Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Convergence and rate optimality of adaptive multilevel stochastic Galerkin FEM

CONVERGENCE AND RATE OPTIMALITY OF ADAPTIVE MULTILEVEL STOCHASTIC GALERKIN FEM

ALEX BESPALOV, DIRK PRAETORIUS, AND MICHELE RUGGERI

ABSTRACT. We analyze an adaptive algorithm for the numerical solution of parametric elliptic partial differential equations in two-dimensional physical domains, with coefficients and the right-hand side functions depending on infinitely many (stochastic) parameters. The algorithm generates multilevel stochastic Galerkin approximations; these are represented in terms of a sparse generalized polynomial chaos expansion with coefficients residing in finite element spaces associated with different locally refined meshes. Adaptivity is driven by a two-level *a posteriori* error estimator and employs a Dörfler-type marking on the joint set of spatial and parametric error indicators. We show that, under an appropriate saturation assumption, the proposed adaptive strategy yields optimal convergence rates with respect to the overall dimension of the underlying multilevel approximation spaces.

1. INTRODUCTION

Adaptive solution algorithms for partial differential equations (PDEs) with parametric or uncertain inputs is an active and topical research area. Adaptive algorithms are especially useful in the case of inputs depending on a large or countably infinite number of uncertain parameters; they provide mechanisms for incorporating a finite set of parameters into discretizations, enriching this set incrementally, and tuning the resulting parametric approximations to their spatial counterparts.

In particular, adaptivity is the key to efficient stochastic Galerkin finite element method (SGFEM), where approximations are typically represented as finite (sparse) generalized polynomial chaos (gPC) expansions with spatial coefficients residing in finite element spaces. While in the simplest (so-called *single-level*) SGFEM all spatial coefficients reside in the same finite element space, a more flexible *multilevel* construction allows spatial gPC-coefficients to reside in different finite element spaces.

Several adaptive algorithms driven by bespoke *a posteriori* error estimators have been proposed in the framework of *single-level* SGFEM (see [EGSZ15, EM16, BS16, BR18, BPRR19]), with convergence analysis presented in [EGSZ15, BPRR19]. At each iteration of the adaptive loop, these algorithms incrementally enrich *either* the finite element space *or* the gPC expansion. Note that *combined* enrichments of spatial and parametric components of single-level SGFEM approximations at each iteration of the adaptive algorithm

are prohibitively expensive due to the multiplicative increase of the total number of degrees of freedom. Furthermore, it is evident from the numerical experiments presented in the above works that, staying within the *single-level* framework, SGFEM cannot achieve the convergence rate of the chosen FEM for parameter-free problems.

Multilevel SGFEMs have emerged in the works by Cohen, DeVore and Schwab [CDS10, CDS11] and Gittelson [Git13]. These works have provided theoretical benchmarks for convergence analysis of the SGFEM by proving the existence of a sequence of multilevel approximation spaces such that the errors in the associated Galerkin solutions converge to zero with the same rate as the errors in the chosen FEM for the corresponding parameter-free problem. Practical adaptive algorithms generating such sequences of approximation spaces and Galerkin solutions have been developed in [EGSZ14, CPB19, BPR20]. In particular, an algorithm with *combined* enrichment of spatial and parametric components has been proposed and implemented in [BPR20] (see Algorithm 7 with Marking Criterion C therein). This algorithm is driven by a two-level *a posteriori* error estimator (that was also introduced in [BPR20]) and employs a Dörfler-type marking on the joint set of all spatial and parametric error indicators. In the numerical experiments reported in [EGSZ14, CPB19, BPR20], the theoretically predicted convergence rates have been observed for adaptive multilevel SGFEM approximations. However, a provable convergence or optimality result for any of the developed adaptive multilevel algorithms has been an open problem.

In this paper, we consider the same parametric model problem as in the above cited works—the steady-state diffusion equation with a spatially varying coefficient that has affine dependence on infinitely many parameters. We study convergence and rate optimality of multilevel SGFEM approximations generated by the adaptive algorithm with combined marking/enrichment as proposed in [BPR20]. To the best of our knowledge, this is the first time when convergence and rate optimality are analyzed for a specific adaptive algorithm in the framework of *multilevel* SGFEM. Assuming appropriate saturation assumptions, the main result of this work is twofold (see Theorem 5): (i) the adaptive algorithm with combined marking/enrichment ensures linear convergence of generated multilevel SGFEM approximations; (ii) the decay of the energy errors in these approximations is rate optimal with respect to the overall dimension of the underlying multilevel approximation spaces (here, the rate optimality is understood as the best algebraic rate that can possibly be achieved for a given approximation class).

There are some interesting by-products of our theoretical analysis. We have introduced a new concept of *multilevel structure* (a spatio-parametric discrete structure that underpins a multilevel SGFEM approximation space in the same way as a mesh-degree combination underpins a finite element space). We have proved (see Lemmas 7 and 8) that refinements of multilevel structures satisfy the overlay and closure estimates—the well-known properties of spatial meshes refined by newest vertex bisection. We have also proved that under an appropriate saturation assumption, the combined Dörfler marking is optimal, in the sense that it is equivalent to linear error reduction (see Proposition 6).

The rest of the paper is organized as follows. Section 2 introduces the model parametric problem and its weak formulation. In section 3, we describe spatial and parametric components of discrete multilevel structures, the associated finite-dimensional spaces, multilevel SGFEM discretizations, and *a posteriori* error estimators. The adaptive algorithm and

the main result of this work are formulated in section 4. The proofs of linear convergence and rate optimality are given, respectively, in sections 5 and 6. The results of some numerical experiments are reported in section 7.

2. Problem formulation

Let $D \subset \mathbb{R}^2$ be a bounded Lipschitz domain with polygonal boundary ∂D and let $\Gamma := \prod_{m=1}^{\infty}[-1,1]$ denote the infinitely-dimensional hypercube. We will refer to D and Γ as the physical domain and the parameter domain, respectively. We consider the elliptic boundary value problem

$$
\begin{aligned}
-\nabla \cdot (\boldsymbol{a}\nabla\boldsymbol{u}) &= \boldsymbol{f} && \text{in } D \times \Gamma, \\
\boldsymbol{u} &= 0 && \text{on } \partial D \times \Gamma.
\end{aligned}
\tag{1}
$$

Here, the scalar coefficient $\boldsymbol{a}$ and the right-hand side $\boldsymbol{f}$ (and, hence, the solution $\boldsymbol{u}$) depend on a countably infinite number of scalar parameters, i.e., $\boldsymbol{a} = \boldsymbol{a}(x, \mathbf{y})$, $\boldsymbol{f} = \boldsymbol{f}(x, \mathbf{y})$, and $\boldsymbol{u} = \boldsymbol{u}(x, \mathbf{y})$ with $x \in D$ and $\mathbf{y} \in \Gamma$. We assume affine dependence of the coefficient $\boldsymbol{a}$ on the parameters, i.e.,

$$
\boldsymbol{a}(x, \mathbf{y}) = a_0(x) + \sum_{m=1}^{\infty} y_m a_m(x) \quad \text{for all } x \in D \text{ and } \mathbf{y} = (y_m)_{m \in \mathbb{N}} \in \Gamma,
\tag{2}
$$

and that $\boldsymbol{f} \in L^2_\pi(\Gamma; H^{-1}(D))$, where $\pi = \pi(\mathbf{y})$ is a measure on $(\Gamma, \mathcal{B}(\Gamma))$ with $\mathcal{B}(\Gamma)$ being the Borel σ-algebra on Γ. Moreover, we assume that $\pi(\mathbf{y})$ is the product of symmetric Borel probability measures π_m on $[-1, 1]$, i.e., $\pi(\mathbf{y}) = \prod_{m=1}^{\infty} \pi_m(y_m)$.

For each $m \in \mathbb{N}_0$, the scalar functions $a_m \in L^\infty(D)$ in (2) are required to satisfy the following inequalities (cf. [SG11, Section 2.3]):

$$
0 < a_0^{\min} \le a_0(x) \le a_0^{\max} < \infty \quad \text{for almost all } x \in D,
\tag{3}
$$

$$
\tau := \frac{1}{a_0^{\min}} \left\| \sum_{m=1}^{\infty} |a_m| \right\|_{L^\infty(D)} < 1 \quad \text{and} \quad \sum_{m=1}^{\infty} \|a_m\|_{L^\infty(D)} < \infty.
\tag{4}
$$

With $\mathbb{X} := H_0^1(D)$, we consider the Bochner space $\mathbb{V} := L^2_\pi(\Gamma; \mathbb{X})$ and define the following bilinear forms on $\mathbb{V}$:

$$
B_0(\boldsymbol{u}, \boldsymbol{v}) := \int_\Gamma \int_D a_0(x) \nabla \boldsymbol{u}(x, \mathbf{y}) \cdot \nabla \boldsymbol{v}(x, \mathbf{y}) \, dx \, d\pi(\mathbf{y}),
\tag{5}
$$

$$
B(\boldsymbol{u}, \boldsymbol{v}) := B_0(\boldsymbol{u}, \boldsymbol{v}) + \sum_{m=1}^{\infty} \int_\Gamma \int_D y_m a_m(x) \nabla \boldsymbol{u}(x, \mathbf{y}) \cdot \nabla \boldsymbol{v}(x, \mathbf{y}) \, dx \, d\pi(\mathbf{y}).
\tag{6}
$$

It is easy to see that assumptions (2)–(4) imply that the bilinear forms $B_0(\cdot, \cdot)$ and $B(\cdot, \cdot)$ are symmetric, continuous, and elliptic on $\mathbb{V}$. Moreover, let $\|\!|\cdot|\!\|$ (resp., $\|\!|\cdot|\!\|_0$) denote the norm induced by $B(\cdot, \cdot)$ (resp., $B_0(\cdot, \cdot)$). Then, there holds

$$
\lambda \|\!|\boldsymbol{v}|\!\|_0^2 \le \|\!|\boldsymbol{v}|\!\|^2 \le \Lambda \|\!|\boldsymbol{v}|\!\|_0^2 \quad \text{for all } \boldsymbol{v} \in \mathbb{V},
\tag{7}
$$

where $\lambda := 1 - \tau$ and $\Lambda := 1 + \tau$. Note that $0 < \lambda < 1 < \Lambda < 2$.

The weak formulation of problem (1) reads as follows: Find $\boldsymbol{u} \in \mathbb{V}$ such that

$$
B(\boldsymbol{u}, \boldsymbol{v}) = F(\boldsymbol{v}) := \int_\Gamma \int_D \boldsymbol{f}(x, \mathbf{y}) \boldsymbol{v}(x, \mathbf{y}) \, dx \, d\pi(\mathbf{y}) \quad \text{for all } \boldsymbol{v} \in \mathbb{V}.
\tag{8}
$$

The existence and uniqueness of the solution $\boldsymbol{u} \in \mathbb{V}$ to (8) follow by the Riesz theorem.

3. Multilevel stochastic Galerkin FEM discretization

3.1. Discretization in the physical domain and mesh refinement. Let $\mathcal{T}_\bullet$ be a *mesh*, i.e., a conforming triangulation of $D \subset \mathbb{R}^2$ into compact non-degenerate triangles $T \in \mathcal{T}_\bullet$. Let $\mathcal{N}_\bullet$ be the set of vertices of $\mathcal{T}_\bullet$. For mesh refinement, we employ 2D newest vertex bisection (NVB); see, e.g., [Ste08, KPP13]. We assume that any mesh $\mathcal{T}_\bullet$ employed for the spatial discretization can be obtained by applying NVB refinement(s) to a given initial (coarse) mesh $\mathcal{T}_0$. In particular, we denote by refine($\mathcal{T}_0$) the set of all meshes obtained from $\mathcal{T}_0$ by finitely many steps of refinement.

For a given mesh $\mathcal{T}_\bullet$, let $\widehat{\mathcal{T}}_\bullet$ denote its uniform refinement, where all elements of $\mathcal{T}_\bullet$ are refined by three bisections. Let $\widehat{\mathcal{N}}_\bullet$ be the set of vertices of $\widehat{\mathcal{T}}_\bullet$. Let $\mathcal{N}_\bullet^+ := (\widehat{\mathcal{N}}_\bullet \setminus \mathcal{N}_\bullet) \setminus \partial D$ be the set of new interior vertices created by uniform refinement of $\mathcal{T}_\bullet$. For a set of marked vertices $\mathcal{M}_\bullet \subseteq \mathcal{N}_\bullet^+$, let $\mathcal{T}_\circ := \text{refine}(\mathcal{T}_\bullet, \mathcal{M}_\bullet)$ be the coarsest mesh such that $\mathcal{M}_\bullet \subseteq \mathcal{N}_\circ$, i.e., all marked vertices are vertices of $\mathcal{T}_\circ$. Since NVB is a binary refinement rule, this implies that $\mathcal{N}_\circ \subseteq \widehat{\mathcal{N}}_\bullet$ and $(\mathcal{N}_\circ \setminus \mathcal{N}_\bullet) \setminus \partial D = \mathcal{N}_\bullet^+ \cap \mathcal{N}_\circ$. In particular, the choices $\mathcal{M}_\bullet = \emptyset$ and $\mathcal{M}_\bullet = \mathcal{N}_\bullet^+$ lead to the meshes $\mathcal{T}_\bullet = \text{refine}(\mathcal{T}_\bullet, \emptyset)$ and $\widehat{\mathcal{T}}_\bullet = \text{refine}(\mathcal{T}_\bullet, \mathcal{N}_\bullet^+)$, respectively.

For a given mesh $\mathcal{T}_\bullet \in \text{refine}(\mathcal{T}_0)$, we consider the space of piecewise affine and globally continuous finite elements
$$\mathbb{X}_\bullet := \mathcal{S}_0^1(\mathcal{T}_\bullet) := \{v_\bullet \in \mathbb{X} : v_\bullet|_T \text{ is affine for all } T \in \mathcal{T}_\bullet\} \subset \mathbb{X} = H_0^1(D).$$
For $z \in \mathcal{N}_\bullet$, let $\varphi_{\bullet,z}$ be the associated hat function, i.e., $\varphi_{\bullet,z}$ is piecewise affine, globally continuous, and satisfies the Kronecker property $\varphi_{\bullet,z}(z') = \delta_{zz'}$ for all $z' \in \mathcal{N}_\bullet$. Recall that $\{\varphi_{\bullet,z} : z \in \mathcal{N}_\bullet \setminus \partial D\}$ is the standard basis of $\mathbb{X}_\bullet$.

