ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Adaptive stochastic Galerkin FEM

M. Eigel and C. Gittelson and C. Schwab and E. Zander

Research Report No. 2013-01
January 2013

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

- Funding SNF: 200021-120290/1
- Funding ERC: Advanced Grant grant STAHDPDE 247277
- Funding: MATHEON project C33

ADAPTIVE STOCHASTIC GALERKIN FEM

MARTIN EIGEL, CLAUDE JEFFREY GITTELSON, CHRISTOPH SCHWAB, AND ELMAR ZANDER

Abstract. A framework for residual-based a posteriori error estimation and adaptive mesh refinement and polynomial chaos expansion for general second order linear elliptic PDEs with random coefficients is presented. A parametric, deterministic elliptic boundary value problem on an infinite-dimensional parameter space is discretized by means of a Galerkin projection onto finite generalized polynomial chaos (gpc) expansions, and by discretizing each gpc coefficient by a FEM in the physical domain.

An anisotropic residual-based a posteriori error estimator is developed. It contains bounds for both contributions to the overall error: the error due to gpc discretization and the error due to Finite Element discretization of the gpc coefficients in the expansion. The reliability of the residual estimator is established.

Based on the explicit form of the residual estimator, an adaptive refinement strategy is presented which allows to steer the polynomial degree adaptation and the dimension adaptation in the stochastic Galerkin discretization, and, embedded in the gpc adaptation loop, also the Finite Element mesh refinement of the gpc coefficients in the physical domain. Asynchronous mesh adaptation for different gpc coefficients is permitted, subject to a minimal compatibility requirement on meshes used for different gpc coefficients.

Details on the implementation in the software environment FEniCS are presented; it is generic, and is based on available stiffness and mass matrices of a FEM for the deterministic, nonparametric nominal problem.

Preconditioning of the resulting matrix equation and iterative solution are discussed. Numerical experiments in two spatial dimensions for membrane and plane stress boundary value problems on polygons are presented. They indicate substantial savings in total computational complexity due to FE mesh coarsening in high gpc coefficients.

Date: December 18, 2012.

2010 *Mathematics Subject Classification.* 35R60,47B80,60H35,65C20,65N12,65N22,65J10.

Key words and phrases. partial differential equations with random coefficients, parabolic differential equations, uncertainty quantification, stochastic finite element methods, operator equations, FEniCS, adaptive methods.

Research supported in part by the MATHEON project C33.

Research supported in part by the Swiss National Science Foundation grant No. 200021-120290/1.

Research supported in part by the ERC Advanced Grant grant STAHDPDE 247277.

The authors thank Roman Andreev for his support early in the project.

1. Introduction

The origins of the problems examined here are in uncertainty quantification (UQ) when elliptic PDEs with random field inputs are considered and the input's random coefficients are given in terms of a Karhunen–Loève expansion. By now, the reduction of a PDE with random inputs to a parametric, deterministic PDE on a possibly infinite-dimensional parameter space is standard (see, *e.g.*, [22] and the references therein). In recent years, the efficient numerical solution of such PDEs has received substantial attention. Two broad classes of algorithms to this end have emerged: stochastic Galerkin and stochastic collocation discretizations.

Stochastic collocation is algorithmically reminiscent of statistical sampling methods which require repeated execution of a given, deterministic solver with instances (similar to Monte Carlo samples) of input data, and is thus termed *nonintrusive*. This means that a given forward simulation software does not mandate modifications in order to execute numerical uncertainty quantification analysis. Recent progress in mathematical formulation has provided some theoretical basis for the convergence properties of stochastic collocation (*e.g.* [2, 4, 5, 22] and the references therein).

In applications, stochastic Galerkin discretizations have been labelled as *intrusive* since they require reformulation of the problem, and are perceived to require at least partial redesign of code for the generation and assembly of the stochastic Galerkin stiffness matrix and the load vector(s). There are also delicate mathematical issues as to what constitutes a proper *mathematical stochastic Galerkin formulation*, in particular in connection with probability measures whose densities are either unbounded or nearly degenerate. We refer to [12] and the references there for results and (counter) examples as well as for discussion of (nonequivalent) mathematical formulations of stochastic Galerkin FEM. These issues are absent in stochastic Galerkin formulations whose random input variables have bounded supports. In theses cases, there is a solid mathematical foundation of formulation and convergence of stochastic Galerkin FEM (see, *e.g.*, [16, 14, 15, 22] and the references therein). In particular, natural advantages of Galerkin discretizations can be brought to bear: Galerkin orthogonality and weak residuals are available for a posteriori error estimation and adaptive refinement of discretizations in physical space (and time, for evolution problems) as well as of truncated polynomial chaos expansions in probability space are available. These advantages of stochastic Galerkin FEM are offset, however, to some extent by the seeming substantial additional coding effort and the need to solve a massive, tensor-structured linear system of equations which arise from the spectral stochastic Galerkin discretizations.

One purpose of this paper is to demonstrate that at least for elliptic PDEs with random inputs, this view is not entirely accurate: we derive *residual a posteriori error estimators* for stochastic Galerkin Finite Element discretizations and a *novel matrix assembly algorithm* for Galerkin discretizations of PDEs with random input data which, in conjunction with a block-diagonal preconditioning technique used *e.g.* in [16], will allow for the efficient iterative solution of the stochastic Galerkin equations and, in addition, never requires the actual assembly of the entire stochastic Galerkin matrix. In addition, we exhibit a tensor structure of the matrices which arise in the stochastic Galerkin discretization and explain how Galerkin orthogonality in the discretization error allows for *error separation* in residual error estimators: the weak residual allows orthogonal (in mean-square) decomposition into contributions from the stochastic Galerkin discretization and from the Finite Element approximation of the gpc coefficients. This error separation is verified in the present paper for a rather straightforward, residual a posteriori error estimator, which is also found to perform well in our numerical experiments. We emphasize,

however, that *the basic orthogonality property which we used in deriving the error estimator and the error separation in the weak residual can reasonably be expected to hold for any of the by now numerous a posteriori error estimation methodologies* which are available in the Finite Element Method. We refer to [7] for a survey and further references. We show in particular for linear elliptic problems arising in computational mechanics how the stochastic Galerkin FEM can, indeed, be implemented with complexity that is comparable to that of a stochastic collocation solution. In addition, we show also how mesh adaptivity with different levels of Finite Element mesh refinement for each coefficient of an approximate stochastic Galerkin solution can be realized algorithmically. In numerical experiments we confirm the reliability of the adaptive strategy and also the substantial savings gained by the possibility of nonuniform mesh adaptation for different gpc coefficients which was theoretically predicted in [9].

The outline of this paper is as follows. In Section 2, we present a class of model parametric elliptic diffusion problems in a polygonal or polyhedral, "physical" domain D (to be distinguished from the "stochastic" domain) whose differential operators depend in an affine fashion on a sequence $y = (y_m)_{m=1}^\infty$ of parameters.

Section 3 present the gpc expansion of the solution. Specifically, finite spans of tensorized Legendre polynomials are used for the Galerkin projection. Section 4 presents the stochastic Galerkin formulation. Section 5 presents the derivation of the weak residual, in particular with a decomposition of the overall weak residual into a part from gpc discretization and a second, orthogonal part due to FE discretization of "active" gpc coefficients in the domain D. Section 6 contains the derivation of the residual error estimator. Section 7 presents the adaptive refinement criteria, in particular the marking of elements for subdivision and new gpc modes to be appended to the Galerkin approximation. Section 8 discusses the iterative solver, which consists of a pcg block iteration. Termination criteria are developed which balance the iteration error with the Finite Element discretization and the gpc truncation error. Section 9 then discusses the extension of the developed concepts to general linear elliptic problems in two space dimensions. As an illustration, the FE discretization of plane, linearized elastostatics with stochastic Poisson ratio is presented. Section 10 finally adresses the concrete realization of the Galerkin operator. Due to the iterative solver, the *matrix in the stochastic Galerkin FEM is never explicitly assembled*, and we present a "dynamic" assembly which is realized in the matrix-vector multiplication in our solver. The paper concludes with several numerical experiments in Section 11, among others a Cook's membrane test with stochastic Poisson ratio.

2. A parametric boundary value problem

2.1. **Parametric form.** We consider the model parametric elliptic boundary value problem

$$\begin{cases} -\nabla\cdot(a\nabla u) = f & \text{in } D \\ u = 0 & \text{on } \partial D \end{cases} \tag{2.1}$$

on a bounded Lipschitz domain $D \subset \mathbb{R}^d$. Points in the physical domain D shall be denoted by x with coordinates $(x_1, \ldots, x_d)$. In (2.1), the coefficient a is permitted to depend on a sequence of scalar parameters y_m in an affine fashion

$$a(y,x) = \bar{a}(x) + \sum_{m=1}^{\infty} y_m a_m(x)\,, \quad x \in D\,, \tag{2.2}$$

with $\bar{a}, a_m \in W^{1,\infty}(D)$ and $|y_m| \leq 1$, which entails $y \coloneqq (y_m)_{m=1}^\infty \in \Gamma \coloneqq [-1,1]^\infty$. Parametric coefficients such as (2.2) arise, for example, as Karhunen–Loève expansions of random fields $a(\omega, x)$ in D. In order to ensure convergence in (2.2) and positivity of a, we assume

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

additional summability assumptions are made in Section 5.2. We will use the notation $\Sigma_m \coloneqq \operatorname{supp} a_m \subset D$.

The variational formulation of (2.1) without the parameter y is set in the space $V \coloneqq H^1_0(D)$, which we endow with the $\bar{a}$-dependent scalar product

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

and the induced norm $\|\cdot\|_V$. More generally, for any measurable subset $G \subset D$, we define the semi-definite form $(\cdot,\cdot)_{V,G}$ and corresponding seminorm $|\cdot|_{V,G}$ by restricting the integral in (2.4) to G.

The operator

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

can be interpreted as the Riesz isomorphism from V to V^*, and is thus boundedly invertible. We also define the bounded linear maps

$$A_m\colon H^1_0(D) \to H^{-1}(D)\,, \quad v \mapsto -\nabla\cdot(a_m\nabla v)\,, \quad m \in \mathbb{N}\,, \tag{2.6}$$

through which we can express

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

as

$$A(y) = \bar{A} + \sum_{m=1}^\infty y_m A_m\,, \quad y \in \Gamma\,, \tag{2.8}$$

with unconditional convergence in $\mathcal{L}(V, V^*)$. Then equation (2.1) is expressed succinctly as

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

2.2. **Weak formulation.** Anticipating the approximation of u by a Galerkin projection simultaneously in $x \in D$ and $y \in \Gamma$, we integrate (2.9) with respect to a measure on Γ. This could be a statistically meaningful probability distribution if a is modeled as a random field, or an auxiliary measure used only for numerical purposes.

For all $m \in \mathbb{N}$, let π_m be a symmetric Borel probability measure on $[-1,1]$, *i.e.* π_m is invariant under the transformation $y_m \mapsto -y_m$. We assume for simplicity that the support of π_m in $[-1,1]$ has infinite cardinality. Then

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

is a probability measure on Γ with the Borel σ-algebra.

