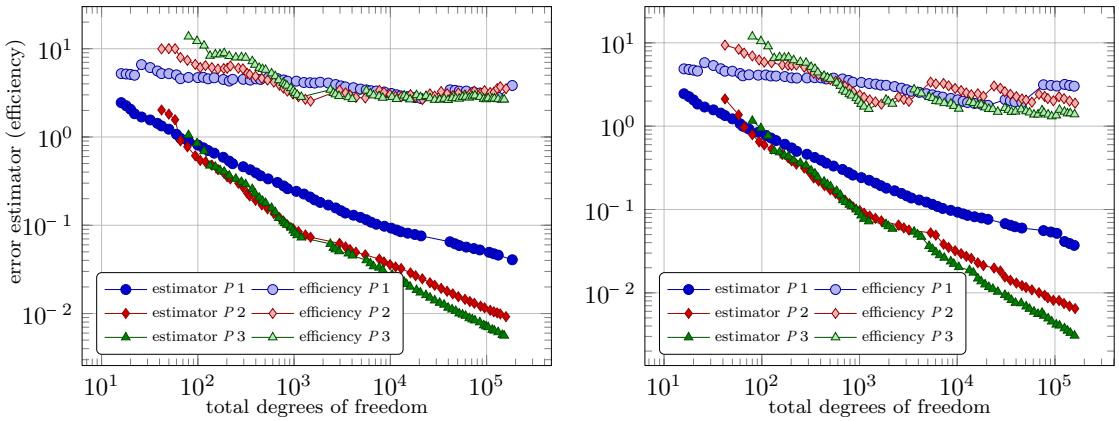

FIGURE 5. Convergence of the error estimator in the energy norm with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the L-shaped domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay. Total number of degrees of freedom and efficiency of the error estimator with respect to the MC reference error.

A comparison with regard to the two decay rates reveals that the adaptive algorithm activates more multi-indices in the case of slower decay (left-hand side in all figures with $\tilde{\sigma} = 2$) since more terms in (2.1) are required for an accurate representation than for faster decay (right-hand side in all figures with $\tilde{\sigma} = 4$).

9.2.2. L-shaped domain

A standard benchmark problem for deterministic *a posteriori* error estimators is the stationary diffusion problem (2.2) on the L-shaped domain $D = (-1,1)^2 \setminus (0,1) \times (-1,0)$. It is well-known that the solution exhibits a singularity at the reentrant corner at $(0,0)$ which is resolved by a pronounced mesh refinement in its vicinity. The convergence of the error estimator and its efficiency with regard to the error determined by (9.1) are depicted in Figure 5. In Figure 6, the error and the number of active multi-indices are shown. The relation of active multi-indices to total energy error is depicted in Figure 8. As before, the multi-index activation is (nearly) independent of the polynomial degree of the spatial approximation.

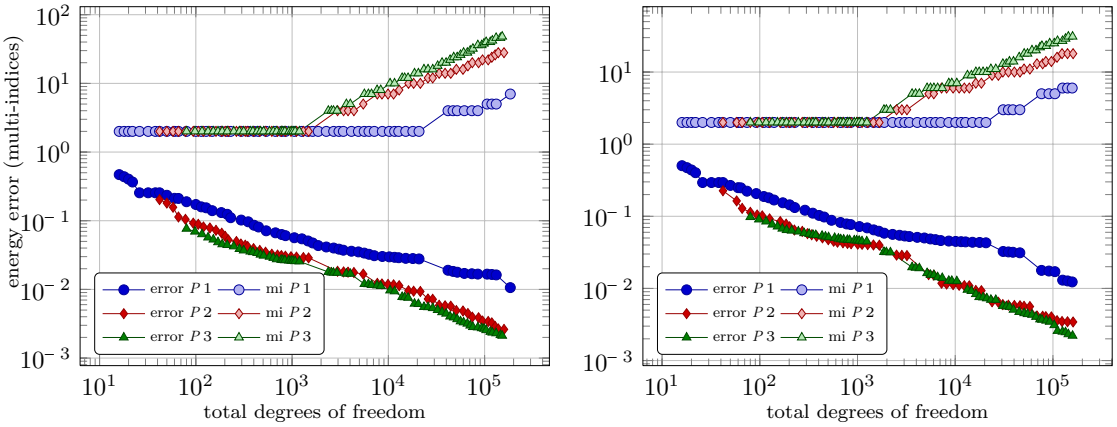

FIGURE 6. Convergence of the error in the energy norm with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the L-shaped domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay. Total number of degrees of freedom and active multi-indices.