The finite element space associated with $\widehat{\mathcal{T}}_\bullet$ is denoted by $\widehat{\mathbb{X}}_\bullet := \mathcal{S}_0^1(\widehat{\mathcal{T}}_\bullet)$ and $\{\widehat{\varphi}_{\bullet,z} : z \in \widehat{\mathcal{N}}_\bullet \setminus \partial D\}$ is the corresponding basis of hat functions. Later, we shall exploit the (H^1-stable) two-level decomposition $\widehat{\mathbb{X}}_\bullet = \mathbb{X}_\bullet \oplus \text{span}\{\widehat{\varphi}_{\bullet,z} : z \in \mathcal{N}_\bullet^+\}$.

3.2. Discretization in the parameter domain and parametric enrichment. For all $m \in \mathbb{N}$, we denote by $(P_n^m)_{n \in \mathbb{N}_0}$ the sequence of univariate polynomials which are orthogonal with respect to π_m such that P_n^m is a polynomial of degree $n \in \mathbb{N}_0$ with $\|P_n^m\|_{L^2_{\pi_m}(-1,1)} = 1$ and $P_0^m \equiv 1$. It is well-known that $\{P_n^m : n \in \mathbb{N}_0\}$ is an orthonormal basis of $L^2_{\pi_m}(-1,1)$. With $\mathbb{N}_0^{\mathbb{N}} := \{\nu = (\nu_m)_{m \in \mathbb{N}} : \nu_m \in \mathbb{N}_0 \text{ for all } m \in \mathbb{N}\}$ and $\text{supp}(\nu) := \{m \in \mathbb{N} : \nu_m \neq 0\}$, let $\mathfrak{I} := \{\nu \in \mathbb{N}_0^{\mathbb{N}} : \# \text{supp}(\nu) < \infty\}$ be the set of all finitely supported multi-indices. Note that $\mathfrak{I}$ is countable. With
$$P_\nu(\mathbf{y}) := \prod_{m \in \mathbb{N}} P_{\nu_m}^m(y_m) = \prod_{m \in \text{supp}(\nu)} P_{\nu_m}^m(y_m) \quad \text{for all } \nu \in \mathfrak{I} \text{ and all } \mathbf{y} \in \Gamma,$$
the set $\{P_\nu : \nu \in \mathfrak{I}\}$ is an orthonormal basis of $L^2_\pi(\Gamma)$; see [SG11, Theorem 2.12].

The Bochner space $\mathbb{V} = L^2_\pi(\Gamma; \mathbb{X})$ is isometrically isomorphic to $\mathbb{X} \otimes L^2_\pi(\Gamma)$ and each function $\boldsymbol{v} \in \mathbb{V}$ can be represented in the form
$$\boldsymbol{v}(x, \mathbf{y}) = \sum_{\nu \in \mathfrak{I}} v_\nu(x) P_\nu(\mathbf{y}) \quad \text{with unique coefficients } v_\nu \in \mathbb{X}. \tag{9}$$

Parametric discretization is based on a finite index set $\mathfrak{P}_\bullet \subset \mathfrak{I}$ and the associated subspace $\text{span}\{P_\nu : \nu \in \mathfrak{P}_\bullet\} \subset L^2_\pi(\Gamma)$. We denote by $\text{supp}(\mathfrak{P}_\bullet) := \bigcup_{\nu \in \mathfrak{P}_\bullet} \text{supp}(\nu)$ the

4

set of active parameters in $\mathfrak{P}_\bullet$. We assume that any finite index set $\mathfrak{P}_\bullet$ employed for parametric discretization contains the zero index $\mathbf{0} = (0, 0, \dots)$; in particular, we set $\mathfrak{P}_0 := \{\mathbf{0}\} \subseteq \mathfrak{P}_\bullet$.

For parametric enrichment, we follow [BS16] and consider the *detail index set*

$$\mathfrak{Q}_\bullet := \{\mu \in \mathfrak{I} \setminus \mathfrak{P}_\bullet : \mu = \nu \pm \varepsilon_m \text{ for all } \nu \in \mathfrak{P}_\bullet \text{ and all } m = 1, \dots, M_{\mathfrak{P}_\bullet} + 1\}, \quad (10)$$

where $M_{\mathfrak{P}_\bullet} := \# \operatorname{supp}(\mathfrak{P}_\bullet) \in \mathbb{N}_0$ is the number of active parameters in the index set $\mathfrak{P}_\bullet$ and, for any $m \in \mathbb{N}$, $\varepsilon_m \in \mathfrak{I}$ denotes the m-th unit sequence, i.e., $(\varepsilon_m)_i = \delta_{mi}$ for all $i \in \mathbb{N}$. Then, a parametric enrichment is obtained by adding some marked indices $\mathfrak{M}_\bullet \subseteq \mathfrak{Q}_\bullet$ to the current index set $\mathfrak{P}_\bullet$, i.e., $\mathfrak{P}_\circ := \mathfrak{P}_\bullet \cup \mathfrak{M}_\bullet$. We note that $\mathfrak{P}_\bullet \subseteq \mathfrak{P}_\circ \subseteq \mathfrak{P}_\bullet \cup \mathfrak{Q}_\bullet$ and at least one of these inclusions is strict.

3.3. Multilevel approximation and multilevel refinement. Let us consider a discrete structure $\mathbb{P}_\bullet = [\mathfrak{P}_\bullet, (\mathcal{T}_{\bullet\nu})_{\nu \in \mathfrak{I}}]$ that consists of a finite index set $\mathfrak{P}_\bullet \subset \mathfrak{I}$ and a family of spatial meshes $(\mathcal{T}_{\bullet\nu})_{\nu \in \mathfrak{I}}$, where $\mathcal{T}_{\bullet\nu} \in \operatorname{refine}(\mathcal{T}_0)$ for all $\nu \in \mathfrak{P}_\bullet$, while $\mathcal{T}_{\bullet\nu} = \mathcal{T}_0$ for all $\nu \in \mathfrak{I} \backslash \mathfrak{P}_\bullet$. We will call $\mathbb{P}_\bullet$ a *multilevel structure*. In particular, let $\mathbb{P}_0 := [\mathfrak{P}_0, (\mathcal{T}_{0\nu})_{\nu \in \mathfrak{I}}]$ with $\mathcal{T}_{0\nu} = \mathcal{T}_0$ for all $\nu \in \mathfrak{I}$, be the initial multilevel structure.

For two multilevel structures $\mathbb{P}_\bullet = [\mathfrak{P}_\bullet, (\mathcal{T}_{\bullet\nu})_{\nu \in \mathfrak{I}}]$ and $\mathbb{P}_\circ = [\mathfrak{P}_\circ, (\mathcal{T}_{\circ\nu})_{\nu \in \mathfrak{I}}]$, we write $\mathbb{P}_\circ = \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$ if $\mathbb{P}_\circ$ is obtained from $\mathbb{P}_\bullet$ using *one step of multilevel refinement* defined as follows:

- $\mathbf{M}_\bullet = [\mathfrak{M}_\bullet, (\mathcal{M}_{\bullet\nu})_{\nu \in \mathfrak{P}_\bullet}]$ with $\mathfrak{M}_\bullet \subseteq \mathfrak{Q}_\bullet$ and $\mathcal{M}_{\bullet\nu} \subseteq \mathcal{N}_{\bullet\nu}^+$ for all $\nu \in \mathfrak{P}_\bullet$;
- $\mathfrak{P}_\circ = \mathfrak{P}_\bullet \cup \mathfrak{M}_\bullet$;
- for all $\nu \in \mathfrak{P}_\bullet$, there holds $\mathcal{T}_{\circ\nu} = \operatorname{refine}(\mathcal{T}_{\bullet\nu}, \mathcal{M}_{\bullet\nu})$;
- for all $\nu \in \mathfrak{I} \backslash \mathfrak{P}_\bullet$, there holds $\mathcal{T}_{\circ\nu} = \mathcal{T}_{\bullet\nu} = \mathcal{T}_0$.

In particular, we denote by $\mathbf{REFINE}(\mathbb{P}_0)$ the set of all multilevel structures obtained from $\mathbb{P}_0$ by finitely many steps of multilevel refinement. Throughout the remainder of this work, we implicitly assume that all occurring multilevel structures belong to $\mathbf{REFINE}(\mathbb{P}_0)$.

With each multi-level structure $\mathbb{P}_\bullet = [\mathfrak{P}_\bullet, (\mathcal{T}_{\bullet\nu})_{\nu \in \mathfrak{I}}]$, we associate the finite dimensional subspace

$$\mathbb{V}_\bullet := \bigoplus_{\nu \in \mathfrak{P}_\bullet} \mathbb{V}_{\bullet\nu} \subset \mathbb{V} \quad \text{with} \quad \mathbb{V}_{\bullet\nu} := \mathbb{X}_{\bullet\nu} \otimes \operatorname{span}\{P_\nu\} = \operatorname{span}\{\varphi_{\bullet\nu,z} P_\nu : z \in \mathcal{N}_{\bullet\nu}\}, \quad (11)$$

where $\mathbb{X}_{\bullet\nu} = \mathcal{S}_0^1(\mathcal{T}_{\bullet\nu})$ for all $\nu \in \mathfrak{P}_\bullet$. We note that the sum of the spaces $\mathbb{V}_{\bullet\nu}$ is orthogonal (and hence direct) and that each function $\boldsymbol{v}_\bullet \in \mathbb{V}_\bullet$ can be represented in the form (cf. (9))

$$\boldsymbol{v}_\bullet(x, \mathbf{y}) = \sum_{\nu \in \mathfrak{P}_\bullet} v_{\bullet\nu}(x) P_\nu(\mathbf{y}) \quad \text{with unique coefficients } v_{\bullet\nu} \in \mathbb{X}_{\bullet\nu}.$$

The Galerkin discretization of (8) reads as follows: Find $\boldsymbol{u}_\bullet \in \mathbb{V}_\bullet$ such that

$$B(\boldsymbol{u}_\bullet, \boldsymbol{v}_\bullet) = F(\boldsymbol{v}_\bullet) \quad \text{for all } \boldsymbol{v}_\bullet \in \mathbb{V}_\bullet. \quad (12)$$

As in the continuous case, the Riesz theorem proves the existence and uniqueness of the solution $\boldsymbol{u}_\bullet \in \mathbb{V}_\bullet$. Moreover, there holds the Céa lemma

$$\|\|\boldsymbol{u} - \boldsymbol{u}_\bullet\|\| = \min_{\boldsymbol{v}_\bullet \in \mathbb{V}_\bullet} \|\|\boldsymbol{u} - \boldsymbol{v}_\bullet\|\|. \quad (13)$$

We stress that multilevel refinement $\mathbb{P}_\circ \in \mathbf{REFINE}(\mathbb{P}_\bullet)$ implies nestedness of the associated multilevel spaces $\mathbb{V}_\bullet \subseteq \mathbb{V}_\circ$ and hence, in particular, $\|\|\boldsymbol{u} - \boldsymbol{u}_\circ\|\| \leq \|\|\boldsymbol{u} - \boldsymbol{u}_\bullet\|\|$.

3.4. Saturation assumption. For *a posteriori* error estimation, we follow our approach in [BPR20]: For a given multilevel structure $\mathbb{P}_\bullet$ and the associated finite-dimensional subspace $\mathbb{V}_\bullet$, we consider the enriched subspace $\widehat{\mathbb{V}}_\bullet \subset \mathbb{V}$ defined as

$$\widehat{\mathbb{V}}_\bullet := \bigoplus_{\nu \in \mathfrak{P}_\bullet} \left[\widehat{\mathbb{X}}_{\bullet\nu} \otimes \mathrm{span}\{P_\nu\} \right] \oplus \bigoplus_{\nu \in \mathfrak{Q}_\bullet} \left[\mathbb{X}_0 \otimes \mathrm{span}\{P_\nu\} \right], \tag{14}$$

where we recall that $\mathcal{T}_{\bullet\nu} = \mathcal{T}_0$ for all $\nu \in \mathfrak{Q}_\bullet \subset \mathfrak{I}\backslash\mathfrak{P}_\bullet$. Note that $\mathbb{V}_\bullet \subseteq \mathbb{V}_\circ \subseteq \widehat{\mathbb{V}}_\bullet$ for any $\mathbb{P}_\circ = \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$. Moreover, $\widehat{\mathbb{V}}_\bullet$ corresponds to the multilevel structure $\widehat{\mathbb{P}}_\bullet = \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$ with $\mathbf{M}_\bullet = [\mathfrak{Q}_\bullet, (\mathcal{N}_{\bullet\nu}^+)_{\nu\in\mathfrak{P}_\bullet}]$.