The weak formulation of (2.9) is to find $u \in L^2_\pi(\Gamma; V)$ such that

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

Existence and uniqueness of the solution u are a consequence of the Riesz isomorphism since the bilinear form in (2.11) defines a scalar product on $L^2_\pi(\Gamma; V)$. Furthermore, the solution coincides with that of (2.9) for π-a.e. $y \in \Gamma$.

The weak formulation (2.11) can be cast as an operator equation $\mathcal{A}u = f$ with

$$\mathcal{A}\colon L^2_\pi(\Gamma;V) \to L^2_\pi(\Gamma;V^*)\,, \quad v \mapsto [y \mapsto A(y)v(y)]\,. \tag{2.12}$$

Due to (2.8), identifying $L^2_\pi(\Gamma;V)$ with the Hilbert tensor product $L^2_\pi(\Gamma)\otimes V$, and similarly for V^* in place of V, $\mathcal{A}$ has the tensor product expansion

$$\mathcal{A} = \mathrm{id}_{L^2_\pi(\Gamma)} \otimes \bar{A} + \sum_{m=1}^\infty K_m \otimes A_m\,, \tag{2.13}$$

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

We define the energy norm $\|\cdot\|_{\mathcal{A}}$ on $L^2_\pi(\Gamma;V)$ through the scalar product

$$(w,v)_{\mathcal{A}} \coloneqq \langle \mathcal{A}w, v\rangle = \int_\Gamma \langle A(y)w(y), v(y)\rangle \,\mathrm{d}\pi(y)\,. \tag{2.14}$$

3. Tensor product polynomial expansion

3.1. **Orthonormal polynomials.** By definition, for every $m \in \mathbb{N}$, π_m is a symmetric probability measure on $[-1,1]$ whose support has infinite cardinality. We denote by $(P^m_n)_{n=0}^\infty$ an orthonormal basis of $L^2_{\pi_m}([-1,1])$, where P^m_n is a polynomial of degree n. Such a basis can be constructed by Gram–Schmidt orthogonalization of the monomial basis, which leads to a recursion

$$\beta^m_n P^m_n(y_m) = y_m P^m_{n-1}(y_m) - \beta^m_{n-1} P^m_{n-2}(y_m)\,, \quad n \ge 1\,, \tag{3.1}$$

with the initialization $P^m_0 \coloneqq 1$ and $\beta^m_0 \coloneqq 0$. The polynomials P^m_n are unique *e.g.* if β^m_n are chosen as positive for all $n \ge 1$.

For example, if $\mathrm{d}\pi_m(y_m) = \frac{1}{2}\,\mathrm{d}y_m$ is the uniform distribution, then $(P^m_n)_{n=0}^\infty$ are Legendre polynomials, and $\beta^m_n = (4-n^{-2})^{-1/2}$. Alternatively, if $\mathrm{d}\pi_m(y_m) = \frac{1}{\pi}(1-y_m^2)^{-1/2}\,\mathrm{d}y_m$, then $(P^m_n)_{n=0}^\infty$ are Chebyshev polynomials of the first kind, with $\beta^m_1 = 1/\sqrt{2}$ and $\beta^m_n = 1/2$ for $n \ge 2$. We refer to [13, 16] for details and further examples.

More generally, if π_m is not symmetric, an additional term appears in (3.1). Furthermore, if π_m is a convex combination of point masses, then $L^2_{\pi_m}([-1,1])$ is finite dimensional, and thus the polynomial basis is also finite. Our results extend to these cases, but we restrict to symmetric measures with infinite support to simplify notation.

3.2. **Tensorized basis.** Tensor products of the polynomial bases in each coordinate of $\Gamma = [-1,1]^\infty$ form a basis of $L^2_\pi(\Gamma)$. Let $\mathcal{F}$ denote the set of finitely supported sequences in $\mathbb{N}_0$,

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

where $\operatorname{supp}\mu \coloneqq \{m \in \mathbb{N}\,;\, \mu_m \ne 0\}$. For any $\mu \in \mathcal{F}$, the countable tensor product polynomial $P_\mu \coloneqq \bigotimes_{m=1}^\infty P^m_{\mu_m}$ is given by

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

since $P^m_0 = 1$ for all $m \in \mathbb{N}$. The countable set $(P_\mu)_{\mu\in\mathcal{F}}$ is an orthonormal basis of $L^2_\pi(\Gamma)$, see *e.g.* [16, 12].

We denote the element of $\mathcal{F}$ consisting only of zeros by 0. The corresponding basis function is $P_0 = 1$. Also, for any $m \in \mathbb{N}$, $\epsilon_m \coloneqq (\delta_{mn})_{n=1}^\infty$ denotes the Kronecker sequence for the coordinate m.

Due to (3.1) and (3.3), for any $\mu \in \mathcal{F}$ and $m \in \mathbb{N}$,

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

where we use the convention $P_\mu = 0$ if any $\mu_m < 0$. Thus multiplication by y_m has a particularly simple representation with respect to the basis $(P_\mu)_{\mu\in\mathcal{F}}$.

3.3. **Reformulation of the parametric equation.** Since $(P_\mu)_{\mu\in\mathcal{F}}$ is an orthonormal basis of $L^2_\pi(\Gamma)$, the solution u of (2.11) can be expanded as

$$u(y,x) = \sum_{\mu\in\mathcal{F}} u_\mu(x) P_\mu(y)\,, \tag{3.5}$$

with coefficients u_μ in $V = H^1_0(D)$ and convergence in $L^2_\pi(\Gamma;V)$.

Inserting (3.5) and an analogous expansion of v into (2.11) leads to the countably infinite coupled system of deterministic equations

$$\bar{A}u_\mu + \sum_{m=1}^{\infty} A_m(\beta^m_{\mu_m+1} u_{\mu+\epsilon_m} + \beta^m_{\mu_m} u_{\mu-\epsilon_m}) = f\delta_{\mu 0} \qquad \forall \mu\in\mathcal{F} \tag{3.6}$$

for the coefficient vector $(u_\mu)_{\mu\in\mathcal{F}} \in \ell^2(\mathcal{F};V)$. We refer to [16] for a mathematically rigorous derivation.

4. Galerkin projection

4.1. **General approximation.** Let $\Lambda \subset \mathcal{F}$ be a finite subset, and for each $\mu\in\Lambda$, let $V_\mu \subset V$ be a finite dimensional subspace. Then

$$\mathcal{V}_N \coloneqq \left\{ v(y,x) = \sum_{\mu\in\Lambda} v_\mu(x) P_\mu(y)\,;\ v_\mu \in V_\mu\ \forall \mu\in\Lambda \right\} \subset L^2_\pi(\Gamma;V) \tag{4.1}$$

is a finite dimensional subspace. The Galerkin approximation of u in $\mathcal{V}_N$ is the unique $u_N \in \mathcal{V}_N$ satisfying

$$\int_\Gamma \langle A(y)u_N(y), v(y)\rangle \,\mathrm{d}\pi(y) = \int_\Gamma \int_D f(x)v(y,x)\,\mathrm{d}x\,\mathrm{d}\pi(y) \quad \forall v\in\mathcal{V}_N\,, \tag{4.2}$$

i.e. u_N is the orthogonal projection of u onto $\mathcal{V}_N$ with respect to the scalar product $(\cdot,\cdot)_{\mathcal{A}}$.

Let $\mathcal{A}_N\colon \mathcal{V}_N \to \mathcal{V}_N^*$ be the restriction of $\mathcal{A}$ to $\mathcal{V}_N$, and let f_N be equal to f, interpreted as an element of $\mathcal{V}_N^*$. Then the Galerkin projection u_N is the solution of

$$\mathcal{A}_N u_N = f_N\,. \tag{4.3}$$

For later use, we also define $\bar{\mathcal{A}}_N\colon \mathcal{V}_N \to \mathcal{V}_N^*$ as the restriction of $\mathrm{id}_{L^2_\pi(\Gamma)} \otimes \bar{A}$ to $\mathcal{V}_N$. We will always tacitly assume $0\in\Lambda$.

As in Section 3.3, the coefficients $(u_{N,\mu})_{\mu\in\mathcal{F}}$ of u_N are characterized by $u_{N,\mu} = 0$ for $\mu\in\mathcal{F}\setminus\Lambda$ and

$$\langle \bar{A}u_{N,\mu}, v\rangle + \sum_{m=1}^{\infty} \langle A_m(\beta^m_{\mu_m+1} u_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} u_{N,\mu-\epsilon_m}), v\rangle = \langle f\delta_{\mu 0}, v\rangle \qquad \forall v\in V_\mu \tag{4.4}$$

for all $\mu\in\Lambda$. The coefficent vector $(u_{N,\mu})_{\mu\in\mathcal{F}}$ can be interpreted as an element of

$$\prod_{\mu\in\Lambda} V_\mu \subset \ell^2(\mathcal{F};V)\,, \tag{4.5}$$

and equation (4.3) can be interpreted on this space in place of $\mathcal{V}_N$. We refrain from introducing a different notation for this equivalent formulation.

4.2. **Finite element spaces.** We construct V_μ for $\mu \in \Lambda$ as finite element spaces. In order to ensure a degree of compatibility between these spaces, and uniform shape regularity constants, we assume that the underlying meshes $\mathcal{T}_\mu$ are constructed by some refinement of a given initial mesh $\mathring{\mathcal{T}}$.

Let $\mathring{\mathcal{T}}$ be a conforming simplicial mesh of D. For each element $T \in \mathring{\mathcal{T}}$, we prescribe a sequence of bisections of T into simplices which are uniformly shape regular. Let the set $\mathbb{T}$ consist of all conforming simplicial meshes of D attainable through the prescribed local refinements. For example, if $d = 2$, we may consider triangular meshes generated by newest vertex bisection. We refer to [8] and the references therein for details.

For any $\mathcal{T} \in \mathbb{T}$, we consider $\mathcal{T}$ to be the set of elements of the mesh, and denote the set of faces by $\mathcal{S}$. Interior faces are collected in $\mathcal{S} \cap D$, and the set $\mathcal{S} \cap \partial D$ consists of all boundary faces. Similarly, for any $T \in \mathcal{T}$, the set $\mathcal{S} \cap \partial T$ contains the faces of $\mathcal{T}$ in the boundary of T.

For each $\mu \in \Lambda$, let $\mathcal{T}_\mu \in \mathbb{T}$ with faces $\mathcal{S}_\mu$. In principle, V_μ may be any conforming finite element space on $\mathcal{T}_\mu$. We focus on the simplest such setting and define V_μ to be the space of continuous piecewise affine functions on $\mathcal{T}_\mu$ which vanish on ∂D. In order to avoid the technicalities of boundary approximations, we assume that D is a polytope. These assumptions are not critical to our method, and we make them only to keep the development as clear as possible.