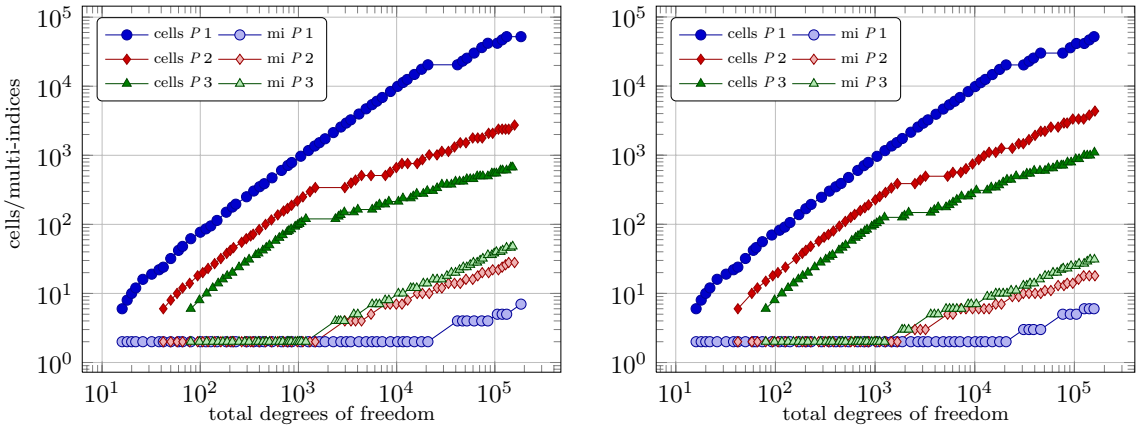

FIGURE 7. Number of mesh cells and active multi-indices with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the L-shaped domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay with respect to total number of degrees of freedom.

In order to assess the relation between deterministic and stochastic refinement, Figure 7 depicts the number of mesh cells and active multi-indices in the course of the adaptive algorithm. As compared to the experiment on the square in Subsection 9.2.1, now the mesh is strongly refined for all polynomial degrees up to about 10^3 degrees of freedom to resolve the singularity at the reentrant corner. Subsequently, the higher order spatial discretisations favour the refinement of the stochastic space by activation of new multi-indices while the P1 FEM results in a continued strong refinement of the mesh. Similar to the previous experiment, the efficiency indices lie closely together between 1 and 10. Preasymptotically, the difference between the two decay rates with regard to the activated multi-indices is less pronounced than before. This is due to the delayed stochastic refinement which is an effect of the initial singularity resolution of the adaptive algorithm. The P3 FEM only leads to marginal improvements of the empirical error convergence as compared to P2 FEM, also see Figure 10.