Let $\widehat{\boldsymbol{u}}_\bullet \in \widehat{\mathbb{V}}_\bullet$ be the unique Galerkin solution to

$$B(\widehat{\boldsymbol{u}}_\bullet, \widehat{\boldsymbol{v}}_\bullet) = F(\widehat{\boldsymbol{v}}_\bullet) \quad \text{for all } \widehat{\boldsymbol{v}}_\bullet \in \widehat{\mathbb{V}}_\bullet. \tag{15}$$

Existence and uniqueness of the solution $\widehat{\boldsymbol{u}}_\bullet \in \widehat{\mathbb{V}}_\bullet$ follow from the Riesz theorem. We stress, however, that $\widehat{\boldsymbol{u}}_\bullet \in \widehat{\mathbb{V}}_\bullet$ is only needed for the theoretical analysis and will not be computed throughout. We suppose that there exists a uniform constant $0 < q_{\mathrm{sat}} < 1$ such that the following saturation assumption holds:

$$\|\, \boldsymbol{u} - \widehat{\boldsymbol{u}}_\bullet \,\| \leq q_{\mathrm{sat}} \,\|\, \boldsymbol{u} - \boldsymbol{u}_\bullet \,\|. \tag{16}$$

Remark 1. *While the saturation assumption can be verified empirically as soon as approximations exhibit some asymptotic behavior, the existing rigorous proofs in the context of FEM are tailored to deterministic problems with constant coefficients (see, e.g., [DN02] or [CGG16]). When required for generic discrete spaces $\mathbb{V}_\bullet$, the saturation assumption (16) is a strong restriction, which may even fail in general (see [BEK96] for a counterexample in the deterministic setting). In our analysis, however, it will be required only for the sequence of nested discrete subspaces $(\mathbb{V}_\ell)_{\ell\in\mathbb{N}_0}$ generated by Algorithm 3 below.*

3.5. *A posteriori* error estimation. To abbreviate notation, we denote the inner product on $\mathbb{X} = H_0^1(D)$ by $\langle w\,,\, v\rangle_D := \int_D a_0 \nabla w \cdot \nabla v\, \mathrm{d}x$ and the induced energy norm by $\|\cdot\|_D := \|a_0^{1/2}\nabla(\cdot)\|_{L^2(D)}$.

The *parametric* error is estimated by means of hierarchical error indicators

$$\tau_\bullet(\nu) := \|e_{\bullet\nu}\|_D \quad \text{for all } \nu \in \mathfrak{Q}_\bullet, \tag{17a}$$

where $e_{\bullet\nu} \in \mathbb{X}_0$ is the unique solution of

$$\langle e_{\bullet\nu}\,,\, v_0\rangle_D = F(v_0 P_\nu) - B(\boldsymbol{u}_\bullet, v_0 P_\nu) \quad \text{for all } v_0 \in \mathbb{X}_0. \tag{17b}$$

In order to estimate the errors due to *spatial* discretizations, we employ the two-level error estimation strategy. Specifically, we define the two-level error indicators

$$\tau_\bullet(\nu, z) := \frac{|F(\widehat{\varphi}_{\bullet\nu,z} P_\nu) - B(\boldsymbol{u}_\bullet, \widehat{\varphi}_{\bullet\nu,z} P_\nu)|}{\|\widehat{\varphi}_{\bullet\nu,z}\|_D} \quad \text{for all } \nu \in \mathfrak{P}_\bullet \text{ and all } z \in \mathcal{N}_{\bullet\nu}^+. \tag{18}$$

Overall, we thus consider the computable *a posteriori* error estimate

$$\tau_\bullet := \left(\sum_{\nu\in\mathfrak{P}_\bullet} \sum_{z\in\mathcal{N}_{\bullet\nu}^+} \tau_\bullet(\nu, z)^2 + \sum_{\nu\in\mathfrak{Q}_\bullet} \tau_\bullet(\nu)^2 \right)^{1/2}. \tag{19}$$

We recall the following main result from [BPR20], where we note that the validity of (20) hinges on 2D newest vertex bisection (since the hat functions satisfy $\widehat{\varphi}_{\bullet\nu,z} = \varphi_{\circ\nu,z} \in \mathbb{X}_{\circ\nu}$

6

for all $\nu \in \mathfrak{P}_\bullet$ if $\mathcal{T}_{\circ\nu} = \text{refine}(\mathcal{T}_{\bullet\nu}, \mathcal{M}_{\bullet\nu})$ and $z \in \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\circ\nu}$); we refer to [BPR20, Theorem 2 and Remark 6].

Theorem 2. *Let* $\mathbb{P}_\bullet \in \textbf{\textsc{refine}}(\mathbb{P}_0)$ *and* $\mathbb{P}_\circ \in \textbf{\textsc{refine}}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$, *where* $\mathbf{M}_\bullet = [\mathfrak{M}_\bullet, (\mathcal{M}_{\bullet\nu})_{\nu \in \mathfrak{P}_\bullet}]$ *satisfies* $\mathfrak{M}_\bullet \subseteq \mathfrak{Q}_\bullet$ *and* $\mathcal{M}_{\bullet\nu} \subseteq \mathcal{N}_{\bullet\nu}^+$ *for all* $\nu \in \mathfrak{P}_\bullet$. *Let* $\mathbb{V}_\bullet \subseteq \mathbb{V}_\circ \subseteq \widehat{\mathbb{V}}_\bullet$ *be the corresponding multilevel spaces with associated Galerkin solutions* $\boldsymbol{u}_\bullet \in \mathbb{V}_\bullet$ *(solving* (12)*)*, $\boldsymbol{u}_\circ \in \mathbb{V}_\circ$ *(solving* (12) *with* $\mathbb{V}_\bullet$ *being replaced by* $\mathbb{V}_\circ$*), and* $\widehat{\boldsymbol{u}}_\bullet \in \widehat{\mathbb{V}}_\bullet$ *(solving* (15)*). Then, there holds*

$$C_{\text{est}}^{-1} \left\| \boldsymbol{u}_\circ - \boldsymbol{u}_\bullet \right\| \leq \left(\sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\circ\nu}} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{Q}_\bullet \cap \mathfrak{P}_\circ} \tau_\bullet(\nu)^2 \right)^{1/2} \leq C_{\text{est}} \left\| \boldsymbol{u}_\circ - \boldsymbol{u}_\bullet \right\|. \quad (20)$$

In particular, this also guarantees that

$$C_{\text{est}}^{-1} \left\| \widehat{\boldsymbol{u}}_\bullet - \boldsymbol{u}_\bullet \right\| \leq \tau_\bullet \overset{(19)}{=} \left(\sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{N}_{\bullet\nu}^+} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{Q}_\bullet} \tau_\bullet(\nu)^2 \right)^{1/2} \leq C_{\text{est}} \left\| \widehat{\boldsymbol{u}}_\bullet - \boldsymbol{u}_\bullet \right\|. \quad (21)$$

Furthermore, let $\boldsymbol{u} \in \mathbb{V}$ *be the solution to problem* (8)*. Then, under the saturation assumption* (16)*, the estimates* (21) *are equivalent to*

$$\frac{(1 - q_{\text{sat}}^2)^{1/2}}{C_{\text{est}}} \left\| \boldsymbol{u} - \boldsymbol{u}_\bullet \right\| \leq \tau_\bullet \leq C_{\text{est}} \left\| \boldsymbol{u} - \boldsymbol{u}_\bullet \right\|, \quad (22)$$

i.e., the proposed error estimator is reliable (under the saturation assumption (16)*) and (always) efficient. The constant* $C_{\text{est}} \geq 1$ *in* (20)–(22) *is generic and depends only on* $\mathcal{T}_0$, *the mean field* a_0, *and the constants* $\lambda, \Lambda > 0$ *from* (7).

4. Adaptive algorithm and main result

The main results of this work concern the following algorithm proposed in [BPR20] (see Algorithm 7.C therein). In this algorithm, the enhancement of the approximation space $\mathbb{V}_\ell$ for each $\ell \in \mathbb{N}_0$ is steered by the Dörfler marking performed on the joint set of all spatial and parametric error indicators (see step (iv) in the algorithm below).

Algorithm 3. *Input:* $\mathfrak{P}_0 = \{\mathbf{0}\}$ *and* $\mathcal{T}_{0\nu} := \mathcal{T}_0$ *for all* $\nu \in \mathfrak{P}_0 \cup \mathfrak{Q}_0$, *as well as marking parameter* $0 < \theta \leq 1$.
Loop: For all $\ell = 0, 1, 2, \ldots,$ *iterate the following steps:*

    (i) *Compute the discrete solution* $\boldsymbol{u}_\ell \in \mathbb{V}_\ell$ *associated with* $\mathbb{P}_\ell = [\mathfrak{P}_\ell, (\mathcal{T}_{\ell\nu})_{\nu \in \mathfrak{I}}]$.
    (ii) *Compute parametric error indicators* $\tau_\ell(\nu)$ *from* (17) *for all* $\nu \in \mathfrak{Q}_\ell$.
    (iii) *Compute spatial error indicators* $\tau_\ell(\nu, z)$ *from* (18) *for all* $\nu \in \mathfrak{P}_\ell$ *and all* $z \in \mathcal{N}_{\ell\nu}^+$.
    (iv) *Determine the sets* $\mathcal{M}_{\ell\nu} \subseteq \mathcal{N}_{\ell\nu}^+$ *for all* $\nu \in \mathfrak{P}_\ell$ *and the set* $\mathfrak{M}_\ell \subseteq \mathfrak{Q}_\ell$ *such that*

$$\theta \left(\sum_{\nu \in \mathfrak{P}_\ell} \sum_{z \in \mathcal{N}_{\ell\nu}^+} \tau_\ell(\nu, z)^2 + \sum_{\nu \in \mathfrak{Q}_\ell} \tau_\ell(\nu)^2 \right) \leq \sum_{\nu \in \mathfrak{P}_\ell} \sum_{z \in \mathcal{M}_{\ell\nu}} \tau_\ell(\nu, z)^2 + \sum_{\nu \in \mathfrak{M}_\ell} \tau_\ell(\nu)^2, \quad (23)$$

    where the overall cardinality $\#\mathfrak{M}_\ell + \sum_{\nu \in \mathfrak{P}_\ell} \#\mathcal{M}_{\ell\nu}$ *is minimal amongst all tuples* $\mathbf{M}_\ell = [\mathfrak{M}_\ell, (\mathcal{M}_{\ell\nu})_{\nu \in \mathfrak{P}_\ell}]$ *satisfying the marking criterion* (23).
    (v) *For all* $\nu \in \mathfrak{P}_\ell$, *let* $\mathcal{T}_{\ell+1,\nu} := \text{refine}(\mathcal{T}_{\ell\nu}, \mathcal{M}_{\ell\nu})$.
    (vi) *Define* $\mathfrak{P}_{\ell+1} := \mathfrak{P}_\ell \cup \mathfrak{M}_\ell$ *and* $\mathcal{T}_{(\ell+1)\nu} := \mathcal{T}_0$ *for all* $\nu \in \mathfrak{Q}_{\ell+1}$.

Output: *For all $\ell \in \mathbb{N}_0$, the algorithm returns the multilevel stochastic Galerkin approximation $\boldsymbol{u}_\ell \in \mathbb{V}_\ell$ as well as the corresponding error estimate τ_ℓ.*

While linear convergence of the adaptive algorithm will rely only on the above saturation assumption (16), the proof of optimal convergence rates requires the following *strong saturation assumption* [PRS20]: There exist constants $0 < \kappa_{\mathrm{sat}} \leq q_{\mathrm{sat}} < 1$ such that for all multilevel structures $\mathbb{P}_\bullet \in \mathbf{REFINE}(\mathbb{P}_0)$ and $\mathbb{P}_\star \in \mathbf{REFINE}(\mathbb{P}_\bullet)$, one step of multilevel refinement $\mathbb{P}_\circ := \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$ with $\mathbf{M}_\bullet := [\mathfrak{P}_\star \cap \mathfrak{Q}_\bullet, (\mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu})_{\nu \in \mathfrak{P}_\bullet}]$ satisfies the following implication:

$$\| \boldsymbol{u} - \boldsymbol{u}_\star \| \leq \kappa_{\mathrm{sat}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \| \quad \Longrightarrow \quad \| \boldsymbol{u} - \boldsymbol{u}_\circ \| \leq q_{\mathrm{sat}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|. \tag{24}$$

In explicit terms, the strong saturation assumption (24) states that, if $\mathbb{P}_\star \in \mathbf{REFINE}(\mathbb{P}_\bullet)$ leads to a sufficient improvement of the error, then already one step of multilevel refinement of $\mathbb{P}_\bullet$ towards $\mathbb{P}_\star$ provides a uniform improvement of the error.

Remark 4. *We note that the strong saturation assumption (24) is, in fact, stronger than the saturation assumption (16). To see this, let us suppose that (24) is satisfied. Since convergence of stochastic Galerkin FEM is known, there exists $\mathbb{P}_\star \in \mathbf{REFINE}(\mathbb{P}_\bullet)$ with $\| \boldsymbol{u} - \boldsymbol{u}_\star \| \leq \kappa_{\mathrm{sat}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|$. Let $\widehat{\mathbb{P}}_\star = \mathbf{REFINE}(\mathbb{P}_\star, \mathbf{M}_\star)$, where $\mathbf{M}_\star = [\mathfrak{Q}_\star, (\mathcal{N}_{\star\nu}^+)_{\nu \in \mathfrak{P}_\star}]$. According to the Céa lemma (13), there holds*

$$\| \boldsymbol{u} - \widehat{\boldsymbol{u}}_\star \| \leq \| \boldsymbol{u} - \boldsymbol{u}_\star \| \leq \kappa_{\mathrm{sat}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|.$$

According to (24) (now applied to $\widehat{\mathbb{P}}_\star$ and $\mathbb{P}_\bullet$), the multilevel structure $\widetilde{\mathbb{P}}_\circ := \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$ with $\mathbf{M}_\bullet := [\widehat{\mathfrak{P}}_\star \cap \mathfrak{Q}_\bullet, (\widehat{\mathcal{N}}_{\star\nu} \cap \mathcal{N}_{\bullet\nu}^+)_{\nu \in \mathfrak{P}_\bullet}]$ thus satisfies

$$\| \boldsymbol{u} - \widetilde{\boldsymbol{u}}_\circ \| \leq q_{\mathrm{sat}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|.$$

Since $\widehat{\mathfrak{P}}_\star \cap \mathfrak{Q}_\bullet = \mathfrak{Q}_\bullet$ and $\widehat{\mathcal{N}}_{\star\nu} \cap \mathcal{N}_{\bullet\nu}^+ = \mathcal{N}_{\bullet\nu}^+$ for all $\nu \in \mathfrak{P}_\bullet$, we conclude that $\widetilde{\mathbb{P}}_\circ = \widehat{\mathbb{P}}_\bullet$. Therefore, the last estimate is precisely (16).