Let $\mathcal{T} \in \mathbb{T}$. For any $T \in \mathcal{T}$ and $S \in \mathcal{S}$, let $h_T \coloneqq \operatorname{diam} T$ and $h_S \coloneqq \operatorname{diam} S$ describe the element and face sizes, and let $\tilde{\omega}_T$ and $\tilde{\omega}_S$ denote the union of all elements of $\mathcal{T}$ sharing at least a vertex with T or S, respectively. We note that the number of these neighborhoods to which any element $T \in \mathcal{T}$ belongs is bounded uniformly on $\mathbb{T}$. Consequently, the Clément interpolation operators $\mathcal{I}_\mu \colon H^1_0(D) \to V_\mu$ satisfy

$$\|\bar{a}^{1/2}(v - \mathcal{I}_\mu v)\|_{L^2(T)} \leq c_{\mathcal{T}} h_T |v|_{V,\tilde{\omega}_T} \qquad \forall T \in \mathcal{T}_\mu \tag{4.6}$$

and

$$\|\bar{a}^{1/2}(v - \mathcal{I}_\mu v)\|_{L^2(S)} \leq c_{\mathcal{S}} h_S^{1/2} |v|_{V,\tilde{\omega}_S} \qquad \forall S \in \mathcal{S}_\mu \tag{4.7}$$

with uniform constants $c_{\mathcal{T}}$ and $c_{\mathcal{S}}$, see *e.g.* [6]. The weight $\bar{a}^{1/2}$ does not affect this standard result since it is uniformly bounded from above and below. Of course, in the case $d = 1$, we have $h_S = 0$, and (4.6) holds for the standard nodal interpolant with T in place of $\tilde{\omega}_T$.

5. Decomposition of the error

5.1. **The residual.** For any $w \in L^2_\pi(\Gamma; V)$, the residual $\mathcal{R}(w) \in L^2_\pi(\Gamma; V^*)$ is

$$\mathcal{R}(w) \coloneqq f - \mathcal{A}w = \mathcal{A}(u - w)\,. \tag{5.1}$$

By the Riesz representation theorem,

$$\|u - w\|_{\mathcal{A}} = \sup_{v \in L^2_\pi(\Gamma;V)} \frac{\langle \mathcal{A}(u-w), v\rangle}{\|v\|_{\mathcal{A}}} = \sup_{v \in L^2_\pi(\Gamma;V)} \frac{\langle \mathcal{R}(w), v\rangle}{\|v\|_{\mathcal{A}}}\,, \tag{5.2}$$

i.e. the error in the energy norm is equal to a dual norm of the residual.

Theorem 5.1. *Let $\mathcal{V}_N \subset L^2_\pi(\Gamma; V)$ be a closed subspace, $w_N \in \mathcal{V}_N$, and let u_N denote the Galerkin projection of u onto $\mathcal{V}_N$. Then for any bounded linear map $Q \colon L^2_\pi(\Gamma; V) \to \mathcal{V}_N$,*

$$\begin{aligned}\|w_N - u\|^2_{\mathcal{A}} \leq{}& \left(\frac{1}{\sqrt{1-\gamma}} \sup_{v \in L^2_\pi(\Gamma;V)} \frac{|\langle \mathcal{R}(w_N), v - Qv\rangle|}{\|v\|_{L^2_\pi(\Gamma;V)}} + c_Q \|w_N - u_N\|_{\mathcal{A}}\right)^2 \\ &+ \|w_N - u_N\|^2_{\mathcal{A}}\,,\end{aligned} \tag{5.3}$$

where c_Q is the operator norm of $\mathrm{id} - Q$ *with respect to the energy norm $\|\cdot\|_{\mathcal{A}}$.*

Proof. Since u_N is the $\mathcal{A}$-orthogonal projection of u onto $\mathcal{V}_N$,

$$\|w_N - u\|_{\mathcal{A}}^2 = \|u_N - u\|_{\mathcal{A}}^2 + \|w_N - u_N\|_{\mathcal{A}}^2 .$$

Using Galerkin orthogonality again, the first term can be written as

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

Furthermore, by Cauchy–Schwarz,

$$|\langle \mathcal{R}(u_N) - \mathcal{R}(w_N), v - v_N\rangle| = |\langle \mathcal{A}(w_N - u_N), v - v_N\rangle| \le \|w_N - u_N\|_{\mathcal{A}} \|v - v_N\|_{\mathcal{A}} .$$

Finally, $\|v\|_{\mathcal{A}}^2 \ge (1-\gamma)\|v\|_{L^2_\pi(\Gamma;V)}^2$ due to (2.3), see [16], and the claim follows with $v_N \coloneqq Qv$. □

5.2. **Projection errors.** We represent the residual $\mathcal{R}(w) \in L^2_\pi(\Gamma; V^*)$ with respect to the tensorized polynomial basis $(P_\mu)_{\mu\in\mathcal{F}}$. If $(w_\mu)_{\mu\in\mathcal{F}} \in \ell^2(\mathcal{F}; V)$ are the coefficients of w, then the coefficients of $\mathcal{R}(w)$ are

$$[\mathcal{R}(w)]_\mu = f\delta_{\mu 0} - \bar{A} w_\mu - \sum_{m=1}^{\infty} A_m(\beta^m_{\mu_m+1} w_{\mu+\epsilon_m} + \beta^m_{\mu_m} w_{\mu-\epsilon_m}) , \quad \mu \in \mathcal{F} . \tag{5.4}$$

Let $\mathcal{V}_N$ be of the form (4.1). For every pair $\mu, \nu \in \Lambda$, let $\Pi^\nu_\mu \colon V_\nu \to V_\mu$ be an arbitrary map, and let $\Pi^\nu_\mu \coloneqq 0$ if either μ or ν is in $\mathcal{F} \setminus \Lambda$. We think of Π^ν_μ as a projection of V_ν into V_μ, but do not require this property rigorously since the spaces need not be nested. For example, Π^ν_μ could be a nodal interpolation operator.

For $w_N \in \mathcal{V}_N$ with coefficients $w_{N,\mu} \in V_\mu$, we approximate $[\mathcal{R}(w_N)]_\mu$ by

$$r_\mu(w_N) \coloneqq f\delta_{\mu 0} - \bar{A} w_{N,\mu} - \sum_{m=1}^{\infty} A_m(\beta^m_{\mu_m+1} \Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} \Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m}) \tag{5.5}$$

for $\mu \in \Lambda$ and $r_\mu(w_N) \coloneqq 0$ for $\mu \in \mathcal{F} \setminus \Lambda$.

Lemma 5.2. *For any $w_N \in \mathcal{V}_N$ and any $\mu \in \mathcal{F}$,*

$$\|r_\mu(w_N) - [\mathcal{R}(w_N)]_\mu\|_{V^*} \le \delta_\mu(w_N) \tag{5.6}$$

for

$$\begin{aligned}\delta_\mu(w_N) \coloneqq \sum_{m=1}^{\infty} \Big\| \frac{a_m}{\bar{a}} \Big\|_{L^\infty(D)} \big(&\beta^m_{\mu_m+1} |\Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} - w_{N,\mu+\epsilon_m}|_{V,\Sigma_m} \\ &+ \beta^m_{\mu_m} |\Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m} - w_{N,\mu-\epsilon_m}|_{V,\Sigma_m} \big) .\end{aligned} \tag{5.7}$$

Proof. The claim follows by triangle inequality using the estimate $\|A_m v\|_{V^*} \le \|a_m/\bar{a}\|_{L^\infty(D)} |v|_{V,\Sigma_m}$ for $v \in V$. □

Although the vector $(r_\mu(w_N))_{\mu\in\mathcal{F}} \in \ell^2(\mathcal{F}; V^*)$ is supported on the finite set Λ, $\delta_\mu(w_N) \ne 0$ also for some $\mu \in \mathcal{F} \setminus \Lambda$. In fact, for any $\mu \in \mathcal{F} \setminus \Lambda$,

$$\delta_\mu(w_N) = \sum_{m=1}^{\infty} \Big\| \frac{a_m}{\bar{a}} \Big\|_{L^\infty(D)} \big(\beta^m_{\mu_m+1} |w_{N,\mu+\epsilon_m}|_{V,\Sigma_m} + \beta^m_{\mu_m} |w_{N,\mu-\epsilon_m}|_{V,\Sigma_m} \big) . \tag{5.8}$$

The infinite series in (5.7) and (5.8) are actually finite sums since $w_{N,\nu} \ne 0$ only for $\nu \in \Lambda$.

Let

$$\operatorname{supp} \Lambda \coloneqq \bigcup_{\mu\in\Lambda} \operatorname{supp} \mu = \{m \in \mathbb{N} ; \, \exists \mu \in \Lambda \colon \mu_m \ne 0\} \tag{5.9}$$

be the set of active dimensions of Λ. For any $m \in \mathbb{N}$, let

$$\partial_m \Lambda \coloneqq [(\Lambda + \epsilon_m) \cup (\Lambda - \epsilon_m)] \cap \mathcal{F} = \{\mu \pm \epsilon_m \in \mathcal{F} ; \, \mu \in \Lambda\} , \tag{5.10}$$

and define

$$\partial^i \Lambda := \bigcup_{m \in \operatorname{supp} \Lambda} \partial_m \Lambda \qquad \text{and} \qquad \partial^o \Lambda := \bigcup_{m \in \mathbb{N} \setminus \operatorname{supp} \Lambda} \partial_m \Lambda \,. \tag{5.11}$$

Then $\delta_\mu(w_N)$ takes the general form (5.7) on the finite set Λ and the simplified form (5.8) on the finite set $\partial^i \Lambda$.

If $\nu \in \partial^o \Lambda$, then there is exactly one $\mu \in \Lambda$ and one $m \in \mathbb{N} \setminus \operatorname{supp} \Lambda$ such that $\nu_n = \mu_n$ for all $n \neq m$, and $\nu_m = 1$, *i.e.* $\nu = \mu + \epsilon_m$. Consequently, $\partial^o \Lambda$ can be decomposed into the finite disjoint union

$$\partial^o \Lambda = \bigsqcup_{\mu \in \Lambda} \partial^o \mu \,, \quad \partial^o \mu := \{\mu + \epsilon_m \,;\, m \in \mathbb{N} \setminus \operatorname{supp} \Lambda\} \,. \tag{5.12}$$

For each $\mu \in \Lambda$, due to (5.8),

$$\begin{aligned} \sum_{\nu \in \partial^o \mu} \delta_\nu(w_N)^2 &= \sum_{m \in \mathbb{N} \setminus \operatorname{supp} \Lambda} \left(\beta_1^m \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(D)} |w_{N,\mu}|_{V,\Sigma_m} \right)^2 \\ &\leq \|w_{N,\mu}\|_V^2 \sum_{m \in \mathbb{N} \setminus \operatorname{supp} \Lambda} \left(\beta_1^m \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(D)} \right)^2 . \end{aligned} \tag{5.13}$$

The infinite sum remaining in the last term of (5.13) is independent of μ, and thus

$$\sum_{\nu \in \partial^o \Lambda} \delta_\nu(w_N)^2 \leq \left(\sum_{\mu \in \Lambda} \|w_{N,\mu}\|_V^2 \right) \sum_{m \in \mathbb{N} \setminus \operatorname{supp} \Lambda} \left(\beta_1^m \left\| \frac{a_m}{\bar{a}} \right\|_{L^\infty(D)} \right)^2 . \tag{5.14}$$

We are thereby led to require that $(\beta_1^m \|a_m/\bar{a}\|_{L^\infty(D)})_{m=1}^\infty$ is in ℓ^2. We also assume that this sequence is arranged in decreasing order.