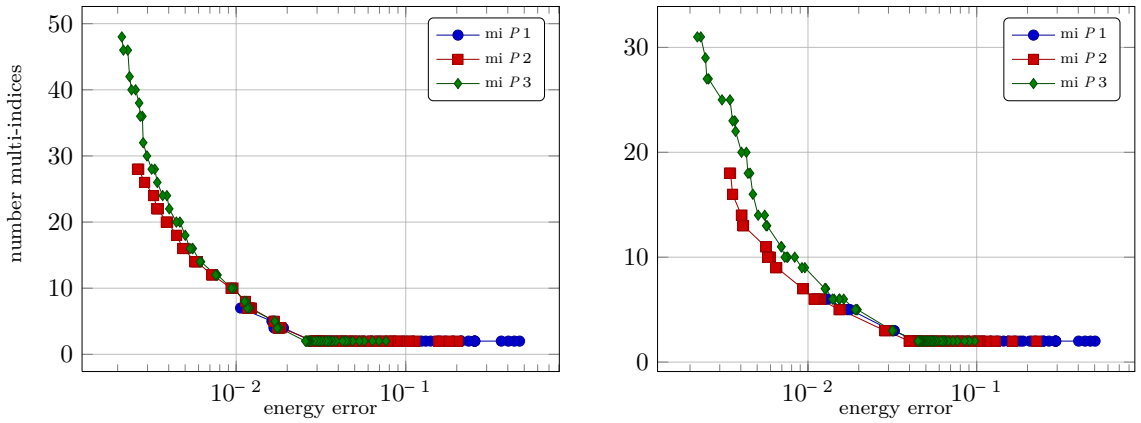

FIGURE 8. Number of active multi-indices with FEM of degree 1, 2 and 3 for the stationary diffusion problem on the L-shaped domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay with respect to the energy error.

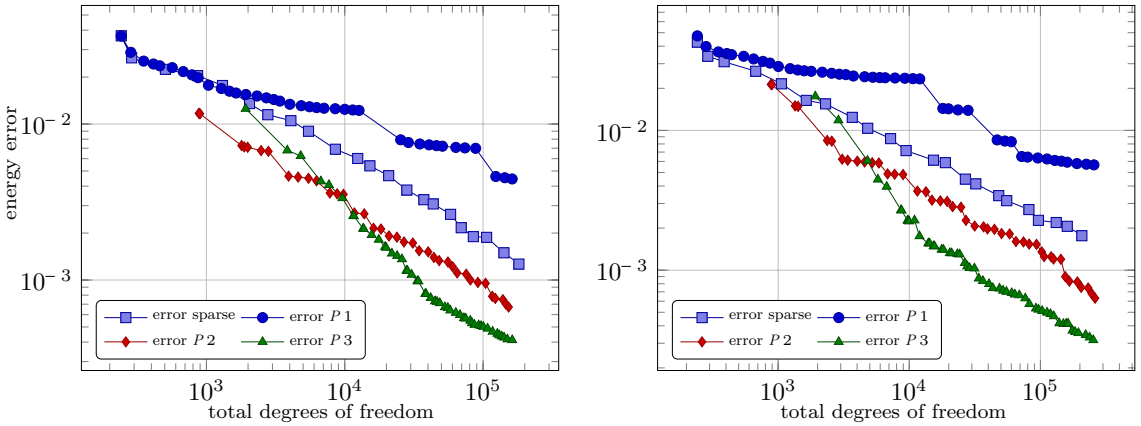

FIGURE 9. Convergence of the error in the energy norm for the stationary diffusion problem on the square domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay. Comparison of ASGFEM1 (sparse) and ASGFEM2 for polynomial degrees 1, 2 and 3.

9.3. Comparison of adaptive algorithms

This section is devoted to the comparison of the adaptive algorithms ASGFEM1 of [7] and ASGFEM2 of Section 6.

In Figure 9, the error graphs for the stationary diffusion problem of Section 9.2.1 for $\tilde{\sigma} = 2$ and $\tilde{\sigma} = 4$ are depicted for the sparse ASGEM1 and ASGFEM2 with polynomial degrees 1, 2 and 3. The parameters for ASGFEM1 are set to

$$\bar{c}_Q = 1, \quad \bar{c}_\eta = 1, \quad \vartheta_\eta = 2/5, \quad \vartheta_\zeta = 10^{-1}, \quad \vartheta_\delta = 10, \quad \chi = 1/10, \quad \epsilon = 10^{-8}$$

with the same ASGFEM2 parameters as above.

It can be observed that the sparse ASGFEM1 with different adapted meshes performs better than ASGEM2 with affine FEM. In particular, the error reduction seems more uniform and the error is smaller than the