Similarly to Remark 1 for the saturation assumption (16), we point out that the proof of our main result in Theorem 5 will exploit the strong saturation assumption (24) with $\mathbb{V}_\bullet = \mathbb{V}_\ell$ for all $\ell \in \mathbb{N}_0$, where $(\mathbb{V}_\ell)_{\ell \in \mathbb{N}_0}$ is the sequence of nested discrete subspaces generated by Algorithm 3.

For a multilevel structure $\mathbb{P}_\bullet = [\mathfrak{P}_\bullet, (\mathcal{T}_{\bullet\nu})_{\nu \in \mathfrak{I}}]$, let

$$\#\mathbb{P}_\bullet := \sum_{\nu \in \mathfrak{P}_\bullet} \big[\#\mathcal{T}_{\bullet\nu} - \#\mathcal{T}_0 + 1 \big]. \tag{25}$$

Note that this definition is motivated by the equivalence

$$\#\mathbb{P}_\bullet = \sum_{\nu \in \mathfrak{P}_\bullet} \big[\#\mathcal{T}_{\bullet\nu} - \#\mathcal{T}_0 + 1 \big] \simeq \sum_{\nu \in \mathfrak{P}_\bullet} \#\mathcal{T}_{\bullet\nu} \simeq \sum_{\nu \in \mathfrak{P}_\bullet} \dim \mathcal{S}_0^1(\mathcal{T}_{\bullet\nu}) = \dim \mathbb{V}_\bullet, \tag{26}$$

i.e., up to some multiplicative constants (depending only on $\#\mathcal{T}_0$), $\#\mathbb{P}_\bullet$ is equivalent to the overall dimension of the corresponding multilevel finite element space $\mathbb{V}_\bullet$ (in fact, the last equivalence in (26) holds only for non-pathological meshes $\mathcal{T}_0$ that have at least one interior vertex).

For $s > 0$, we can now introduce the following notion of approximability:

$$\| \boldsymbol{u} \|_{\mathbb{A}_s} := \sup_{N \in \mathbb{N}_0} (N+1)^s \min_{\substack{\mathbb{P}_{\mathrm{opt}} \in \mathbf{REFINE}(\mathbb{P}_0) \\ \#\mathbb{P}_{\mathrm{opt}} - \#\mathbb{P}_0 \leq N}} \min_{\boldsymbol{v}_{\mathrm{opt}} \in \mathbb{V}_{\mathrm{opt}}} \| \boldsymbol{u} - \boldsymbol{v}_{\mathrm{opt}} \| \in \mathbb{R}_{\geq 0} \cup \{\infty\}. \tag{27}$$

Recalling the equivalence (26), the definition of $\|\boldsymbol{u}\|_{\mathbb{A}_s}$ in (27) is understood as follows: There holds $\|\boldsymbol{u}\|_{\mathbb{A}_s} < \infty$ if and only if there exists a sequence of multilevel structures $(\mathbb{P}_\ell^\star)_{\ell \in \mathbb{N}_0}$ with $\mathbb{P}_0^\star = \mathbb{P}_0$ (but not necessarily nested multilevel spaces $\mathbb{V}_\ell^\star$) such that the corresponding error $\|\boldsymbol{u} - \boldsymbol{u}_\ell^\star\| = \min_{\boldsymbol{v}_\ell \in \mathbb{V}_\ell^\star} \|\boldsymbol{u} - \boldsymbol{v}_\ell\|$ decays at least with an algebraic rate $s > 0$ with respect to the dimensions of the corresponding multilevel spaces $\mathbb{V}_\ell^\star$.

The following theorem is the main result of this work. It shows that Algorithm 3 is linearly convergent under the saturation assumption (16) and even rate-optimal under the strong saturation assumption (24), i.e., the energy errors decay with any possible algebraic rate $s > 0$.

Theorem 5. *Let $C_{\mathrm{est}} \geq 1$ be the constant from Theorem 2. Under the saturation assumption (16) and for each $0 < \theta \leq 1$, Algorithm 3 leads to linear convergence in the sense that*

$$\| \boldsymbol{u} - \boldsymbol{u}_{\ell+n} \| \leq q_{\mathrm{lin}}^n \| \boldsymbol{u} - \boldsymbol{u}_\ell \| \quad \text{for all } \ell, n \in \mathbb{N}_0, \quad \text{where } 0 < q_{\mathrm{lin}}^2 = 1 - \frac{1 - q_{\mathrm{sat}}^2}{C_{\mathrm{est}}^4} \theta < 1. \quad (28)$$

Furthermore, under the strong saturation assumption (24), there exists a constant $0 < \theta_{\mathrm{opt}} < 1$ depending only on C_{est} and q_{sat} such that the following holds whenever $0 < \theta \leq \theta_{\mathrm{opt}}$ is satisfied: If $s > 0$ and $\|\boldsymbol{u}\|_{\mathbb{A}_s} < \infty$, then

$$\sup_{\ell \in \mathbb{N}_0} (\#\mathbb{P}_\ell - \#\mathbb{P}_0 + 1)^s \| \boldsymbol{u} - \boldsymbol{u}_\ell \| \leq C_{\mathrm{opt}} \|\boldsymbol{u}\|_{\mathbb{A}_s}, \quad (29)$$

where $C_{\mathrm{opt}} \geq 1$ depends only on s, $\mathcal{T}_0$, κ_{sat}, and q_{lin}.

The proof of Theorem 5 is postponed to the next two sections.

5. Proof of linear convergence

The following proposition shows that, under the saturation assumption (16), the combined Dörfler marking (23) (or, (30) with $C_{\mathrm{lin}} = 1/\theta$) leads (and is essentially equivalent) to linear error reduction. In particular, the result emphasizes the fact that the combined Dörfler marking (23) is the weakest marking criterion which ensures linear convergence.

Proposition 6. *Let $\mathbb{P}_\bullet := [\mathfrak{P}_\bullet, (\mathcal{T}_{\bullet\nu})_{\nu \in \mathfrak{I}}]$ be a multilevel structure with the associated Galerkin solution $\boldsymbol{u}_\bullet \in \mathbb{V}_\bullet$.*
(i) Suppose that the saturation assumption (16) holds and there exists $C_{\mathrm{lin}} > 1$ as well as subsets $\mathfrak{M}_\bullet \subseteq \mathfrak{Q}_\bullet$ and $\mathcal{M}_{\bullet\nu} \subseteq \mathcal{N}_{\bullet\nu}^+$ for all $\nu \in \mathfrak{P}_\bullet$ such that

$$\tau_\bullet^2 = \sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{N}_{\bullet\nu}^+} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{Q}_\bullet} \tau_\bullet(\nu)^2 \leq C_{\mathrm{lin}} \left(\sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{M}_{\bullet\nu}} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{M}_\bullet} \tau_\bullet(\nu)^2 \right). \quad (30)$$

Let $\mathbb{P}_\circ = \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$, where $\mathbf{M}_\bullet = [\mathfrak{M}_\bullet, (\mathcal{M}_{\bullet\nu})_{\nu \in \mathfrak{I}}]$ and $\mathcal{M}_{\bullet\nu} := \emptyset$ for all $\nu \in \mathfrak{I}\backslash\mathfrak{P}_\bullet$. Then, there holds linear error reduction

$$\| \boldsymbol{u} - \boldsymbol{u}_\circ \| \leq q_{\mathrm{lin}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|, \quad (31)$$

where $0 < q_{\mathrm{lin}}^2 = 1 - (1 - q_{\mathrm{sat}}^2)/(C_{\mathrm{est}}^4 C_{\mathrm{lin}}) < 1$ and $\boldsymbol{u}_\circ \in \mathbb{V}_\circ$ is the Galerkin solution associated with $\mathbb{P}_\circ$.
(ii) Conversely, let $\mathbb{P}_\circ \in \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$ be a multilevel structure obtained from $\mathbb{P}_\bullet$ using one step of multilevel refinement and suppose that the associated Galerkin solution $u_\circ \in \mathbb{V}_\circ$ satisfies (31) with arbitrary $0 < q_{\mathrm{lin}} < 1$. Then, there holds (30) with $C_{\mathrm{lin}} = \frac{C_{\mathrm{est}}^4}{1 - q_{\mathrm{lin}}^2}$,

9

$\mathfrak{M}_\bullet = \mathfrak{Q}_\bullet \cap \mathfrak{P}_\circ$, and $\mathcal{M}_{\bullet\nu} = \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\circ\nu}$ for all $\nu \in \mathfrak{P}_\bullet$ (even without appealing to the saturation assumption (16)).

Proof. According to (20) and (22), there holds

$$\| \boldsymbol{u}_\circ - \boldsymbol{u}_\bullet \|^2 \overset{(20)}{\geq} C_{\mathrm{est}}^{-2} \left(\sum_{\nu\in\mathfrak{P}_\bullet} \sum_{z\in\mathcal{N}_{\bullet\nu}^+\cap\mathcal{N}_{\circ\nu}} \tau_\bullet(\nu,z)^2 + \sum_{\nu\in\mathfrak{Q}_\bullet\cap\mathfrak{P}_\circ} \tau_\bullet(\nu)^2 \right)$$

$$\geq C_{\mathrm{est}}^{-2} \left(\sum_{\nu\in\mathfrak{P}_\bullet} \sum_{z\in\mathcal{M}_{\bullet\nu}} \tau_\bullet(\nu,z)^2 + \sum_{\nu\in\mathfrak{M}_\bullet} \tau_\bullet(\nu)^2 \right)$$

$$\overset{(30)}{\geq} C_{\mathrm{est}}^{-2} C_{\mathrm{lin}}^{-1} \tau_\bullet^2 \overset{(22)}{\geq} \frac{1-q_{\mathrm{sat}}^2}{C_{\mathrm{est}}^4 C_{\mathrm{lin}}} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|^2.$$

With the Galerkin orthogonality for $\mathbb{V}_\bullet \subseteq \mathbb{V}_\circ \subset \mathbb{V}$, it follows that

$$\| \boldsymbol{u} - \boldsymbol{u}_\circ \|^2 = \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|^2 - \| \boldsymbol{u}_\circ - \boldsymbol{u}_\bullet \|^2 \leq \left(1 - \frac{1-q_{\mathrm{sat}}^2}{C_{\mathrm{est}}^4 C_{\mathrm{lin}}} \right) \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|^2.$$

Overall, we have seen that (30) implies (31) (under the saturation assumption (16)). This proves part (i).

To see the converse implication in part (ii), note that the estimate

$$\| \boldsymbol{u} - \boldsymbol{u}_\bullet \|^2 - \| \boldsymbol{u}_\circ - \boldsymbol{u}_\bullet \|^2 = \| \boldsymbol{u} - \boldsymbol{u}_\circ \|^2 \overset{(31)}{\leq} q_{\mathrm{lin}}^2 \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|^2$$

yields that

$$C_{\mathrm{est}}^{-2} \tau_\bullet^2 \overset{(22)}{\leq} \| \boldsymbol{u} - \boldsymbol{u}_\bullet \|^2 \leq \frac{1}{1-q_{\mathrm{lin}}^2} \| \boldsymbol{u}_\circ - \boldsymbol{u}_\bullet \|^2$$

$$\overset{(20)}{\leq} \frac{C_{\mathrm{est}}^2}{1-q_{\mathrm{lin}}^2} \left(\sum_{\nu\in\mathfrak{P}_\bullet} \sum_{z\in\mathcal{N}_{\bullet\nu}^+\cap\mathcal{N}_{\circ\nu}} \tau_\bullet(\nu,z)^2 + \sum_{\nu\in\mathfrak{Q}_\bullet\cap\mathfrak{P}_\circ} \tau_\bullet(\nu)^2 \right).$$

This concludes the proof. □

Proposition 6 yields linear convergence of the iterates of Algorithm 3.

Proof of estimate (28) in Theorem 5. The assumption (30) with $C_{\mathrm{lin}} = 1/\theta$ coincides with the combined Dörfler marking in step (iv) of Algorithm 3. Hence, Proposition 6 yields $\| \boldsymbol{u} - \boldsymbol{u}_{\ell+1} \| \leq q_{\mathrm{lin}} \| \boldsymbol{u} - \boldsymbol{u}_\ell \|$, and induction on n proves (28). □

6. Proof of optimal convergence rates

6.1. Fine properties of multilevel structures. In this subsection, we show that the proposed refinement of multilevel structures satisfies the *overlay estimate* as well as the *closure estimate*, both known for spatial meshes refined by NVB.

Lemma 7 (overlay estimate). *For any two multilevel structures $\mathbb{P}_\bullet, \mathbb{P}_\star \in \mathtt{REFINE}(\mathbb{P}_0)$, there exists a unique multilevel structure $\mathbb{P}_\bullet \oplus \mathbb{P}_\star \in \mathtt{REFINE}(\mathbb{P}_\bullet) \cap \mathtt{REFINE}(\mathbb{P}_\star) \subseteq \mathtt{REFINE}(\mathbb{P}_0)$, the so-called overlay, satisfying*

$$\#(\mathbb{P}_\bullet \oplus \mathbb{P}_\star) = \min\left\{ \#\mathbb{P}_\circ : \mathbb{P}_\circ \in \mathtt{REFINE}(\mathbb{P}_\bullet) \cap \mathtt{REFINE}(\mathbb{P}_\star) \right\} \leq \#\mathbb{P}_\bullet + \#\mathbb{P}_\star - \#\mathbb{P}_0. \quad (32)$$

Proof. For two meshes $\mathcal{T}_\bullet, \mathcal{T}_\star \in \text{refine}(\mathcal{T}_0)$ and NVB refinement, let $\mathcal{T}_\bullet \oplus \mathcal{T}_\star \in \text{refine}(\mathcal{T}_\bullet) \cap \text{refine}(\mathcal{T}_\star) \subseteq \text{refine}(\mathcal{T}_0)$ denote the (unique) coarsest common refinement of $\mathcal{T}_\bullet$ and $\mathcal{T}_\star$. Then, there holds the so-called *overlay estimate*

$$\#(\mathcal{T}_\bullet \oplus \mathcal{T}_\star) = \min\left\{ \#\mathcal{T}_\circ : \mathbb{T}_\circ \in \text{refine}(\mathcal{T}_\bullet) \cap \text{refine}(\mathcal{T}_\star) \right\} \leq \#\mathcal{T}_\bullet + \#\mathcal{T}_\star - \#\mathcal{T}_0; \quad (33)$$

see [Ste07, CKNS08].