Remark 5.3. If the functions a_m are locally supported, the estimate (5.13) can be sharpened, and the resulting summability requirement weakened. We assume that $\mathbb{N}$ can be decomposed into countably many disjoint finite sets Δ_l in such a way that there is only a fixed finite overlap between the supports Σ_m of the functions a_m in each level l. Also, we assume that $\beta_1^m \|a_m/\bar{a}\|_{L^\infty(D)} \leq \alpha_l$ for all $m \in \Delta_l$ for some $\alpha_l > 0$. For all $l \in \mathbb{N}$, let

$$n_l(\Lambda) := \max_{x \in D} \sum_{m \in \Delta_l \setminus \operatorname{supp} \Lambda} 1_{\Sigma_m}(x) \,. \tag{5.15}$$

This denotes the maximal overlap within Δ_l of sets Σ_m for currently inactive dimensions m, and is uniformly bounded in l by assumption. By (5.13), for each $\mu \in \Lambda$,

$$\sum_{\nu \in \partial^o \mu} \delta_\nu(w_N)^2 \leq \|w_N\|_V^2 \sum_{l=1}^\infty n_l(\Lambda) \alpha_l^2 \,, \tag{5.16}$$

and thus

$$\sum_{\nu \in \partial^o \Lambda} \delta_\nu(w_N)^2 \leq \left(\sum_{\mu \in \Lambda} \|w_{N,\mu}\|_V^2 \right) \sum_{l=1}^\infty n_l(\Lambda) \alpha_l^2 \,. \tag{5.17}$$

Therefore, we only need to assume $(\alpha_l)_l \in \ell^2$ in order to ensure convergence in (5.17).

6. Error estimator

6.1. **A residual-based error estimator.** We derive a reliable estimator for $r_\mu(w_N)$ from (5.5) following the standard argument from [20, 1, 23], see also [14]. For all $\mu \in \Lambda$, let

$$\sigma_\mu(w_N) \coloneqq \bar{a}\nabla w_{N,\mu} + \sum_{m=1}^{\infty} a_m \nabla(\beta^m_{\mu_m+1} \Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} \Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m}) \,. \tag{6.1}$$

The sum in (6.1) is finite since only finitely many $w_{N,\nu}$ are different from zero. Then the residual $r_\mu(w_N)$ is given by

$$\langle r_\mu(w_N), v\rangle = \int_D f\delta_{\mu 0} v - \sigma_\mu(w_N)\cdot\nabla v \,\mathrm{d}x \,, \quad v \in H^1_0(D) \,. \tag{6.2}$$

For any $T \in \mathcal{T}_\mu$, let

$$\eta_{\mu,T}(w_N) \coloneqq h_T \|\bar{a}^{-1/2}(f\delta_{\mu 0} + \nabla\cdot\sigma_\mu(w_N))\|_{L^2(T)} \,. \tag{6.3}$$

Since $w_{N,\mu}$ and $\Pi^\nu_\mu w_{N,\nu}$ are affine on T, the divergence of $\sigma_\mu(w_N)$ on T is

$$\begin{aligned}\nabla\cdot\sigma_\mu(w_N) = {}& \nabla\bar{a}\cdot\nabla w_{N,\mu} \\ &+ \sum_{m=1}^{\infty} \nabla a_m \cdot \nabla(\beta^m_{\mu_m+1} \Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} \Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m}) \,.\end{aligned} \tag{6.4}$$

Also, for any $S \in \mathcal{S}_\mu$, let

$$\eta_{\mu,S}(w_N) \coloneqq h_S^{1/2} \|\bar{a}^{-1/2}[\![\sigma_\mu(w_N)]\!]\|_{L^2(S)} \,, \tag{6.5}$$

where $[\![\cdot]\!]$ denotes the normal jump over S, *i.e.* if $S = \bar{T}_1 \cap \bar{T}_2$ and n_i is the exterior unit normal to T_i, then

$$[\![\sigma]\!] \coloneqq \sigma|_{T_1}\cdot n_1 + \sigma|_{T_2}\cdot n_2 \,, \tag{6.6}$$

and if $S \in \mathcal{S} \cap \partial D$, then $[\![\sigma]\!] \coloneqq \sigma\cdot n_D$, where n_D is the exterior unit normal to D. These terms combine to

$$\eta_\mu(w_N) \coloneqq \Big(\sum_{T\in\mathcal{T}_\mu} \eta_{\mu,T}(w_N)^2 + \sum_{S\in\mathcal{S}_\mu} \eta_{\mu,S}(w_N)^2 \Big)^{1/2} \,. \tag{6.7}$$

Theorem 6.1. *For all $w_N \in \mathcal{V}_N$, $\mu \in \Lambda$ and $v \in H^1_0(D)$,*

$$|\langle r_\mu(w_N), v - \mathcal{I}_\mu v\rangle| \le c_\eta \eta_\mu(w_N)\|v\|_V \tag{6.8}$$

with a constant c_η depending only on $\bar{a}$ and the shape regularity of $\mathbb{T}$.

Proof. We abbreviate $z \coloneqq v - \mathcal{I}_\mu v$ and $\sigma_\mu \coloneqq \sigma_\mu(w_N)$, and denote by n_T the exterior unit normal to $T \in \mathcal{T}$. Integrating (6.2) by parts on each $T \in \mathcal{T}_\mu$ leads to

$$\begin{aligned}\langle r_\mu(w_N), z\rangle &= \sum_{T\in\mathcal{T}_\mu} \int_T f\delta_{\mu 0} z - \sigma_\mu\cdot\nabla z \,\mathrm{d}x \\ &= \sum_{T\in\mathcal{T}_\mu} \Big[\int_T (f\delta_{\mu 0} + \nabla\cdot\sigma_\mu) z \,\mathrm{d}x - \sum_{S\in\mathcal{S}_\mu\cap\partial T} \int_S \sigma_\mu\cdot n_T z \,\mathrm{d}S \Big] \\ &= \sum_{T\in\mathcal{T}_\mu} \int_T (f\delta_{\mu 0} + \nabla\cdot\sigma_\mu) z \,\mathrm{d}x - \sum_{S\in\mathcal{S}_\mu} \int_S [\![\sigma_\mu]\!] z \,\mathrm{d}S \,.\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$|\langle r_\mu(w_N), z\rangle| \le \sum_{T\in\mathcal{T}_\mu} \|\bar a^{-1/2}(f\delta_{\mu 0} + \nabla\cdot\sigma_\mu)\|_{L^2(T)}\|\bar a^{1/2} z\|_{L^2(T)} + \sum_{S\in\mathcal{S}_\mu} \|\bar a^{-1/2}[\![\sigma_\mu]\!]\|_{L^2(S)}\|\bar a^{1/2} z\|_{L^2(S)}\,,$$

and due to (4.6) and (4.7),

$$\begin{aligned}|\langle r_\mu(w_N), z\rangle| &\le c_{\mathcal{T}} \sum_{T\in\mathcal{T}_\mu} h_T\|\bar a^{-1/2}(f\delta_{\mu 0} + \nabla\cdot\sigma_\mu)\|_{L^2(T)}|v|_{V,\tilde\omega_T} \\ &\quad + c_{\mathcal{S}} \sum_{S\in\mathcal{S}_\mu} h_S^{1/2}\|\bar a^{-1/2}[\![\sigma_\mu]\!]\|_{L^2(S)}|v|_{V,\tilde\omega_S} \\ &\le (c_{\mathcal{T}} + c_{\mathcal{S}}) \sum_{T\in\mathcal{T}_\mu} \Bigl[h_T\|\bar a^{-1/2}(f\delta_{\mu 0} + \nabla\cdot\sigma_\mu)\|_{L^2(T)} \\ &\qquad + \sum_{S\in\mathcal{S}_\mu\cap\partial T} h_S^{1/2}\|\bar a^{-1/2}[\![\sigma_\mu]\!]\|_{L^2(S)}\Bigr]|v|_{V,\tilde\omega_T}\,.\end{aligned}$$

Since the number of domains $\tilde\omega_T$ and $\tilde\omega_S$ that overlap at any point is uniformly bounded on $\mathbb{T}$,

$$\begin{aligned}|\langle r_\mu(w_N), z\rangle| &\le C\biggl(\sum_{T\in\mathcal{T}_\mu}\Bigl[h_T\|\bar a^{-1/2}(f\delta_{\mu 0} + \nabla\cdot\sigma_\mu)\|_{L^2(T)} \\ &\qquad + \sum_{S\in\mathcal{S}_\mu\cap\partial T} h_S^{1/2}\|\bar a^{-1/2}[\![\sigma_\mu]\!]\|_{L^2(S)}\Bigr]^2\biggr)^{1/2}\|v\|_V \\ &\le c_\eta\eta_\mu(w_N)\|v\|_V\,.\end{aligned}$$

□

6.2. **Upper bound of the total error.** We combine Theorems 5.1 and 6.1 to derive an upper bound for the global error in the energy norm.

Let the space $\mathcal{V}_N$ be as in (4.1), and let $Q\colon L^2_\pi(\Gamma;V)\to\mathcal{V}_N$ be given by

$$Qv \coloneqq \sum_{\mu\in\Lambda}(\mathcal{I}_\mu v_\mu)P_\mu\,,\quad v = \sum_{\mu\in\mathcal{F}} v_\mu P_\mu \in L^2_\pi(\Gamma;V)\,. \tag{6.9}$$

Then the constant c_Q from Theorem 5.1 is bounded uniformly on $\mathbb{T}$. We continue to denote by u_N the Galerkin projection of u onto $\mathcal{V}_N$.

Theorem 6.2. *For any $w_N\in\mathcal{V}_N$,*

$$\begin{aligned}\|w_N - u\|^2_{\mathcal{A}} \le \Biggl[&\frac{c_\eta}{\sqrt{1-\gamma}}\Bigl(\sum_{\mu\in\Lambda}\eta_\mu(w_N)^2\Bigr)^{1/2} + \frac{c_Q}{\sqrt{1-\gamma}}\Bigl(\sum_{\mu\in\mathcal{F}}\delta_\mu(w_N)^2\Bigr)^{1/2} \\ &+ c_Q\|w_N - u_N\|_{\mathcal{A}}\Biggr]^2 + \|w_N - u_N\|^2_{\mathcal{A}}\,.\end{aligned} \tag{6.10}$$

Proof. For any $v\in L^2_\pi(\Gamma;V)$, using (6.9),

$$\begin{aligned}\langle\mathcal{R}(w_N), v - Qv\rangle &= \sum_{\mu\in\mathcal{F}}\langle[\mathcal{R}(w_N)]_\mu, v_\mu - \mathcal{I}_\mu v_\mu\rangle \\ &\le \sum_{\mu\in\Lambda}\langle r_\mu(w_N), v_\mu - \mathcal{I}_\mu v_\mu\rangle \\ &\quad + \sum_{\mu\in\mathcal{F}}\|r_\mu(w_N) - \mathcal{R}(w_N)\|_{V^*}\|v_\mu - \mathcal{I}_\mu v_\mu\|_V\,.\end{aligned}$$

Applying Cauchy–Schwarz, Theorem 6.1 and Lemma 5.2 leads to

$$\frac{|\langle \mathcal{R}(w_N), v - Qv\rangle|}{\|v\|_{L^2_\pi(\Gamma;V)}} \leq c_\eta \Big(\sum_{\mu\in\Lambda} \eta_\mu(w_N)^2 \Big)^{1/2} + c_Q \Big(\sum_{\mu\in\mathcal{F}} \delta_\mu(w_N)^2 \Big)^{1/2}.$$

Then the claim follows from Theorem 5.1. □

Remark 6.3. The estimate in (6.10) contains a sum over the infinite set $\mathcal{F}$. As described in Section 5.2, this can be reduced to a finite sum and the sum over $\mathbb{N}\setminus\operatorname{supp}\Lambda$ from (5.14), which can be approximated more easily. See also Remark 5.3 for a similar reduction for locally supported a_m.