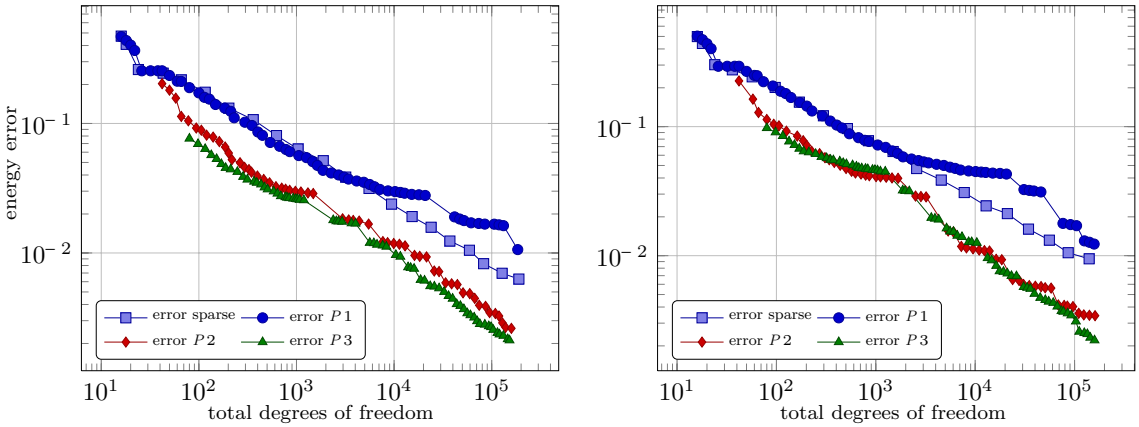

FIGURE 10. Convergence of the error in the energy norm for the stationary diffusion problem on the L-shaped domain with homogeneous Dirichlet boundary conditions for slow ($\tilde{\sigma} = 2$, *left*) and fast ($\tilde{\sigma} = 4$, *right*) decay. Comparison of ASGFEM1 (sparse) and ASGFEM2 for polynomial degrees 1, 2 and 3.

one obtained with ASGFEM2 for affine FEM. However, for higher order approximations, the new adaptive algorithm with a single joint mesh outperforms the adapted sparse ASGFEM1 approximations by nearly an order of magnitude for P3 FEM. Moreover, in terms of the total number of degrees of freedom, the error reduction rate increases as the polynomial degree used is increased.

In the next experiment, whose results are shown in Figure 10, we examine the two adaptive algorithms for the stationary diffusion problem on the L-shaped domain introduced in Section 9.2.2. The parameters for ASGFEM1 are set to

$$\bar{c}_Q = 1, \quad \bar{c}_\eta = 1, \quad \vartheta_\eta = 3/5, \quad \vartheta_\zeta = 10^{-2}, \quad \vartheta_\delta = 1, \quad \chi = 1/10, \quad \epsilon = 10^{-8}$$

with the parameters of ASGFEM2 as before.

We observe that for this example, ASGFEM1 and ASGFEM2 exhibit nearly identical convergence of the error for affine finite element spaces. Unlike what we found in the previous comparison, the error graphs for the P1 FEM lie closely together. Again, for higher order FEM, both the convergence rate and the constants exhibited with ASGFEM2 are improved over ASGFEM1. However, as mentioned earlier, the error reduction rate of P3 does not appear to improve significantly over P2 FEM.

10. Conclusions

We analyzed the convergence for a class of adaptive Galerkin discretizations of countably-parametric, self-adjoint scalar diffusion problems. The Galerkin discretizations are based on a mean-square (with respect to a probability measure on the infinite-dimensional parameter space) energy (with respect to a variational formulation of the problem in physical space) projection of the parametric solution onto a tensor product of a polynomial chaos on parameter space and standard, H^1-conforming Finite Element spaces on families of adaptively refined, regular simplicial triangulations of the physical domain D, subject to the constraint that the same Finite Element subspaces of $H^1_0(D)$ are used for the approximation of all active polynomial chaos coefficients. A residual error estimator was proposed which allows to distinguish between error contributions from the polynomial chaos discretization in the parameter space and the Finite Element discretization in physical space. The estimator was shown to be reliable and, for any fixed set of active gpc modes, efficient (up to a suitable data oscillation term).