If $\mathbb{P}_\bullet = [\mathfrak{P}_\bullet, (\mathcal{T}_{\bullet\nu})_{\nu\in\mathfrak{I}}]$ and $\mathbb{P}_\star = [\mathfrak{P}_\star, (\mathcal{T}_{\star\nu})_{\nu\in\mathfrak{I}}]$, define the overlay $\mathbb{P}_\bullet \oplus \mathbb{P}_\star := [\mathfrak{P}_\bullet \cup \mathfrak{P}_\star, (\mathcal{T}_{\bullet\nu} \oplus \mathcal{T}_{\star\nu})_{\nu\in\mathfrak{I}}]$. Then, $\mathbb{P}_\bullet \oplus \mathbb{P}_\star \in \textbf{REFINE}(\mathbb{P}_\bullet) \cap \textbf{REFINE}(\mathbb{P}_\star)$ and, clearly, $\#(\mathbb{P}_\bullet \oplus \mathbb{P}_\star) \leq \#\mathbb{P}_\circ$ for any $\mathbb{P}_\circ \in \textbf{REFINE}(\mathbb{P}_\bullet) \cap \textbf{REFINE}(\mathbb{P}_\star)$. Moreover, there holds

$$\#(\mathbb{P}_\bullet \oplus \mathbb{P}_\star) \overset{(25)}{=} \sum_{\nu\in\mathfrak{P}_\bullet\cup\mathfrak{P}_\star} \left[\#(\mathcal{T}_{\bullet\nu} \oplus \mathcal{T}_{\star\nu}) - \#\mathcal{T}_0 + 1 \right]$$

$$\overset{(33)}{\leq} \sum_{\nu\in\mathfrak{P}_\bullet\cup\mathfrak{P}_\star} \left[(\#\mathcal{T}_{\bullet\nu} - \#\mathcal{T}_0) + (\#\mathcal{T}_{\star\nu} - \#\mathcal{T}_0) + 1 \right]$$

$$= \sum_{\nu\in\mathfrak{P}_\bullet} \left[\#\mathcal{T}_{\bullet\nu} - \#\mathcal{T}_0 + 1 \right] + \sum_{\nu\in\mathfrak{P}_\star} \left[\#\mathcal{T}_{\star\nu} - \#\mathcal{T}_0 + 1 \right] - \sum_{\nu\in\mathfrak{P}_\bullet\cap\mathfrak{P}_\star} 1$$

$$\overset{(25)}{\leq} \#\mathbb{P}_\bullet + \#\mathbb{P}_\star - \#\mathbb{P}_0,$$

since $\mathcal{T}_{\bullet\nu} = \mathcal{T}_0$ for all $\nu \in \mathfrak{I}\backslash\mathfrak{P}_\bullet$ (resp., $\mathcal{T}_{\star\nu} = \mathcal{T}_0$ for all $\nu \in \mathfrak{I}\backslash\mathfrak{P}_\star$). $\qquad\square$

Lemma 8 (closure estimate). *Suppose that $(\mathbb{P}_\ell)_{\ell\in\mathbb{N}_0}$ is a sequence of successively refined multilevel structures, i.e., $\mathbb{P}_{\ell+1} = \textbf{REFINE}(\mathbb{P}_\ell, \mathbf{M}_\ell)$ with $\mathbf{M}_\ell = [\mathfrak{M}_\ell, (\mathcal{M}_{\ell\nu})_{\nu\in\mathfrak{P}_\ell}]$ for all $\ell \in \mathbb{N}_0$. Then, there exists a constant $C_{\text{cls}} \geq 1$ depending only on $\mathcal{T}_0$ such that*

$$\#\mathbb{P}_\ell - \#\mathbb{P}_0 \leq C_{\text{cls}} \sum_{k=0}^{\ell-1} \left(\#\mathfrak{M}_k + \sum_{\nu\in\mathfrak{P}_k} \#\mathcal{M}_{k\nu} \right) \quad \text{for all } \ell \in \mathbb{N}_0. \quad (34)$$

Proof. It is known that for any sequence $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}_0}$ of successively refined meshes (i.e., $\mathcal{T}_{\ell+1} = \text{refine}(\mathcal{T}_\ell, \mathcal{M}_\ell)$ with appropriate $\mathcal{M}_\ell \subseteq \mathcal{N}_\ell^+$ for all $\ell \in \mathbb{N}_0$), NVB refinement guarantees the *closure estimate*

$$\#\mathcal{T}_\ell - \#\mathcal{T}_0 \leq C_{\text{cls}} \sum_{k=0}^{\ell-1} \#\mathcal{M}_k \quad \text{for all } \ell \in \mathbb{N}_0,$$

where $C_{\text{cls}} \geq 1$ depends only on $\mathcal{T}_0$; see [BDD04, Ste08, KPP13].

By definition of multilevel refinement, there holds $\mathcal{T}_{\ell+1,\nu} = \text{refine}(\mathcal{T}_{\ell\nu}, \mathcal{M}_{\ell\nu})$ for all $\ell \in \mathbb{N}_0$ and all $\nu \in \mathfrak{I}$, where $\mathcal{M}_{\ell\nu} = \emptyset$ for $\nu \in \mathfrak{I} \setminus \mathfrak{P}_\ell$. Hence, the above closure estimate yields that

$$\#\mathcal{T}_{\ell\nu} - \#\mathcal{T}_0 \leq C_{\text{cls}} \sum_{k=0}^{\ell-1} \#\mathcal{M}_{k\nu} \quad \text{for all } \nu \in \mathfrak{I}. \quad (35)$$

Moreover, due to the definition of parametric enrichment, there holds

$$\#\mathfrak{P}_\ell = \#\mathfrak{P}_0 + \sum_{k=0}^{\ell-1} \#\mathfrak{M}_k. \quad (36)$$

11

Recall that $\mathcal{M}_{k\nu} = \emptyset$ if $\nu \in \mathfrak{I} \setminus \mathfrak{P}_k$. Therefore, we are led to

$$\#\mathbb{P}_\ell \overset{(25)}{=} \sum_{\nu \in \mathfrak{P}_\ell} \left[\#\mathcal{T}_{\ell\nu} - \#\mathcal{T}_0 + 1 \right] = \#\mathfrak{P}_\ell + \sum_{\nu \in \mathfrak{P}_\ell} \left[\#\mathcal{T}_{\ell\nu} - \#\mathcal{T}_0 \right]$$

$$\overset{(35)}{\leq} \#\mathfrak{P}_\ell + C_{\mathrm{cls}} \sum_{\nu \in \mathfrak{P}_\ell} \sum_{k=0}^{\ell-1} \#\mathcal{M}_{k\nu}$$

$$\overset{(36)}{=} \#\mathfrak{P}_0 + \sum_{k=0}^{\ell-1} \left(\#\mathfrak{M}_k + C_{\mathrm{cls}} \sum_{\nu \in \mathfrak{P}_k} \#\mathcal{M}_{k\nu} \right).$$

With $\#\mathfrak{P}_0 = \#\mathbb{P}_0$ and $C_{\mathrm{cls}} \geq 1$, this concludes the proof. $\qquad\square$

6.2. Comparison lemma and proof of rate optimality. To prove optimal rates, it remains to compare the multilevel structures $(\mathbb{P}_\ell)_{\ell \in \mathbb{N}_0}$ generated by Algorithm 3 with respective optimal choices provided by the definition of the approximation class $\mathbb{A}_s$ in (27). This is (implicitly) done in the following lemma, which will be exploited in the proof of estimate (29) of Theorem 5.

Lemma 9 (comparison lemma). *Suppose the strong saturation assumption* (24). *Then, there exists $0 < \theta_{\mathrm{opt}} \leq 1$ depending only on C_{est} and q_{sat} such that the following statement holds: Let $s > 0$ and $\|\boldsymbol{u}\|_{\mathbb{A}_s} < \infty$; then, for all $\mathbb{P}_\bullet \in \mathbf{REFINE}(\mathbb{P}_0)$, there exist $\mathfrak{R}_\bullet \subseteq \mathfrak{Q}_\bullet$ and $\mathcal{R}_{\bullet\nu} \subseteq \mathcal{N}_{\bullet\nu}^+$ for all $\nu \in \mathfrak{P}_\bullet$ such that*

$$\#\mathfrak{R}_\bullet + \sum_{\nu \in \mathfrak{P}_\bullet} \#\mathcal{R}_{\bullet\nu} \leq 3\, \kappa_{\mathrm{sat}}^{-1/s} \|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \vertiii{\boldsymbol{u} - \boldsymbol{u}_\bullet}^{-1/s} \tag{37}$$

as well as

$$\theta_{\mathrm{opt}}\, \tau_\bullet^2 \leq \left(\sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{R}_{\bullet\nu}} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{R}_\bullet} \tau_\bullet(\nu)^2 \right). \tag{38}$$

Proof. If $\vertiii{\boldsymbol{u} - \boldsymbol{u}_\bullet} = 0$ or $\|\boldsymbol{u}\|_{\mathbb{A}_s} = \infty$, we may simply choose $\mathfrak{R}_\bullet = \mathfrak{Q}_\bullet$ and $\mathcal{R}_{\bullet\nu} = \mathcal{N}_{\bullet\nu}^+$ for all $\nu \in \mathfrak{P}_\bullet$ so that (37)–(38) are trivially satisfied.

Hence, we may suppose that $\vertiii{\boldsymbol{u} - \boldsymbol{u}_\bullet} > 0$ and $\|\boldsymbol{u}\|_{\mathbb{A}_s} < \infty$. With $0 < \kappa_{\mathrm{sat}} < 1$ from the strong saturation assumption (24), define

$$0 < \varepsilon := \kappa_{\mathrm{sat}} \vertiii{\boldsymbol{u} - \boldsymbol{u}_\bullet} \overset{(13)}{<} \vertiii{\boldsymbol{u} - \boldsymbol{u}_0} \overset{(27)}{\leq} \|\boldsymbol{u}\|_{\mathbb{A}_s} < \infty. \tag{39}$$

By construction, it thus holds that $\|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \varepsilon^{-1/s} > 1$. Choose $N \in \mathbb{N}$ such that

$$1 \leq N < \|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \varepsilon^{-1/s} \leq N + 1. \tag{40}$$

Choose $\mathbb{P}_\varepsilon \in \mathbf{REFINE}(\mathbb{P}_0)$ such that

$$\#\mathbb{P}_\varepsilon - \#\mathbb{P}_0 \leq N \quad \text{and} \quad \vertiii{\boldsymbol{u} - \boldsymbol{u}_\varepsilon} = \min_{\substack{\mathbb{P}_{\mathrm{opt}} \in \mathbf{REFINE}(\mathbb{P}_0) \\ \#\mathbb{P}_{\mathrm{opt}} - \#\mathbb{P}_0 \leq N}} \vertiii{\boldsymbol{u} - \boldsymbol{u}_{\mathrm{opt}}}. \tag{41}$$

Define the overlay $\mathbb{P}_\star := \mathbb{P}_\varepsilon \oplus \mathbb{P}_\bullet$. Then, it follows that

$$\#\mathbb{P}_\star - \#\mathbb{P}_\bullet \overset{(32)}{\leq} \#\mathbb{P}_\varepsilon - \#\mathbb{P}_0 \overset{(41)}{\leq} N \overset{(40)}{<} \|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \varepsilon^{-1/s} \overset{(39)}{=} \kappa_{\mathrm{sat}}^{-1/s} \|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \vertiii{\boldsymbol{u} - \boldsymbol{u}_\bullet}^{-1/s}$$

12

as well as

$$\|\, \boldsymbol{u} - \boldsymbol{u}_\star \,\| \overset{(13)}{\leq} \|\, \boldsymbol{u} - \boldsymbol{u}_\varepsilon \,\| \overset{(41)}{=} \min_{\substack{\mathbb{P}_{\mathrm{opt}} \in \mathbf{REFINE}(\mathbb{P}_0) \\ \#\mathbb{P}_{\mathrm{opt}} - \#\mathbb{P}_0 \leq N}} \|\, \boldsymbol{u} - \boldsymbol{u}_{\mathrm{opt}} \,\| \overset{(27)}{\leq} (N+1)^{-s}\, \|\boldsymbol{u}\|_{\mathbb{A}_s}$$

$$\overset{(40)}{\leq} \varepsilon \overset{(39)}{=} \kappa_{\mathrm{sat}} \|\, \boldsymbol{u} - \boldsymbol{u}_\bullet \,\|.$$

Define $\mathbb{P}_\circ := \mathbf{REFINE}(\mathbb{P}_\bullet, \mathbf{M}_\bullet)$ with $\mathbf{M}_\bullet := [\mathfrak{P}_\star \cap \mathfrak{Q}_\bullet, (\mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu})_{\nu \in \mathfrak{P}_\bullet}]$. Then, the strong saturation assumption (24) yields that $\|\, \boldsymbol{u} - \boldsymbol{u}_\circ \,\| \leq q_{\mathrm{sat}} \|\, \boldsymbol{u} - \boldsymbol{u}_\bullet \,\|$. According to Proposition 6, it thus follows that

$$\tau_\bullet^2 \overset{(30)}{\leq} C_{\mathrm{lin}} \left(\sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\circ\nu}} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{Q}_\bullet \cap \mathfrak{P}_\circ} \tau_\bullet(\nu)^2 \right) \quad \text{with} \quad C_{\mathrm{lin}} := \frac{C_{\mathrm{est}}^4}{1 - q_{\mathrm{sat}}^2} > 1.$$