7. Refinement strategy

7.1. **A finite element marking strategy.** Following [10, 20], we use a Dörfler strategy to mark elements of $\mathcal{T}_\mu$ for refinement, based on the estimators η_μ. For every $S \in \mathcal{S}_\mu$, let

$$\hat\eta_{\mu,S}(w_N) \coloneqq \Big(\eta_{\mu,S}(w_N)^2 + \frac{1}{d+1} \sum_{T\colon S\in\mathcal{S}\cap\partial T} \eta_{\mu,T}(w_N)^2 \Big)^{1/2}, \tag{7.1}$$

such that

$$\eta_\mu(w_N)^2 = \sum_{S\in\mathcal{S}_\mu} \hat\eta_{\mu,S}(w_N)^2 . \tag{7.2}$$

For a parameter $0 < \vartheta_\eta < 1$, let $\hat{\mathcal{S}}_\eta \subset \bigsqcup_{\mu\in\Lambda}\{\mu\}\times\mathcal{S}_\mu$ be a subset satisfying

$$\sum_{(\mu,S)\in\hat{\mathcal{S}}_\eta} \hat\eta_{\mu,S}(w_N)^2 \geq \vartheta_\eta^2 \sum_{\mu\in\Lambda} \eta_\mu(w_N)^2 . \tag{7.3}$$

Ideally, $\hat{\mathcal{S}}_\eta$ should be chosen as small as possible such that (7.3) holds, but we make no formal restrictions. We collect the elements of $\mathcal{T}_\mu$ for all $\mu\in\Lambda$ in the set $\mathcal{T}_N \coloneqq \bigsqcup_{\mu\in\Lambda}\{\mu\}\times\mathcal{T}_\mu$, where each element $T\in\mathcal{T}_\mu$ is encoded as a pair (μ,T). Let $\hat{\mathcal{T}}_\eta$ be the set of elements in $\mathcal{T}_N$ with at least one face in $\hat{\mathcal{S}}_\eta$. These elements are marked for refinement.

Remark 7.1. It may be useful to mark additional elements of $\mathcal{T}_0$ for refinement based on unresolved components of f; we refer to [20] for details. Similar data oscillation contributions for $\bar a$ and a_m could also be incorporated, but we disregard this in the present work.

7.2. **Localization of projection errors.** We also mark elements for refinement based on the projection errors $\delta_\mu(w_N)$ from (5.7), which we decompose into local contributions. For all $\mu\in\Lambda$, $T\in\mathcal{T}_\mu$ and $m\in\mathbb{N}$, let

$$\zeta^{\mu+\epsilon_m}_{\mu,T,m}(w_N) \coloneqq \Big\| \frac{a_m}{\bar a} \Big\|_{L^\infty(D)} \beta^m_{\mu_m+1} |\Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} - w_{N,\mu+\epsilon_m}|_{V,\Sigma_m\cap T} , \tag{7.4a}$$

$$\zeta^{\mu-\epsilon_m}_{\mu,T,m}(w_N) \coloneqq \Big\| \frac{a_m}{\bar a} \Big\|_{L^\infty(D)} \beta^m_{\mu_m} |\Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m} - w_{N,\mu-\epsilon_m}|_{V,\Sigma_m\cap T} \tag{7.4b}$$

and $\zeta^\nu_{\mu,T,m} \coloneqq 0$ for all other $\nu\in\mathcal{F}$. Then

$$\delta_\mu(w_N) = \sum_{m=1}^\infty \Big(\sum_{T\in\mathcal{T}_\mu} \zeta^{\mu+\epsilon_m}_{\mu,T,m}(w_N)^2 \Big)^{1/2} + \sum_{m\in\operatorname{supp}\mu} \Big(\sum_{T\in\mathcal{T}_\mu} \zeta^{\mu-\epsilon_m}_{\mu,T,m}(w_N)^2 \Big)^{1/2} . \tag{7.5}$$

Unfortunately, (7.5) does not decompose into a sum over $T\in\mathcal{T}_\mu$. However, motivated by (7.5), we define the local error indicators

$$\zeta_{\mu,T}(w_N) \coloneqq \sum_{m=1}^\infty \zeta^{\mu+\epsilon_m}_{\mu,T,m}(w_N) + \sum_{m\in\operatorname{supp}\mu} \zeta^{\mu-\epsilon_m}_{\mu,T,m}(w_N) \tag{7.6}$$

for $\mu \in \Lambda$ and $T \in \mathcal{T}_\mu$. The infinite series in (7.6) is actually a finite sum due to the definition of Π_μ^ν. By triangle inequality,

$$\delta_\mu(w_N) \leq \sum_{T\in\mathcal{T}_\mu} \zeta_{\mu,T}(w_N)\,. \tag{7.7}$$

Instead of a Dörfler marking strategy as in Section 7.1, we suggest to mark elements of $\mathcal{T}_N$ for refinement for which $\zeta_{\mu,T}(w_N)$ exceeds a certain threshold. This threshold can be used also to activate indices in $\mathcal{F}\setminus\Lambda$. Let $0<\vartheta_\zeta<1$, and let

$$\bar\zeta \coloneqq \max\{\zeta_{\mu,T}(w_N)\,;\,(\mu,T)\in\mathcal{T}_N\}\,, \tag{7.8}$$

i.e. the maximum of $\zeta_{\mu,T}(w_N)$ for any $T\in\mathcal{T}_\mu$ and any $\mu\in\Lambda$. We mark the elements

$$\hat{\mathcal{T}}_\zeta \coloneqq \{(\mu,T)\in\mathcal{T}_N\,;\,\zeta_{\mu,T}(w_N)\geq\vartheta_\zeta\bar\zeta\} \tag{7.9}$$

for refinement.

Remark 7.2. One motivation for separating the refinement based on projection errors from the more typical finite element refinement process described in Section 7.1 lies in potential difficulties in applying a Dörfler marking strategy caused by the infinitely many error indicators (7.4), which leads to an infinite sum in the Dörfler property and an infinite set of possible refinements. Furthermore, the separation of these marking steps removes the dependence on the relative scaling of the error indicators η_μ and $\zeta_{\mu,T}$.

7.3. **Selection of new indices.** The marking strategy from Section 7.2 extends to $\mu\in\mathcal{F}\setminus\Lambda$. In this case, there is no need to decompose $\delta_\mu(w_N)$ into local contributions. For a parameter $\vartheta_\delta>0$, we select the new indices

$$\hat\Lambda_\delta \coloneqq \{\mu\in\mathcal{F}\setminus\Lambda\,;\,\delta_\mu(w_N)\geq\vartheta_\delta\bar\zeta\}\,. \tag{7.10}$$

Using the same $\bar\zeta$ in (7.10) as in (7.9) should balance the refinement of the active set Λ with the spatial mesh refinements. To avoid pathological examples and to improve stability, we enforce an upper bound N_δ on the size of $\hat\Lambda_\delta$ proportional to $\#\Lambda$, selecting only the N_δ indices μ with the largest values of $\delta_\mu(w_N)$ if this bound is reached.

Remark 7.3. Although $\hat\Lambda_\delta$ is itself a finite set as a consequence of $(\delta_\mu(w_N))\in\ell^2(\mathcal{F})$, its construction as a subset of the infinite set $\mathcal{F}$ is not trivial. To find the contribution of the infinite set $\partial^o\Lambda$ from (5.12), we note that for fixed $\mu\in\Lambda$ and all $m\in\mathbb{N}\setminus\operatorname{supp}\Lambda$,

$$\delta_{\mu+\epsilon_m}(w_N) = \beta_1^m\left\|\frac{a_m}{\bar a}\right\|_{L^\infty(D)}|w_{N,\mu}|_{V,\Sigma_m} \leq \beta_1^m\left\|\frac{a_m}{\bar a}\right\|_{L^\infty(D)}\|w_{N,\mu}\|_V\,. \tag{7.11}$$

By assumption, a_m are arranged in decreasing order of $\beta_1^m\|a_m/\bar a\|_{L^\infty(D)}$. Therefore, for each $\mu\in\Lambda$, we can iterate through $\mu+\epsilon_m$ in increasing order of m until $\beta_1^m\|a_m/\bar a\|_{L^\infty(D)}\|w_{N,\mu}\|_V$ drops below the threshold $\bar\zeta$, which ensures that no more indices $\mu+\epsilon_m$ will contribute to $\hat\Lambda_\delta$.

7.4. **Local mesh refinements.** Using the marking strategies from Sections 7.1 and 7.2, we refine at least the elements in $\hat{\mathcal{T}}\coloneqq\hat{\mathcal{T}}_\eta\cup\hat{\mathcal{T}}_\zeta$.

Let $\mu\in\Lambda$, $\hat{\mathcal{T}}_\mu\coloneqq\{T\,;\,(\mu,T)\in\hat{\mathcal{T}}\}$, and $\hat{\mathcal{S}}_\mu\coloneqq\{S\,;\,(\mu,S)\in\hat S_\eta\}$. As in [20], to which we refer for details, we refine $\mathcal{T}_\mu$ to an element of $\mathbb{T}$ for which each $T\in\hat{\mathcal{T}}_\mu$ and each $S\in\hat{\mathcal{S}}_\mu$ contain at least one interior node.

We also augment Λ by $\hat\Lambda_\delta$. For all $\mu\in\hat\Lambda_\delta$, $\mathcal{T}_\mu$ is initialized as $\mathring{\mathcal{T}}$ with some uniform refinements applied to it such that the oscillations of the coefficient $a_{\max(\operatorname{supp}\Lambda)}$ are resolved. This refinement procedure is encoded into the routine `Refine` which is based on the refinement indicators $\hat\eta$, $\hat\zeta$ and $\hat\delta$ from the evaluation of the error

estimator in `Error`, see Section 10.4 for further details. We suppress the dependence on the parameters ϑ_η, ϑ_ζ and ϑ_δ. Note that, in order to obtain good starting values for the solver iteration, the vector w_N is prolongated to the refined meshes.