Based on the splitting of the error contributions in the residual error estimator, we proposed modules $\texttt{Estimate}_x$, $\texttt{Estimate}_y$ and corresponding modules $\texttt{Mark}_x$, $\texttt{Mark}_y$ and $\texttt{Refine}_x$, $\texttt{Refine}_y$ in a novel, *anisotropic refinement algorithm*. We proved that the proposed algorithm is convergent, *i.e.* that it produces sequences of finitely supported iterates and that it terminates after a finite number of iterations for any prescribed tolerance. We showed quasi-optimality of the spatial adaptations at any fixed, finite set of activated gpc modes in the Galerkin approximation. On a set of test problems with varying degrees of sparsity in the coefficient sequence of the gpc expansion of the exact solution and with corner singularities in the physical domain D, the proposed strategy identifies correctly the sparsity in the gpc expansion and the corner singularities in the physical domain. The optimality of the combined adaptive algorithm is the subject of further research.

References

[1] P. Binev, W. Dahmen and R. DeVore, Adaptive finite element methods with convergence rates. *Numer. Math.* **97** (2004) 219–268.

[2] J.M. Cascon, C. Kreuzer, R.H. Nochetto and K.G. Siebert, Quasi-optimal convergence rate for an adaptive finite element method. *SIAM J. Numer. Anal.* **46** (2008) 2524–2550.

[3] A. Chkifa, A. Cohen, R. DeVore and C. Schwab, Adaptive algorithms for sparse polynomial approximation of parametric and stochastic elliptic pdes. *ESAIM: M2AN* **47** (2013) 253–280.

[4] A. Chkifa, A. Cohen and C. Schwab, High-dimensional adaptive sparse polynomial interpolation and applications to parametric pdes. *J. Found. Comput. Math.* **14** (2014) 601–633.

[5] A. Cohen, R. Devore and C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl.* **9** (2011) 11–47.

[6] W. Dörfler, A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.* **33** (1996) 1106–1124.

[7] M. Eigel, C. Gittelson, C. Schwab and E. Zander, Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Engrg.* **270** (2014) 247–269.

[8] M. Eigel and E. Zander, ALEA - A Python Framework for Spectral Methods and Low-Rank Approximations in Uncertainty Quantification. Available at: `https://bitbucket.org/aleadev/alea`.

[9] W. Gautschi, Orthogonal polynomials: computation and approximation. *Numer. Math. Sci. Comput.* Oxford University Press, New York (2004).

[10] R.G. Ghanem and P.D. Spanos, Stochastic finite elements: a spectral approach. Springer-Verlag, New York (1991).

[11] C.J. Gittelson, R. Andreev and Ch. Schwab, Optimality of adaptive Galerkin methods for random parabolic partial differential equations. *J. Comput. Appl. Math.* **263** (2014) 189–201.

[12] C.J. Gittelson, *Stochastic Galerkin approximation of operator equations with infinite dimensional noise*. Tech. Report 2011-10. Seminar for Applied Mathematics, ETH Zürich (2011).

[13] C.J. Gittelson, Convergence rates of multilevel and sparse tensor approximations for a random elliptic PDE. *SIAM J. Numer. Anal.* **51** (2013) 2426–2447.

[14] C.J. Gittelson, High-order methods as an alternative to using sparse tensor products for stochastic galerkin FEM. *Comput. Math. Appl.* **67** (2014) 888-898.

[15] F.Y. Kuo, C. Schwab and I.H. Sloan, Quasi-monte carlo finite element methods for a class of elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal.* **50** (2012) 3351–3374.

[16] O.P. Le Maître and O.M. Knio, Spectral methods for uncertainty quantification, Scientific Computation. Springer, New York (2010). With applications to computational fluid dynamics.

[17] P. Morin, R.H. Nochetto and K.G. Siebert, Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.* **38** (2000) 466–488.

[18] R.H. Nochetto, K.G. Siebert and A. Veeser, Theory of adaptive finite element methods: an introduction, in *Multiscale, Nonlinear and Adaptive Approximation*. Springer, Berlin (2009) 409–542.

[19] C. Schillings and Ch. Schwab Sparse, adaptive Smolyak quadratures for Bayesian inverse problems. *Inverse Probl.* **29** (2013) 065011.

[20] R. Stevenson, The completion of locally refined simplicial partitions created by bisection. *Math. Comput.* (2008) 227–241.

[21] R. Verfürth, A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques. Teubner Verlag and J. Wiley, Stuttgart (1996).