Let $0 < \theta_{\mathrm{opt}} := 1/C_{\mathrm{lin}} < 1$. Since $\mathcal{N}_{\circ\nu} \subseteq \mathcal{N}_{\star\nu}$ and $\mathfrak{P}_\circ \subseteq \mathfrak{P}_\star$, we obtain that

$$\theta_{\mathrm{opt}} \, \tau_\bullet^2 \leq \left(\sum_{\nu \in \mathfrak{P}_\bullet} \sum_{z \in \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu}} \tau_\bullet(\nu, z)^2 + \sum_{\nu \in \mathfrak{Q}_\bullet \cap \mathfrak{P}_\star} \tau_\bullet(\nu)^2 \right).$$

Defining $\mathfrak{R}_\bullet := \mathfrak{Q}_\bullet \cap \mathfrak{P}_\star$ and $\mathcal{R}_{\bullet\nu} := \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu}$ for all $\nu \in \mathfrak{P}_\bullet$, we prove (38). To establish (37), it only remains to prove that

$$\#\mathfrak{R}_\bullet + \sum_{\nu \in \mathfrak{P}_\bullet} \#\mathcal{R}_{\bullet\nu} \leq 3\,(\#\mathbb{P}_\star - \#\mathbb{P}_\bullet).$$

To this end, note that

$$\begin{aligned}
\#\mathbb{P}_\star - \#\mathbb{P}_\bullet &= \sum_{\nu \in \mathfrak{P}_\star} (\#\mathcal{T}_{\star\nu} - \#\mathcal{T}_0 + 1) - \sum_{\nu \in \mathfrak{P}_\bullet} (\#\mathcal{T}_{\bullet\nu} - \#\mathcal{T}_0 + 1) \\
&= \sum_{\nu \in \mathfrak{P}_\bullet} (\#\mathcal{T}_{\star\nu} - \#\mathcal{T}_{\bullet\nu}) + \sum_{\nu \in \mathfrak{P}_\star \backslash \mathfrak{P}_\bullet} (\#\mathcal{T}_{\star\nu} - \#\mathcal{T}_0 + 1) \\
&\geq \sum_{\nu \in \mathfrak{P}_\bullet} \#(\mathcal{T}_{\bullet\nu} \backslash \mathcal{T}_{\star\nu}) + \#(\mathfrak{P}_\star \backslash \mathfrak{P}_\bullet).
\end{aligned}$$

First, we note that $\mathfrak{R}_\bullet = \mathfrak{Q}_\bullet \cap \mathfrak{P}_\star \subseteq \mathfrak{P}_\star \backslash \mathfrak{P}_\bullet$. Second, each added vertex $z \in \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu}$ leads to refinement of one edge and hence refinement of (at least) one triangle $T \in \mathcal{T}_{\bullet\nu} \backslash \mathcal{T}_{\star\nu}$. Moreover, each refined triangle $T \in \mathcal{T}_{\bullet\nu} \backslash \mathcal{T}_{\star\nu}$ contains at most three added vertices $z \in \mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu}$. This implies that $\#\mathcal{R}_{\bullet\nu} = \#(\mathcal{N}_{\bullet\nu}^+ \cap \mathcal{N}_{\star\nu}) \leq 3\,\#(\mathcal{T}_{\bullet\nu} \backslash \mathcal{T}_{\star\nu})$ for all $\nu \in \mathfrak{P}_\bullet$. Overall, we thus see that

$$\#\mathfrak{R}_\bullet + \sum_{\nu \in \mathfrak{P}_\bullet} \#\mathcal{R}_{\bullet\nu} \leq \#(\mathfrak{P}_\star \backslash \mathfrak{P}_\bullet) + 3 \sum_{\nu \in \mathfrak{P}_\bullet} \#(\mathcal{T}_{\bullet\nu} \backslash \mathcal{T}_{\star\nu}) \leq 3\,(\#\mathbb{P}_\star - \#\mathbb{P}_\bullet).$$

This concludes the proof. $\qquad\square$

We are now ready to prove the rate optimality property for Algorithm 3.

Proof of estimate (29) *in Theorem 5.* Note that

$$\#\mathbb{P}_\ell - \#\mathbb{P}_0 \overset{(34)}{\leq} C_{\mathrm{cls}} \sum_{k=0}^{\ell-1} \left(\#\mathfrak{M}_k + \sum_{\nu \in \mathfrak{P}_k} \#\mathcal{M}_{k\nu} \right).$$

13

Recall that by choice of Algorithm 3 (see step (iv) therein), the tuple $\mathbf{M}_k = [\mathfrak{M}_k, (\mathcal{M}_{k\nu})_{\nu \in \mathfrak{I}}]$ (with $\mathcal{M}_{k\nu} = \emptyset$ for $\nu \notin \mathfrak{P}_k$) satisfies the Dörfler marking criterion in (23). On the other hand, by virtue of Lemma 9, there exists a tuple $\mathbb{R}_k = [\mathfrak{R}_k, (\mathcal{R}_{k\nu})_{\nu \in \mathfrak{I}}]$ (with $\mathcal{R}_{k\nu} = \emptyset$ for $\nu \notin \mathfrak{P}_k$) satisfying (37) and (38). Since $0 < \theta \leq \theta_{\mathrm{opt}}$, inequality (38) implies that the tuple $\mathbb{R}_k = [\mathfrak{R}_k, (\mathcal{R}_{k\nu})_{\nu \in \mathfrak{I}}]$ satisfies the Dörfler marking criterion in (23). Therefore, according to the minimal cardinality property of $\mathbf{M}_k$, it follows that

$$\#\mathfrak{M}_k + \sum_{\nu \in \mathfrak{P}_k} \#\mathcal{M}_{k\nu} \;\leq\; \#\mathfrak{R}_k + \sum_{\nu \in \mathfrak{P}_k} \#\mathcal{R}_{k\nu} \;\overset{(37)}{\leq}\; 3\,\kappa_{\mathrm{sat}}^{-1/s} \|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \interleave \boldsymbol{u} - \boldsymbol{u}_k \interleave^{-1/s} \text{ for all } k \in \mathbb{N}_0.$$

With linear convergence $\interleave \boldsymbol{u} - \boldsymbol{u}_\ell \interleave \leq q_{\mathrm{lin}}^{\ell-k} \interleave \boldsymbol{u} - \boldsymbol{u}_k \interleave$ for all $0 \leq k < \ell$, the geometric series proves that

$$\#\mathbb{P}_\ell - \#\mathbb{P}_0 \leq 3\,C_{\mathrm{cls}}\,\kappa_{\mathrm{sat}}^{-1/s} \|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \sum_{k=0}^{\ell-1} \interleave \boldsymbol{u} - \boldsymbol{u}_k \interleave^{-1/s}$$

$$\leq 3\,C_{\mathrm{cls}}\,\kappa_{\mathrm{sat}}^{-1/s}\,\frac{q_{\mathrm{lin}}^{1/s}}{1 - q_{\mathrm{lin}}^{1/s}}\,\|\boldsymbol{u}\|_{\mathbb{A}_s}^{1/s} \interleave \boldsymbol{u} - \boldsymbol{u}_\ell \interleave^{-1/s} \quad \text{for all } \ell \in \mathbb{N}_0.$$

For $\ell \geq 1$, it follows that

$$(\#\mathbb{P}_\ell - \#\mathbb{P}_0 + 1)^s \interleave \boldsymbol{u} - \boldsymbol{u}_\ell \interleave \leq 2^s\,(\#\mathbb{P}_\ell - \#\mathbb{P}_0)^s \interleave \boldsymbol{u} - \boldsymbol{u}_\ell \interleave \leq \frac{q_{\mathrm{lin}}}{\kappa_{\mathrm{sat}}}\left(\frac{6\,C_{\mathrm{cls}}}{1 - q_{\mathrm{lin}}^{1/s}}\right)^s \|\boldsymbol{u}\|_{\mathbb{A}_s}.$$

For $\ell = 0$, there holds

$$(\#\mathbb{P}_\ell - \#\mathbb{P}_0 + 1)^s \interleave \boldsymbol{u} - \boldsymbol{u}_\ell \interleave = \interleave \boldsymbol{u} - \boldsymbol{u}_0 \interleave \leq \|\boldsymbol{u}\|_{\mathbb{A}_s}.$$

Combining the last two estimates, we prove (29) with $C_{\mathrm{opt}} = \max\left\{1, \dfrac{q_{\mathrm{lin}}}{\kappa_{\mathrm{sat}}}\left(\dfrac{6\,C_{\mathrm{cls}}}{1 - q_{\mathrm{lin}}^{1/s}}\right)^s\right\}$.

This concludes the proof. □

7. Numerical results

In this section, we present a collection of numerical results that illustrate the performance of Algorithm 3. All computations have been performed using the MATLAB toolbox Stochastic T-IFISS [BR19] (see [BRS20] for a recent review).

The details of the implementation of the proposed adaptive multilevel stochastic Galerkin FEM are presented in our recent work [BPR20] (see section 6 therein). In particular, exploiting the binary tree structure of the NVB refinement, we have shown that it is possible to compute all entries of the resulting linear system exactly (up to quadrature), even though for two multi-indices $\mu, \nu \in \mathfrak{P}_\ell$ ($\mu \neq \nu$), the associated meshes $\mathcal{T}_{\ell\mu}, \mathcal{T}_{\ell\nu}$ are, in general, different. The key observation is that $\mathcal{T}_{\ell\mu}, \mathcal{T}_{\ell\nu}$ are NVB refinements of the same initial mesh $\mathcal{T}_0$, and, therefore, for any $T \in \mathcal{T}_{\ell\mu}$ and $T' \in \mathcal{T}_{\ell\nu}$, there holds one of the following four cases:

- $T \cap T'$ is a set of measure zero,
- $T \subsetneq T'$,
- $T' \subsetneq T$,
- $T = T'$.

One can easily determine which of these four cases occurs by keeping a record of the number of bisections needed to generate the triangles T, T' and using the coordinates of their centers of mass (see [BPR20, Algorithm 9]).

7.1. Benchmark problem. Our first example is the parametric model problem introduced in [EGSZ14, section 11] and used for numerical experiments in, e.g., [EGSZ15, EM16, BR18, BPRR19, BPR20].

For $\boldsymbol{f} \equiv 1$, we consider the boundary value problem (1) on the L-shaped domain $D = (-1, 1)^2 \setminus (-1, 0]^2$. For all $x = (x_1, x_2) \in D$, the coefficients in the expansion (2) of the diffusion coefficient are given by

$$a_0(x) = 1, \quad a_m(x) = Am^{-2} \cos(2\pi\beta_1(m)x_1) \cos(2\pi\beta_2(m)x_2) \text{ for } m \in \mathbb{N}.$$

Here, $A = 0.9/\zeta(2) \approx 0.547$ (with $\zeta(\cdot)$ being the Riemann zeta function), $\beta_1(m) = m - k(m)[k(m) + 1]/2$, $\beta_2(m) = k(m) - \beta_1(m)$, and $k(m) = \lfloor -1/2 + \sqrt{1/2 + 2m} \rfloor$. With this choice, the diffusion coefficient $\boldsymbol{a}(x, \mathbf{y})$ trivially satisfies (3) (with $a_0^{\min} = a_0^{\max} = 1$), and both inequalities in (4) hold. The exact solution $\boldsymbol{u}$ to this problem exhibits a geometric singularity at the reentrant corner.

We run Algorithm 3 with different values of the Dörfler marking parameter $\theta = 0.1, \ldots, 0.9$. For all runs, we use the same initial mesh $\mathcal{T}_0$ (a uniform mesh of 384 right-angled triangles) and we stop the computation when $\tau_\ell \leq \mathtt{tol} := 2 \cdot 10^{-3}$.

In Figure 1, we plot the total error estimates τ_ℓ (left) and the reference energy errors $\|\|\boldsymbol{u}_{\mathrm{ref}} - \boldsymbol{u}_\ell\|\|$ (right) as functions of the total number of degrees of freedom (DOFs) $N_\ell = \dim \mathbb{V}_\ell = \sum_{\nu \in \mathfrak{P}_\ell} \dim \mathbb{X}_{\ell\nu}$. Here, the reference error is computed using a reference solution $\boldsymbol{u}_{\mathrm{ref}}$ obtained by running Algorithm 3 with $\theta = 0.5$ and a smaller tolerance (we set $\mathtt{tol} = 8 \cdot 10^{-4}$). We observe that the adaptive algorithm converges regardless of the value of the marking parameter. Moreover, the computations confirm the result of Theorem 5:
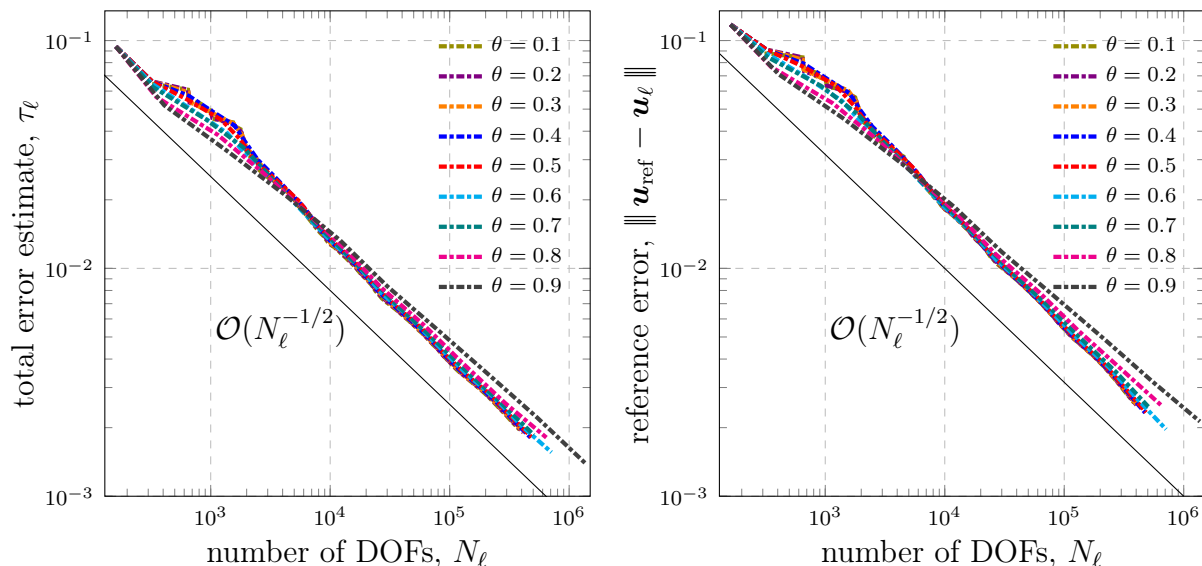