Algorithm 1: `Refine`$[\mathcal{V}_N, w_N, \eta_N, \zeta_N, \delta_N, \bar{\zeta}_N] \mapsto \bar{\mathcal{V}}_N, \bar{w}_N$

$\hat{\mathcal{T}}_\eta \leftarrow \mathrm{construct}(\eta_N)$ `//  residual indicators according to` (7.3)
$\hat{\mathcal{T}}_\zeta \leftarrow \mathrm{construct}(\zeta_N, \bar{\zeta}_N)$ `// projection indicators according to` (7.4)
for $\mu \in \Lambda$ **do**
  $\mathcal{T}_{\bar{\mu}} \leftarrow \mathcal{T}_\mu$
  for $T \in \{T \in \mathcal{T}_\mu \mid (\mu, T) \in \hat{\mathcal{T}}_\eta \cup \hat{\mathcal{T}}_\zeta\}$ **do**
    refine T in $\mathcal{T}_{\bar{\mu}}$
  $\bar{w}_{N,\mu} \coloneqq \Pi^\mu_{\mathcal{T}_{\bar{\mu}}} w_{N,\mu}$
$\hat{\Lambda}_\delta \leftarrow \mathrm{construct}(\delta_N, \bar{\zeta}_N)$
for $\hat{\mu} \in \hat{\Lambda}_\delta$ **do**
  $\bar{w}_{N,\hat{\mu}} \leftarrow$ initialize on $\mathring{\mathcal{T}}$ `// new multi-indices according to` (7.3)
  add $\bar{w}_{N,\hat{\mu}}$ to $\bar{w}_N$

8. An adaptive solver

8.1. **Conjugate gradient iteration.** We use the conjugate gradient method with preconditioner $\bar{\mathcal{A}}_N$ to approximate the Galerkin projection $u_N \in \mathcal{V}_N$. A version of this iteration is given in `PCG`. We note that one application of $\bar{\mathcal{A}}_N^{-1}$ amounts to independent finite element solves in V_μ for all $\mu \in \Lambda$ with the operator $\bar{A}$.

Algorithm 2: `PCG`$[\mathcal{V}_N, w^0, \epsilon] \mapsto w_N, \zeta_N$

$\varrho^0 \coloneqq f - \mathcal{A}_N w^0$
$s^0 \coloneqq \bar{\mathcal{A}}_N^{-1} \varrho^0$
$v^0 \coloneqq s^0$
$\zeta^0 \coloneqq \langle \varrho^0, s^0 \rangle$
for $i \in \mathbb{N}$ **do**
  if $\zeta^{i-1} \leq \epsilon^2$ **then**
    return $w_N \coloneqq w^{i-1}$, $\zeta_N \coloneqq \zeta^{i-1}$
  $z^{i-1} \coloneqq \mathcal{A}_N v^{i-1}$
  $\alpha^{i-1} \coloneqq \langle z^{i-1}, v^{i-1} \rangle$
  $w^i \coloneqq w^{i-1} + \frac{\zeta^{i-1}}{\alpha^{i-1}} v^{i-1}$
  $\varrho^i \coloneqq \varrho^{i-1} - \frac{\zeta^{i-1}}{\alpha^{i-1}} z^{i-1}$
  $s^i \coloneqq \bar{\mathcal{A}}_N^{-1} \varrho^i$
  $\zeta^i \coloneqq \langle \varrho^i, s^i \rangle$
  $v^i \coloneqq s^i + \frac{\zeta^i}{\zeta^{i-1}} v^{i-1}$

Lemma 8.1. *For all $i \in \mathbb{N}_0$,*

$$\frac{1}{1+\gamma} \zeta^i \leq \|w^i - u_N\|^2_{\mathcal{A}} \leq \frac{1}{1-\gamma} \zeta^i . \tag{8.1}$$

Proof. By definition, using the residuals $\varrho^i = \mathcal{A}_N(w^i - u_N)$,

$$\|w^i - u_N\|_{\mathcal{A}}^2 = \langle \varrho^i, \mathcal{A}_N^{-1}\varrho^i\rangle \qquad \text{and} \qquad \zeta^i = \langle \varrho^i, \bar{\mathcal{A}}_N^{-1}\varrho^i\rangle\,.$$

The claim follows using $(1+\gamma)^{-1}\bar{\mathcal{A}}_N^{-1} \le \mathcal{A}_N^{-1} \le (1-\gamma)^{-1}\bar{\mathcal{A}}_N^{-1}$ in the sense of self-adjoint operators, see [16, Prop. 2.10]. □

Theorem 8.2. *The method* $\texttt{PCG}[\mathcal{V}_N, w^0, \epsilon]$ *returns* w_N *and* ζ_N *satisfying*

$$\|w_N - u_N\|_{\mathcal{A}} \le \sqrt{\frac{\zeta_N}{1-\gamma}} \le \frac{\epsilon}{\sqrt{1-\gamma}}\,. \tag{8.2}$$

At most

$$1 + \left\lfloor \frac{\log(2\epsilon^{-1}\sqrt{1+\gamma}\,\|w^0 - u_N\|_{\mathcal{A}})}{\log(\gamma^{-1} + \sqrt{\gamma^{-2}-1})} \right\rfloor \tag{8.3}$$

iterations are performed.

Proof. Equation (8.2) follows from Lemma 8.1 and the termination criterion of PCG. By [18, Satz 9.4.14], for all $i \in \mathbb{N}_0$,

$$\|w^i - u_N\|_{\mathcal{A}} \le 2\frac{q^i}{1+q^{2i}}\,\|w^0 - u_N\|_{\mathcal{A}}\,, \quad q = \frac{\gamma}{1+\sqrt{1-\gamma^2}}\,,$$

see also [16, Thm. 3.7]. Let the final iterate be $w_N = w^j$. Then provided $j \ge 1$, again using Lemma 8.1,

$$\|w^{j-1} - u_N\|_{\mathcal{A}} \ge \sqrt{\frac{\zeta^{j-1}}{1+\gamma}} \ge \frac{\epsilon}{\sqrt{1+\gamma}}\,.$$

Consequently,

$$\epsilon \le \sqrt{1+\gamma}\|w^{j-1} - u_N\|_{\mathcal{A}} \le 2\sqrt{1+\gamma}\,\|w^0 - u_N\|_{\mathcal{A}} q^{j-1}\,,$$

and solving for j leads to (8.3). □

Remark 8.3. The exact computation of $z^i = \mathcal{A}_N v^i$ requires the assembly of matrix representations of A_m with different domains and codomains, *i.e.* as maps from V_ν to V_μ with $\nu \ne \mu$. To circumvent this costly procedure, we suggest first projecting onto the codomain. Let $(v^i_\mu)_{\mu\in\Lambda}$ be the coefficients of v^i, with $v^i_\mu \in V_\mu$. We approximate z^i by the element of $\mathcal{V}_N^*$ with coefficients

$$z^i_\mu \coloneqq \bar{A}v^i_\mu + \sum_{m\in\operatorname{supp}\Lambda} A_m(\beta^m_{\mu_m+1}\Pi_\mu^{\mu+\epsilon_m} v^i_{\mu+\epsilon_m} + \beta^m_{\mu_m}\Pi_\mu^{\mu-\epsilon_m} v^i_{\mu-\epsilon_m})\,, \tag{8.4}$$

interpreted as an element of V_μ^*, for $\mu \in \Lambda$, and similarly for the application of $\mathcal{A}_N$ in the definition of ϱ^0. Let $\mathcal{A}_N^\Pi$ denote this approximation of $\mathcal{A}_N$. It follows by induction that, if $\mathcal{A}_N$ is replaced by $\mathcal{A}_N^\Pi$ in PCG, then $\varrho^i = f - \mathcal{A}_N^\Pi w^i$, *i.e.* the residuals can still be computed recursively. Although we expect the effects of this approximation on the convergence of the conjugate gradient method to be small, especially if only a few iterations are performed, and it has been used successfully in [14, 15], it does break the symmetry of the operator $\mathcal{A}_N$, and Theorem 8.2 is no longer guaranteed to hold.

8.2. **An adaptive solver.** We combine the conjugate gradient iteration from Section 8.1 with the error estimate in Theorem 6.2 and the refinement strategy from Section 7 to construct an adaptive solver for (2.11). Let

$$\begin{aligned}\texttt{Error}[w_N,\zeta_N] \coloneqq \Bigg[&\frac{\bar{c}_\eta}{\sqrt{1-\gamma}}\Bigg(\sum_{\mu\in\Lambda}\eta_\mu(w_N)^2\Bigg)^{1/2} + \frac{\bar{c}_Q}{\sqrt{1-\gamma}}\Bigg(\sum_{\mu\in\mathcal{F}}\delta_\mu(w_N)^2\Bigg)^{1/2} \\ &+ \bar{c}_Q\sqrt{\frac{\zeta_N}{1-\gamma}}\Bigg]^2 + \frac{\zeta_N}{1-\gamma}\end{aligned} \tag{8.5}$$

for $w_N \in \mathcal{V}_N$ and $\zeta_N \geq 0$, with constants $\bar{c}_Q \geq c_Q$ and $\bar{c}_\eta \geq c_\eta$.

Algorithm 3: $\texttt{Solve}[\epsilon, \mathcal{V}_N^1, w_N^0, \xi^0] \mapsto u_\epsilon$

for $i \in \mathbb{N}$ **do**
- $\tilde{w}_N^i, \zeta_N^i \coloneqq \texttt{PCG}[\mathcal{V}_N^i, w_N^{i-1}, \chi\xi_N^{i-1}]$
- $\xi_N^i, \eta_N^i, \delta_N^i \coloneqq \texttt{Error}[w_N^i, \zeta_N^i]$
- **if** $\xi_N^i \leq \epsilon$ **then**
  - **return** $u_\epsilon \coloneqq w_N^i$
- $\mathcal{V}_N^{i+1}, w_N^i \coloneqq \texttt{Refine}[\mathcal{V}_N^i, \tilde{w}_N^i, \eta_N^i, \delta_N^i]$

The total error of the a posteriori error estimator is denoted by ξ_N while the residual and the projection parts are denoted by η_N and δ_N, respectively. We assume that the inputs of `Solve` satisfy $\epsilon > 0$, $w_N^0 \in \mathcal{V}_N^1$ for $\mathcal{V}_N^1$ of the form (4.1) with finite element spaces V_μ as in Section 4.2, and $\|w_N^0 - u\|_{\mathcal{A}} \leq \xi^0$. For example, $\mathcal{V}_N^1$ may have a single active coefficient $0 \in \mathcal{F}$, with $\mathcal{T}_0 = \mathring{\mathcal{T}}$, and $w_N^0 \coloneqq 0$. In this case, we can set $\xi^0 \coloneqq (1-\gamma)^{-1/2}\|f\|_{V^*}$.

Due to Theorems 6.2 and 8.2, if `Solve` terminates, then $\|u_\epsilon - u\|_{\mathcal{A}} \leq \epsilon$. However, convergence of the solver is not proven. It is likely that the parameter $0 < \chi < 1$ must be chosen sufficiently small.

Further details of the implementation are provided in Sections 10.2 and 10.4.

9. Extensions to more general problems

The proposed approach can be used for Galerkin (primal and mixed) FEM for any linear elliptic system in divergence form, for instance the Helmholtz equation (in which case the perturbation of the operator $A(y)$ must be such that it "stays away" from resonance), Reissner–Mindlin plate models and several other equations. The general algorithmic structure remains unchanged and can thus be understood as a generic numerical approach.