FIGURE 1. Experiments in section 7.1: Total error estimates τ_ℓ (left) and reference energy errors $\|\|\boldsymbol{u}_{\mathrm{ref}} - \boldsymbol{u}_\ell\|\|$ (right) versus the total number of DOFs $N_\ell = \dim \mathbb{V}_\ell$ for $\theta = 0.1, \ldots, 0.9$.
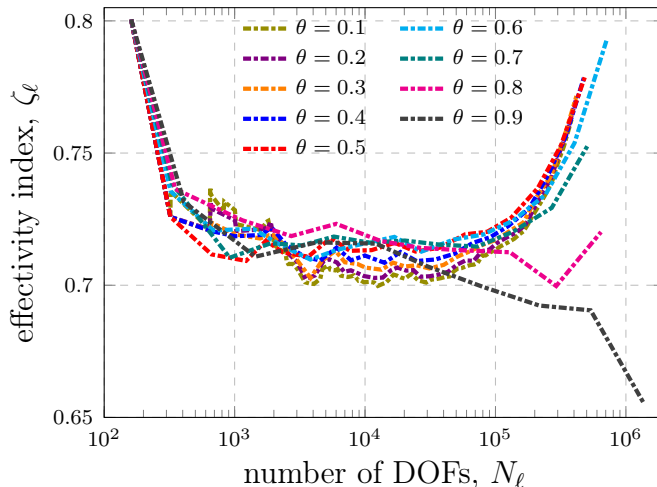
FIGURE 2. Experiments in section 7.1: Effectivity indices ζ_ℓ for the error estimates τ_ℓ for the SGFEM approximations generated by Algorithm 3 with $\theta = 0.1, \ldots, 0.9$.

for sufficiently small values of θ, the error estimates and the reference errors decay with the optimal rate $\mathcal{O}(N_\ell^{-1/2})$, which is the convergence rate of linear (P1) FEM for the corresponding parameter-free problem. For larger values of θ (in particular, for $\theta = 0.8, 0.9$), the convergence of the error estimates and the reference errors appears to be slightly suboptimal.

The suboptimality of the convergence rate for $\theta = 0.8, 0.9$ observed in Figure 1 is more evident for the reference error (right plot) than for the error estimate (left plot). To investigate this further we compute the *effectivity index*

$$\zeta_\ell := \frac{\tau_\ell}{\| \boldsymbol{u}_{\mathrm{ref}} - \boldsymbol{u}_\ell \|} = \frac{\tau_\ell}{\left(\| \boldsymbol{u}_{\mathrm{ref}} \|^2 - \| \boldsymbol{u}_\ell \|^2 \right)^{1/2}}$$

at each iteration of the adaptive loop. In Figure 2, we plot the effectivity indices ζ_ℓ versus the total number of degrees of freedom N_ℓ in computed SGFEM approximations for $\theta = 0.1, \ldots, 0.9$. For all $\theta \in \{0.1, \ldots, 0.7\}$, the effectivity indices exhibit similar behavior and vary in a range between 0.7 and 0.8 throughout all iterations. However, for $\theta \in \{0.8, 0.9\}$ and for large N_ℓ, the effectivity indices are becoming smaller. In particular, for $\theta = 0.9$, we even observe a deterioration of ζ_ℓ.

In Figure 3, we plot the cardinality of the index sets $\mathfrak{P}_\ell$ versus the total number of degrees of freedom N_ℓ for different values of θ. We see that the slope of the curve decreases as θ increases. Furthermore, while the curves for $\theta = 0.1, \ldots, 0.7$ (the values for which we observe optimal convergence rates in Figure 1) are relatively close to each other, those for $\theta = 0.8, 0.9$ exhibit a slower increase in the cardinality of the index set. This plot suggests that the lack of rate optimality observed for $\theta = 0.8, 0.9$ is associated with insufficient enrichments of the index set.

In Table 1, we show the following outputs for each run: the total number of iterations L needed to reach the prescribed tolerance, the dimension of the final approximation space, $N_L = \dim \mathbb{V}_L$, the final value of the total error estimate τ_L, the cardinality of the final index set $\mathfrak{P}_L$, the (total) degree $\deg \mathfrak{P}_L := \max_{\nu \in \mathfrak{P}_L} \sum_{j \geq 1} \nu_j$ of polynomials in the associated polynomial space, and the number of active parameters $M_{\mathfrak{P}_L}$ in $\mathfrak{P}_L$.
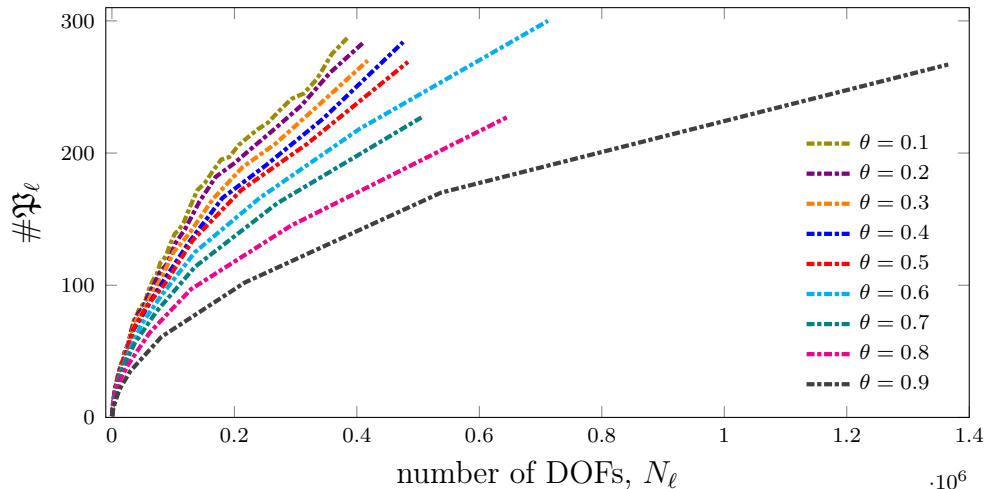
16

FIGURE 3. Experiments in section 7.1: Cardinality of the index set $\mathfrak{P}_\ell$ versus the total number of DOFs, $N_\ell = \dim \mathbb{V}_\ell$, for $\theta = 0.1, \ldots, 0.9$.

θ	L	N_L	τ_L	$\#\mathfrak{P}_L$	$\deg \mathfrak{P}_L$	$M_{\mathfrak{P}_L}$
0.1	95	384 241	$1.961\,225 \cdot 10^{-3}$	287	7	35
0.2	47	411 524	$1.906\,145 \cdot 10^{-3}$	284	7	31
0.3	31	417 773	$1.908\,122 \cdot 10^{-3}$	270	7	24
0.4	23	475 628	$1.809\,402 \cdot 10^{-3}$	284	8	20
0.5	18	483 559	$1.817\,273 \cdot 10^{-3}$	269	8	18
0.6	15	711 989	$1.558\,647 \cdot 10^{-3}$	300	8	15
0.7	12	505 518	$1.860\,177 \cdot 10^{-3}$	227	8	12
0.8	10	644 724	$1.809\,412 \cdot 10^{-3}$	227	8	10
0.9	9	1 365 625	$1.395\,409 \cdot 10^{-3}$	267	8	9

TABLE 1. Experiments in section 7.1: Final outputs of Algorithm 3 as $\tau_L \leq \mathtt{tol} := 2 \cdot 10^{-3}$.

Looking in detail at the results in Figure 3 and Table 1, we see how the value of the Dörfler marking parameter θ influences the convergence behavior of Algorithm 3. Clearly, the larger the value of θ, the smaller the total number of iterations L. Excluding the results for $\theta = 0.6, 0.9$ (for which the tolerance is met with a significantly smaller value of the error estimate), we observe a clear trend: the number of degrees of freedom necessary to achieve the same accuracy increases with θ (see, e.g., the results for $\theta = 0.4, 0.8$, for which the values of τ_L are nearly the same, whereas the value of N_L is significantly smaller for $\theta = 0.4$ than for $\theta = 0.8$). On the other hand, the results show that smaller values of θ yield larger index sets as well as a larger number of active parameters. From this, we infer that different values of θ induce different allocations of the degrees of freedom in adaptively refined SGFEM approximations. Specifically, smaller values of θ lead to
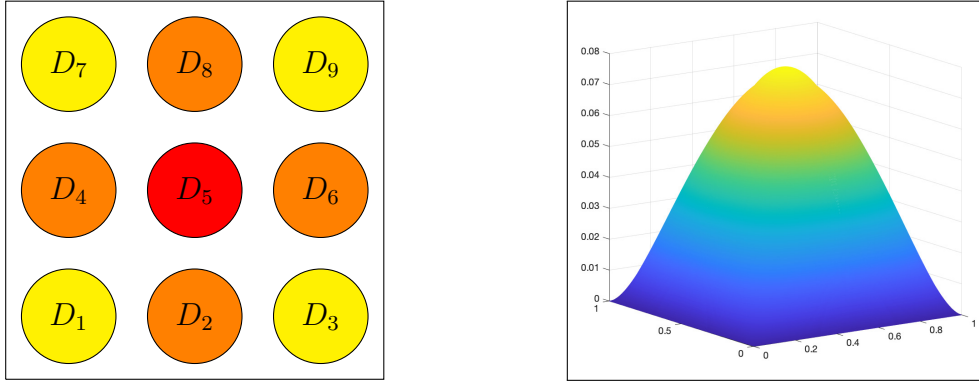
FIGURE 4. Experiments in section 7.2: Computational domain, where the subdomains of equal 'importance' are highlighted with the same color (left); the expectation of the reference SGFEM solution (right).

larger index sets and larger number of active parameters (which both yield more accurate *parametric* discretizations), whereas larger values of θ seem to generate finer meshes over the physical domain (hence, more accurate *spatial* discretizations).

7.2. Cookie problem. Our second example of parametric problem (1)–(2) is a version of the so-called *cookie problem*; cf. [BG15, ENSW19, BPR20]. Here, we adopt the setting of [BPR20, section 7.2] and consider the square domain $D = (0,1)^2$ that contains nine circular inclusions $D_m \subset D$ ($m = 1, \ldots, 9$). For all $i, j \in \{1, 2, 3\}$, the subdomain $D_{i+3(j-1)}$ is the disk with center at the point $((2i-1)/6, (2j-1)/6)$ and radius $r = 1/8$; see Figure 4(left). We set $\boldsymbol{f} \equiv 1$ in (1) and choose the expansion coefficients in (2) as follows:

$$a_0(x) = 1, \quad a_m(x) = k_m\, \chi_{D_m}(x) \text{ for } m = 1, \ldots, 9, \quad a_m(x) = 0 \text{ for } m > 9 \quad (x \in D),$$

where χ_{D_m} denotes the characteristic function of the subdomain D_m and

$$k_m = \begin{cases} 0.5 & \text{if } m = 1, 3, 7, 9, \\ 0.7 & \text{if } m = 2, 4, 6, 8, \\ 0.9 & \text{if } m = 5. \end{cases} \tag{42}$$

Thus, the diffusion coefficient $\boldsymbol{a}(x, \mathbf{y})$ in this example depends on finitely many parameters $y_1, \ldots, y_9 \in [-1, 1]$, and the amplitudes of the corresponding coefficients in the expansion (2) induce a hierarchy of these parameters, whereby y_5 is more 'important' than y_2, y_4, y_6, and y_8, which in turn are more 'important' than y_1, y_3, y_7, and y_9. With these choices, assumptions (3)–(4) are satisfied with $a_0^{\min} = a_0^{\max} = 1$ and $\tau = 0.9$. The expectation of an SGFEM solution to this problem is plotted in Figure 4(right).

We run Algorithm 3 with the same initial mesh $\mathcal{T}_0$ (a uniform mesh of 512 right-angled triangles) and different values of the Dörfler marking parameter $\theta = 0.1, \ldots, 0.9$. Each computation is terminated when the error estimate τ_ℓ falls below the tolerance $\mathtt{tol} := 10^{-3}$. Following [BPR20, section 7.2], we change the definition of the detail index set (10) to

$$\mathfrak{Q}_\bullet := \{\mu \in \mathbb{N}_0^9 \setminus \mathfrak{P}_\bullet : \mu = \nu \pm \varepsilon_m \text{ for all } \nu \in \mathfrak{P}_\bullet \text{ and all } m = 1, \ldots, 9\},$$

so that all relevant parameters are available for activation starting from the first iteration and the computations exhibit a shorter preasymptotic phase. We emphasize that the results of Theorem 2 and Theorem 5 remain valid in this case.
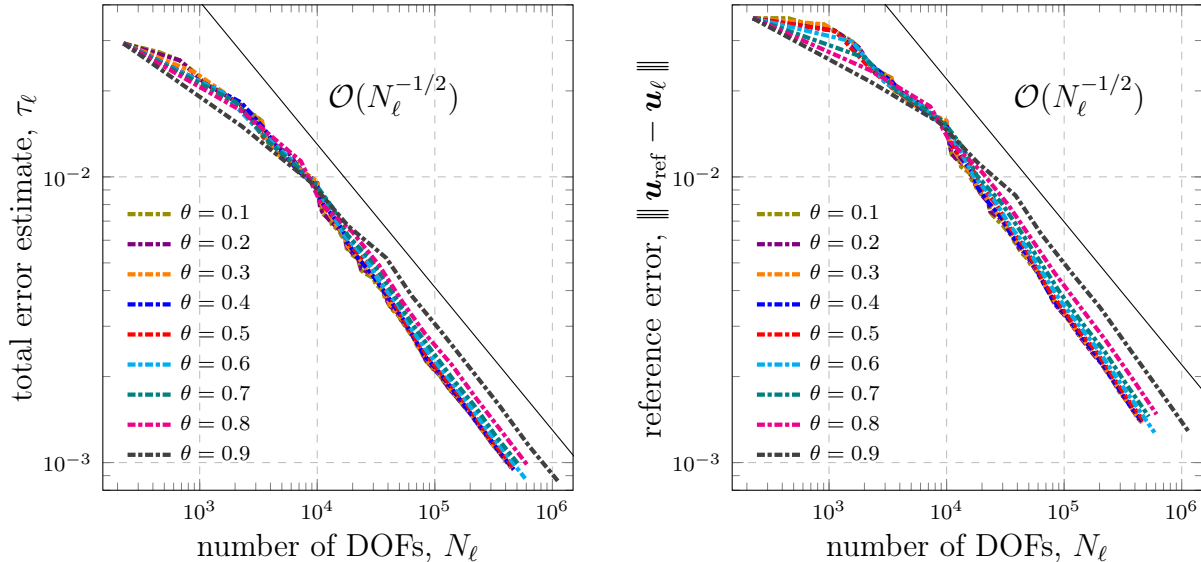

FIGURE 5. Experiments in section 7.2: Total error estimates τ_ℓ (left) and reference energy errors $\| \boldsymbol{u}_{\mathrm{ref}} - \boldsymbol{u}_\ell \|$ (right) versus the total number of DOFs $N_\ell = \dim \mathbb{V}_\ell$ for $\theta = 0.1, \ldots, 0.9$.