9.1. **Inhomogeneous boundary conditions.** If inhomogeneous Dirichlet boundary conditions $u = g$ are imposed on ∂D for a $g \in H^{1/2}(\partial D)$, then the solution u is in $\bar{g} + L^2_\pi(\Gamma; V)$ for any extension $\bar{g}$ of g to $H^1(D)$, rather than in $L^2_\pi(\Gamma; V)$. All of the above goes through in this setting if u and its approximations u_N and w_N are taken to be in the affine space $\bar{g} + L^2_\pi(\Gamma; V)$ or suitable subspaces of the form $\bar{g} + \mathcal{V}_N$.

However, in practice approximate solutions lie in $\tilde{g} + \mathcal{V}_N$ for an extension $\tilde{g}$ of *e.g.* a piecewise linear interpolant of g on the exterior faces of $\mathcal{T}_0$. This nonconformity introduces an additional error that is not captured by the error estimator. We refer to [3] for a priori and a posteriori estimates of the error induced by this additional approximation.

The adaptive method also extends to other typed of boundary conditions. For example, if Neumann boundary conditions are imposed on a part or all of ∂D, the space V and right-hand side f are modified in the usual way. The right-hand side remains deterministic if $a_m = 0$ on the Neumann part of the boundary for all $m \in \mathbb{N}$. Neumann boundary data enters the residual error estimator in $\eta_{\mu,S}$ from (6.5) for $\mu = 0$ and faces S in the appropriate part of ∂D.

9.2. **Stochastic forcing.** In order to handle right-hand sides f that depend on the parameter $y \in \Gamma$, it is necessary to assume that the expansion

$$f(y) = \sum_{\mu\in\mathcal{F}} f_\mu P_\mu(y) \tag{9.1}$$

with $f_\mu \in V^*$ is known. Replacing $f\delta_{\mu 0}$ by f_μ in (5.4) for all $\mu \in \mathcal{F}$ and in (5.5) for all $\mu \in \mathcal{F} \setminus \Lambda$, estimate (5.6) becomes

$$\|r_\mu(w_N) - [\mathcal{R}(w_N)]_\mu\|_{V^*} \le \delta_\mu(w_N) + \|f_\mu\|_{V^*} \tag{9.2}$$

if $\mu \in \mathcal{F} \setminus \Lambda$. The residual error estimator only needs to be modified by replacing $f\delta_{\mu 0}$ by f_μ in (6.3), but the bound on the total error in Theorem 6.2 becomes

$$\begin{aligned}\|w_N - u\|_{\mathcal{A}}^2 \le \Bigg[&\frac{c_\eta}{\sqrt{1-\gamma}}\Bigg(\sum_{\mu\in\Lambda}\eta_\mu(w_N)^2\Bigg)^{1/2} + \frac{c_Q}{\sqrt{1-\gamma}}\Bigg(\sum_{\mu\in\mathcal{F}}\delta_\mu(w_N)^2\Bigg)^{1/2} \\ &+ \frac{c_Q}{\sqrt{1-\gamma}}\Bigg(\sum_{\mu\in\mathcal{F}\setminus\Lambda}\|f_\mu\|_{V^*}^2\Bigg)^{1/2} + c_Q\|w_N - u_N\|_{\mathcal{A}}\Bigg]^2 \\ &+ \|w_N - u_N\|_{\mathcal{A}}^2 .\end{aligned} \tag{9.3}$$

for any $w_N \in \mathcal{V}_N$. Thus some computable bounds on the norms $\|f_\mu\|_{V^*}$ must be available, and these should be used in Section 7.3 similarly to $\delta_\mu(w_N)$ to select new indices $\mu \in \mathcal{F} \setminus \Lambda$ in the refinement process.

For example, if $f(y)$ has the form $f - A(y)z(y)$ for a given

$$z(y,x) = \sum_{\mu\in\mathcal{F}} z_\mu(x)P_\mu(y)\,, \tag{9.4}$$

coming *e.g.* from boundary data or from a domain decomposition method, then

$$f_\mu = f\delta_{\mu 0} - \bar{A}z_\mu - \sum_{m=1}^{\infty} A_m(\beta^m_{\mu_m+1}z_{\mu+\epsilon_m} + \beta^m_{\mu_m}z_{\mu-\epsilon_m}) \tag{9.5}$$

for all $\mu \in \mathcal{F}$, and the series in (9.5) is only a finite sum if the expansion (9.4) is finite. The norm of f_μ can be estimated by triangle inequality.

Alternatively, as in the discussion of inhomogeneous Dirichlet boundary conditions, the equation can be formulated as $\mathcal{A}(z+u) = f$ with a deterministic f, and a solution $z+u$ in the affine space $z + L^2_\pi(\Gamma; V)$. This circumvents the need for the extra terms in (9.2) and (9.3), and leads to a sharper bound on the error since it does not use the triangle inequality to estimate the norm of (9.5).

9.3. **Linearized elasticity.**

9.3.1. *Navier–Lamé equations.* Consider an elastic body $D \subset \mathbb{R}^2$ with boundary $\partial D = \partial D_D \cup \partial D_N$ which is loaded by applied volume forces $f \in L^2(D;\mathbb{R}^2)$ and surface traction $g \in L^2(\partial D_N;\mathbb{R}^2)$ on some relatively open part ∂D_N of the boundary ∂D with exterior unit normal n. The elastic body is supported on the Dirichlet part $\partial D_D \coloneqq \partial D \setminus \partial D_N$ where the displacement field is prescribed by $u_D \in V \coloneqq H^1(D;\mathbb{R}^2)$. In order to obtain uniqueness and existence of weak solutions, ∂D_D is assumed to be closed and to have positive surface measure.

Within the theory of linearized elasticity, the material behavior is modeled with the positive Lamé parameters λ and μ which define the fourth-order isotropic material tensor $\mathbb{C}$. With the displacement field $u \in H^1(D;\mathbb{R}^2)$, the stress tensor $\sigma \in L^2(D;\mathbb{R}^{2\times 2}_{sym})$ is a linear function of the linear Green strain $\varepsilon(u) \coloneqq (Du + (Du)^T)/2$. Here, $Du = (u_{j,k})_{j,k=1,2}$ is the matrix of all first-order partial derivatives $u_{j,k} \coloneqq \partial u_j/\partial x_k$. With an isotropic but possibly inhomogeneous constitutive law with random coefficient λ of the form (2.2), the stress is defined by

$$\sigma(u;\lambda,\mu) \coloneqq \mathbb{C}\varepsilon(u) \coloneqq \lambda \operatorname{tr}(\varepsilon(u))I + 2\mu\varepsilon(u). \tag{9.6}$$

We assume $\bar\lambda, \lambda_m \in W^{1,\infty}(D)$ such that (2.3) holds. The Green strain and the stress tensor are symmetric 2×2 matrices. A different set of material parameters is given by the Young modulus $E > 0$ and the Poisson ratio $0 < \nu < 1/2$. These parameters are related to the Lamé parameters by

$$\lambda = \frac{E\nu}{(1+\nu)(1-2\nu)} \qquad \text{and} \qquad \mu = \frac{E}{2(1+\nu)}. \tag{9.7}$$

The boundary value problem of the Navier–Lamé equations in linear elasticity reads

$$\begin{cases} -\operatorname{div}\sigma(u;\lambda,\mu) = f & \text{in } D, \\ \sigma(u;\lambda,\mu)n = g & \text{on } \partial D_N, \\ u = u_D & \text{on } \partial D_D. \end{cases} \tag{9.8}$$

The variational formulation of (9.8) is set in the space $V \coloneqq H^1_0(D;\mathbb{R}^2)$ which we endow with the scalar product

$$(w,v)_V \coloneqq \int_D \sigma(w;\bar\lambda,\mu) : \varepsilon(v)\,\mathrm{d}x \tag{9.9}$$

where $\bar{\mathbb{C}}\varepsilon(w) = \sigma(w;\bar\lambda,\mu)$ and $\|\cdot\|_V$ is the induced norm. Note that for $u \in V$

$$\|u\|_V = \|\bar{\mathbb{C}}^{1/2}\varepsilon(u)\|_{L^2(D)}. \tag{9.10}$$

9.3.2. *Parametric Operator.* In analogy to Section 2.1, we define the operator

$$\bar A\colon V \to V^*, \quad v \mapsto -\operatorname{div}\sigma(v;\bar\lambda,\mu) \tag{9.11}$$

which is boundedly invertible. Moreover, we define the bounded linear maps

$$A_m\colon V \to V^*, \quad v \mapsto -\operatorname{div}\sigma(v;\lambda_m,0), \quad m \in \mathbb{N}, \tag{9.12}$$

through which we can express

$$A(y)\colon V \to V^*, \quad v \mapsto -\operatorname{div}\sigma(v;\lambda(y),\mu), \quad y \in \Gamma, \tag{9.13}$$

as

$$A(y) = \bar A + \sum_{m=1}^{\infty} y_m A_m, \quad y \in \Gamma, \tag{9.14}$$

with unconditional convergence in $\mathcal{L}(V,V^*)$. Thus, equation (9.8) is expressed succinctly as

$$A(y)u(y) = f, \quad y \in \Gamma. \tag{9.15}$$

9.3.3. *Residual estimator.* The residual error estimator is defined in the same way as in Section 6.1 for the Poisson model problem. Assume $\mathbb{C}_m\varepsilon(w) = \sigma(w;\lambda_m,\mu)$. For all $\mu \in \Lambda$, define

$$\sigma_\mu(w_N) \coloneqq \bar{\mathbb{C}}\varepsilon(w_{N,\mu}) + \sum_{m=1}^{\infty} \mathbb{C}_m\varepsilon(\beta^m_{\mu_m+1} \Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} \Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m}) \tag{9.16}$$

$$= \sigma(w_{N,\mu};\bar{\lambda},\mu) + \sum_{m=1}^{\infty} \sigma(\beta^m_{\mu_m+1} \Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} \Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m};\lambda_m,0)\,. \tag{9.17}$$

Then, as in (6.2), the approximate residual $r_\mu(w_N)$ reads

$$\langle r_\mu(w_N), v\rangle = \int_D f\cdot\delta_{\mu 0} v - \sigma_\mu(w_N) : \varepsilon(v)\,\mathrm{d}x\,, \quad v \in H^1_0(D;\mathbb{R}^2)\,. \tag{9.18}$$

For any $T \in \mathcal{T}_\mu$, let

$$\eta_{\mu,T}(w_N) \coloneqq h_T \|\bar{\mathbb{C}}^{-1/2}(f\delta_{\mu 0} + \operatorname{div}\sigma_\mu(w_N))\|_{L^2(T)} \tag{9.19}$$

and note that on T the divergence of $\sigma_\mu(w_N)$ is

$$\begin{aligned}\operatorname{div}\sigma_\mu(w_N) = {}& \operatorname{div}\bar{\mathbb{C}}\varepsilon(w_{N,\mu}) \\ &+ \sum_{m=1}^{\infty} \operatorname{div}\mathbb{C}_m\varepsilon(\beta^m_{\mu_m+1} \Pi^{\mu+\epsilon_m}_\mu w_{N,\mu+\epsilon_m} + \beta^m_{\mu_m} \Pi^{\mu-\epsilon_m}_\mu w_{N,\mu-\epsilon_m})\,.\end{aligned} \tag{9.20}$$

In particular, for any affine $w_{N,\mu}$ on T,

$$\operatorname{div}\mathbb{C}_m\varepsilon(w_{N,\mu}) = (\nabla\lambda)\operatorname{tr}(\varepsilon(w_{N,\mu}))\,. \tag{9.21}$$

Moreover, for any $S \in \mathcal{S}_\mu$, let

$$\eta_{\mu,S}(w_N) \coloneqq h_S^{1/2} \|\bar{\mathbb{C}}^{-1/2}[\![\sigma_\mu(w_N)]\!]\|_{L^2(S)}\,. \tag{9.22}$$

For faces $S \in \mathcal{S}_\mu$ on the Neumann boundary, *i.e.* $|S \cap \partial D_N| > 0$, the jump term in (9.22) is defined by $g + \sigma(w_N)\cdot n$. With this, the residual error estimator $\eta_\mu(w_N)$ is defined as in (6.7).