In Figure 5, we plot the total error estimates τ_ℓ (left) and the reference energy errors $\| \boldsymbol{u}_{\mathrm{ref}} - \boldsymbol{u}_\ell \|$ (right) versus the total number of degrees of freedom N_ℓ. Here, the reference solution $\boldsymbol{u}_{\mathrm{ref}}$ is computed by running Algorithm 3 with $\theta = 0.5$ to a smaller tolerance (we set $\mathtt{tol} = 4 \cdot 10^{-4}$). For all considered values of θ, the convergence rate of both the error estimates and the reference errors is $\mathcal{O}(N_\ell^{-1/2})$, i.e., the optimal rate for the SGFEM based on P1-FEM approximation in the physical domain. In contrast to what we observed for the test problem in section 7.1, the rate appears to be optimal also for $\theta = 0.8, 0.9$.

In Figure 6, we plot the effectivity indices ζ_ℓ versus the total number of degrees of freedom N_ℓ in computed SGFEM approximations for $\theta = 0.1, \ldots, 0.9$. In contrast to the test problem in section 7.1 (see Figure 2), Figure 6 shows essentially the same behavior of ζ_ℓ for all values of $\theta \in \{0.1, \ldots, 0.9\}$. Furthermore, the effectivity indices vary in a range between 0.6 and 0.82 throughout all iterations, which shows that the underestimation of the reference error in this example is more pronounced than for the test problem in section 7.1.

In Figure 7, we plot the cardinality of the index sets $\mathfrak{P}_\ell$ versus the total number of degrees of freedom N_ℓ for $\theta = 0.1, \ldots, 0.9$. We observe that the relative position of the curves is the inverse of the one in the corresponding plot for the experiment in section 7.1 (cf. Figure 3), with the curve for $\theta = 0.9$ exhibiting the fastest increase. Moreover, all curves are now positioned close to each other, which is consistent with the results presented in Figure 5, where the optimal convergence rate is observed for all values of θ.

The final outputs of Algorithm 3 are collected in Table 2. We can see that for all values of θ, Algorithm 3 identifies and activates all nine relevant parameters. For smaller values
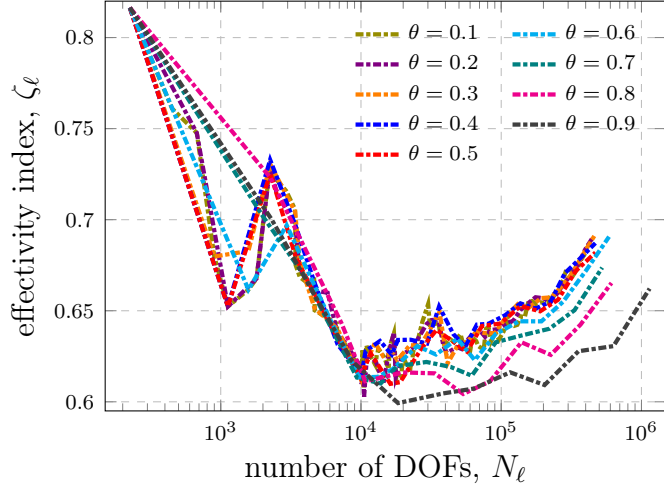
FIGURE 6. Experiments in section 7.2: Effectivity indices ζ_ℓ for the error estimates τ_ℓ for the SGFEM approximations generated by Algorithm 3 with $\theta = 0.1, \ldots, 0.9$.
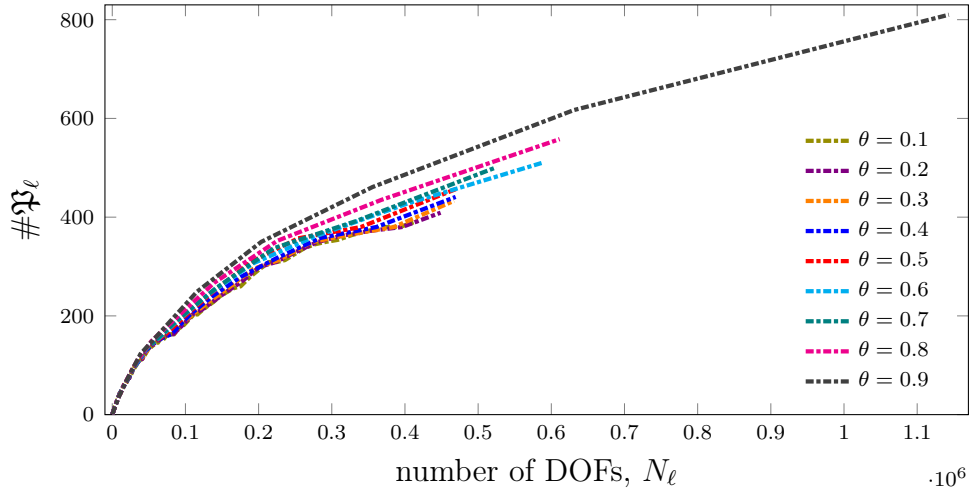

FIGURE 7. Experiments in section 7.2: Cardinality of the index set $\mathfrak{P}_\ell$ versus the total number of DOFs, $N_\ell = \dim \mathbb{V}_\ell$, for $\theta = 0.1, \ldots, 0.9$.

of θ, the algorithm generates final Galerkin approximations with significantly less degrees of freedom but requires significantly more iterations in order to reach the prescribed tolerance. This is in agreement with the results of section 7.1 (cf. Table 1 and the associated discussion). However, in this example, the cardinality of the final index set tends to increase and the total polynomial degree tends to decrease as θ increases, which is different from the behavior observed for the benchmark problem in section 7.1 (cf. Table 1). This difference in behavior is due to the diffusion coefficient in this example depending on *finitely* many parameters and the expansion coefficients in (2) having *local* supports.

Finally, in Table 3, for all $\theta = 0.1, \ldots, 0.9$, we show the maximum polynomial degree activated for each parameter y_m ($m = 1, \ldots, 9$). We can see that Algorithm 3 is effective in capturing the 'importance' of parameters. Indeed, the highest polynomial degree is always used for the parameter y_5 associated with the largest expansion coefficient; cf. (42). On

θ	L	N_L	τ_L	$\#\mathfrak{P}_L$	$\deg\mathfrak{P}_L$	$M_{\mathfrak{P}_L}$
0.1	100	416 841	$9.948\,64\cdot 10^{-4}$	392	15	9
0.2	50	448 722	$9.586\,15\cdot 10^{-4}$	409	16	9
0.3	33	467 023	$9.419\,73\cdot 10^{-4}$	433	16	9
0.4	24	469 294	$9.462\,78\cdot 10^{-4}$	441	16	9
0.5	19	469 862	$9.578\,95\cdot 10^{-4}$	457	15	9
0.6	16	587 344	$8.794\,47\cdot 10^{-4}$	510	15	9
0.7	13	524 901	$9.763\,64\cdot 10^{-4}$	501	12	9
0.8	11	611 416	$9.814\,06\cdot 10^{-4}$	558	11	9
0.9	10	1 143 257	$8.507\,08\cdot 10^{-4}$	810	10	9

TABLE 2. Experiments in section 7.2: Final outputs of Algorithm 3 as $\tau_L \leq \texttt{tol} := 10^{-3}$.

	$m = 1,\,3,\,7,\,9$	$m = 2,\,4,\,6,\,8$	$m = 5$
$\theta = 0.1$	5	8	15
$\theta = 0.2,\,0.3,\,0.4$	6	8	16
$\theta = 0.5,\,0.6$	6	9	15
$\theta = 0.7$	6	9	12
$\theta = 0.8$	6	9	11
$\theta = 0.9$	7	10	10

TABLE 3. Experiments in section 7.2: Maximum polynomial degree $\max_{\nu\in\mathfrak{P}_L} \nu_m$ in the final index set $\mathfrak{P}_L$ generated by Algorithm 3 for $m = 1,\dots,9$ and for $\theta = 0.1,\dots,0.9$.

the other hand, the range of maximum polynomial degrees for different sets of parameters (or, subdomains) shrinks as θ increases.

REFERENCES

[BDD04]   P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2):219–268, 2004.
[BEK96]   F. A. Bornemann, B. Erdmann, and R. Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM J. Numer. Anal.*, 33(3):1188–1204, 1996.
[BG15]   J. Ballani and L. Grasedyck. Hierarchical tensor approximation of output quantities of parameter-dependent PDEs. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):852–872, 2015.
[BPR20]   A. Bespalov, D. Praetorius, and M. Ruggeri. Two-level a posteriori error estimation for adaptive multilevel stochastic Galerkin FEM. arXiv:2006.02255, 2020.
[BPRR19] A. Bespalov, D. Praetorius, L. Rocchi, and M. Ruggeri. Convergence of adaptive stochastic Galerkin FEM. *SIAM J. Numer. Anal.*, 57(5):2359–2382, 2019.
[BR18]   A. Bespalov and L. Rocchi. Efficient adaptive algorithms for elliptic PDEs with random data. *SIAM/ASA J. Uncertain. Quantif.*, 6(1):243–272, 2018.

[BR19]   A. Bespalov and L. Rocchi. Stochastic T-IFISS, February 2019. Available online at `http://web.mat.bham.ac.uk/A.Bespalov/software/index.html#stoch_tifiss`.

[BRS20]   A. Bespalov, L. Rocchi, and D. Silvester. T-IFISS: a toolbox for adaptive FEM computation. *Comput. Math. Appl.*, 2020. In press.

[BS16]   A. Bespalov and D. Silvester. Efficient adaptive stochastic Galerkin methods for parametric operator equations. *SIAM J. Sci. Comput.*, 38(4):A2118–A2140, 2016.

[CDS10]   A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646, 2010.

[CDS11]   A. Cohen, R. DeVore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl.*, 9(1):11–47, 2011.

[CGG16]   C. Carstensen, D. Gallistl, and J. Gedicke. Justification of the saturation assumption. *Numer. Math.*, 134(1):1–25, 2016.

[CKNS08]   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(5):2524–2550, 2008.

[CPB19]   A. J. Crowder, C. E. Powell, and A. Bespalov. Efficient adaptive multilevel stochastic Galerkin approximation using implicit a posteriori error estimation. *SIAM J. Sci. Comput.*, 41(3):A1681–A1705, 2019.

[DN02]   W. Dörfler and R. H. Nochetto. Small data oscillation implies the saturation assumption. *Numer. Math.*, 91(1):1–12, 2002.

[EGSZ14]   M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander. Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Engrg.*, 270:247–269, 2014.

[EGSZ15]   M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander. A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes. *ESAIM Math. Model. Numer. Anal.*, 49(5):1367–1398, 2015.

[EM16]   M. Eigel and C. Merdon. Local equilibration error estimators for guaranteed error control in adaptive stochastic higher-order Galerkin finite element methods. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):1372–1397, 2016.

[ENSW19]   M. Eigel, J. Neumann, R. Schneider, and S. Wolf. Non-intrusive tensor reconstruction for high-dimensional random PDEs. *Comput. Meth. Appl. Mat.*, 19(1):39–53, 2019.

[Git13]   C. J. Gittelson. Convergence rates of multilevel and sparse tensor approximations for a random elliptic PDE. *SIAM J. Numer. Anal.*, 51(4):2426–2447, 2013.

[KPP13]   M. Karkulik, D. Pavlicek, and D. Praetorius. On 2D newest vertex bisection: Optimality of mesh-closure and H^1-stability of L_2-projection. *Constr. Approx.*, 38:213–234, 2013.

[PRS20]   D. Praetorius, M. Ruggeri, and E. P. Stephan. The saturation assumption yields optimal convergence of two-level adaptive BEM. *Appl. Numer. Math.*, 152:105–124, 2020.

[SG11]   C. Schwab and C. J. Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 20:291–467, 2011.

[Ste07]   R. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007.

[Ste08]   R. Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77(261):227–241, 2008.

SCHOOL OF MATHEMATICS, UNIVERSITY OF BIRMINGHAM, EDGBASTON, BIRMINGHAM B15 2TT, UK

Email address: `a.bespalov@bham.ac.uk`

INSTITUTE OF ANALYSIS AND SCIENTIFIC COMPUTING, TU WIEN, WIEDNER HAUPTSTRASSE 8–10, 1040 VIENNA, AUSTRIA

Email address: `dirk.praetorius@asc.tuwien.ac.at`

INSTITUTE OF ANALYSIS AND SCIENTIFIC COMPUTING, TU WIEN, WIEDNER HAUPTSTRASSE 8–10, 1040 VIENNA, AUSTRIA

Email address: `michele.ruggeri@asc.tuwien.ac.at`