10. Implementation

The implementation is based on a new open-source framework for numerical methods in uncertainty quantification [11]. It relies on the public domain FEM package `FEniCS` [19]. The aim of the framework is to provide means to easily test novel numerical methods. Moreover, it enables the comparison of different existing methods within a single programming environment. Although the framework is in an early state of development, all common components for stochastic Galerkin FEM (SGFEM) are readily available. A more in-depth review regarding the efficient and flexible application of SGFEM which also elaborates on aspects of the software design is in preparation.

For the adaptive solver described in Section 8, some specific requirements had to be implemented. In particular, the management of different meshes for different active $\mu \in \Lambda$ and the transfer of vectors between them is required.

In the following, we give a brief overview of some key aspects of the implementation which was used to run the numerical experiments of Section 11. The focus lies on the ingredients of the residual error estimator of Section 6 and the adaptive solver of Section 8.

Since some quantities are very difficult to compute in practice, simplifying assumptions were made where appropriate or necessary.

10.1. **Environment.** The implementation of the stochastic Galerkin discretization is mainly carried out in the programming language Python with an embedded C++ library for the FEM part. The nature of Python allows for rapid prototyping as well as for convenient and efficient development of new numerical techniques. The object-oriented software design enables the separation of generic components which allows for the reuse of tested code. In order to ensure correctness, unit tests are part of the code in many cases. Abstraction is used whenever possible to facilitate a generic implementation of algorithms without having to resort to specific data structures.

For the FEM part, the framework `FEniCS` provides the management of meshes, the construction of discrete spaces and a versatile definition and evaluation of weak forms. In fact, the interface between the stochastic and the discrete discretisation is such that different equations can easily be implemented and tested. A large variety of finite element spaces are available and can be chosen according to the problem at hand. Moreover, coefficients may be inhomogeneous or anisotropic and the iterative solution of nonlinear problems is supported in a general interface.

10.2. **Computation of the discrete operator.** In reference to (4.1), a vector $w_N \in \mathcal{V}_N$ includes components $w_{N,\mu} \in V_\mu$ for $\mu \in \Lambda$. These are based on different discretisation meshes $\mathcal{T}_\mu$, respectively. The meshes and spatial vectors of all $w_{N,\mu}$ in some vector w_N are managed in a dictionary data structure. A projection between two meshes $\Pi_\mu^\nu \colon V_\nu \to V_\mu$ is provided by nodal interpolation or by L^2 projection. For interpolation, a search tree structure is used in case of different meshes to efficiently identify the cells of interpolation points. The projection between meshes is employed in the evaluation of the discrete operator (4.4) and in the evaluation of the projection error (7.4a) and (7.4b). The evaluation of the discrete system (4.4) in practice bears the difficulty that the bilinear form cannot be computed in the original form. This is due to the different meshes for trial and test spaces which occur simultaneously on the left-hand side.

Different approaches to remedy this problem are possible. Recall the discrete operator equation

$$\langle \bar{A} u_{N,\mu}, v\rangle + \sum_{m=1}^{\infty} \langle \beta_{\mu_m+1}^m A_m u_{N,\mu+\epsilon_m}, v\rangle + \langle \beta_{\mu_m}^m A_m u_{N,\mu-\epsilon_m}, v\rangle = \langle f\delta_{\mu 0}, v\rangle, \quad v \in V_\mu. \tag{10.1}$$

Denote by $N_\mu = \dim V_\mu$ the dimension of the discrete space associated with multi-index $\mu \in \Lambda$ and set $N = \sum_{\mu\in\Lambda} N_\mu$. Written as the product of a block matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ with a block vector $\mathbf{u} \in \mathbb{R}^N$, the discrete system $\mathbf{Au} = \mathbf{f}$ takes the form

$$\begin{bmatrix} \ddots & & \vdots & & & \iddots \\ & \mathbf{A}_{\mu-\epsilon_m} & \cdots & \mathbf{C}^{\mu}_{\mu-\epsilon_m} & \cdots & 0 & \\ & \vdots & & \vdots & & \vdots & \\ \cdots & \mathbf{B}^{\mu-\epsilon_m}_{\mu} & \cdots & \mathbf{A}_{\mu} & \cdots & \mathbf{C}^{\mu+\epsilon_m}_{\mu} & \cdots \\ & \vdots & & \vdots & & \vdots & \\ & 0 & \cdots & \mathbf{B}^{\mu}_{\mu+\epsilon_m} & \cdots & \mathbf{A}_{\mu+\epsilon_m} & \\ \iddots & & \vdots & & & & \ddots \end{bmatrix} \begin{bmatrix} \vdots \\ \mathbf{u}_{\mu-\epsilon_m} \\ \vdots \\ \mathbf{u}_{\mu} \\ \vdots \\ \mathbf{u}_{\mu+\epsilon_m} \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \mathbf{f}_{\mu-\epsilon_m} \\ \vdots \\ \mathbf{f}_{\mu} \\ \vdots \\ \mathbf{f}_{\mu+\epsilon_m} \\ \vdots \end{bmatrix}. \tag{10.2}$$

Here, the matrices $\mathbf{A}_\mu$, $\mathbf{B}_\mu^{\mu-\epsilon_m}$ and $\mathbf{C}_\mu^{\mu+\epsilon_m}$ are defined by

$$\begin{aligned}
[\mathbf{A}_\mu]_{ij} &= \langle \bar{A}\Phi_i^\mu, \Phi_j^\mu\rangle, \quad \mathbf{A} \in \mathbb{R}^{N_\mu \times N_\mu},\\
\left[\mathbf{B}_\mu^{\mu-\epsilon_m}\right]_{ij} &= \beta_{\mu_m}^m \langle A_m \Phi_i^\mu, \Phi_j^{\mu-\epsilon_m}\rangle, \quad \mathbf{B} \in \mathbb{R}^{N_\mu \times N_{\mu-\epsilon_m}},\\
\left[\mathbf{C}_\mu^{\mu+\epsilon_m}\right]_{ij} &= \beta_{\mu_m+1}^m \langle A_m \Phi_i^\mu, \Phi_j^{\mu+\epsilon_m}\rangle, \quad \mathbf{C} \in \mathbb{R}^{N_\mu \times N_{\mu+\epsilon_m}},
\end{aligned}$$

where Φ_i^ν is the basis of V_ν for $\nu \in \{\mu, \mu \pm \epsilon_m\}$. The full Galerkin matrix is symmetric, since

$$\left[\mathbf{B}_\mu^{\mu-\epsilon_m}\right]_{ji} = \beta_{\mu_m}^m \langle A_m \Phi_j^\mu, \Phi_i^{\mu-\epsilon_m}\rangle = \beta_{\mu_m}^m \langle A_m \Phi_i^{\mu-\epsilon_m}, \Phi_j^\mu\rangle = \left[\mathbf{C}_{\mu-\epsilon_M}^\mu\right]_{ij}.$$

However, the computation of the matrices $\mathbf{B}_\mu^{\mu-\epsilon_m}$ and $\mathbf{C}_\mu^{\mu+\epsilon_m}$ is not feasible efficiently as it involves integration over basis functions defined on incompatible meshes. In the used implementation, these matrices can be approximated in different ways.

The most efficient approach is to assume

$$\tilde{\mathbf{B}}_\mu^{\mu-\epsilon_m} = \mathbf{A}_m^\mu \Pi_\mu^{\mu-\epsilon_m} \quad \text{and} \quad \tilde{\mathbf{C}}_\mu^{\mu+\epsilon_m} = \mathbf{A}_m^\mu \Pi_\mu^{\mu+\epsilon_m}, \tag{10.3}$$

which coincides with (8.4), see also Remark 8.3. Since the resulting consistency error, which basically requires interpolation error estimates for arbitrary (non-hierarchical) discrete spaces, can not be quantified without further restrictive assumptions, we choose a more elaborate approach which requires the construction of a union of meshes. Denote by $\mathcal{T}_{\tilde{\mu}}$ the union of $\mathcal{T}_\mu$ and $\mathcal{T}_{\mu\mp\epsilon_m}$,[1] *i.e.* $\mathcal{T}_{\tilde{\mu}} \subseteq \mathcal{T}_\mu$ and $\mathcal{T}_{\tilde{\mu}} \subseteq \mathcal{T}_{\mu\mp\epsilon_m}$ which means that for each $\tilde{T} \in \mathcal{T}_{\tilde{\mu}}$ there exists some $T \in \mathcal{T}_\mu$ (or $T \in \mathcal{T}_{\mu\mp\epsilon_m}$, respectively) such that $\tilde{T} \subseteq T$. Assume the approximations

$$\tilde{\mathbf{B}}_\mu^{\mu-\epsilon_m} = \Pi_\mu^{\tilde{\mu}} \mathbf{A}_m^{\tilde{\mu}} \Pi_{\tilde{\mu}}^{\mu-\epsilon_m} \quad \text{and} \quad \tilde{\mathbf{C}}_\mu^{\mu+\epsilon_m} = \Pi_\mu^{\tilde{\mu}} \mathbf{A}_m^{\tilde{\mu}} \Pi_{\tilde{\mu}}^{\mu+\epsilon_m}. \tag{10.4}$$

It is shown in [21] that this approach does not adversely affect the order of convergence of the Galerkin method. Note however that the computation of the union of the meshes $\mathcal{T}_{\tilde{\mu}}$ and the assembly of $\mathbf{A}_m^{\tilde{\mu}}$ can be computationally expensive.

10.3. **Inhomogeneous Dirichlet boundary conditions.** For the proposed solver in Section 8.1, the incorporation of Dirichlet boundary conditions in the spatially discretised form has to be carried out in such a way that the symmetry of the operator is retained. Furthermore, if the Dirichlet boundary conditions are inhomogeneous, care has to be taken that the right hand side is appropriately modified.

Let $\mathbf{A}$ be the discrete (assembled) form of the operator A and denote by $\mathbf{f}$ and $\mathbf{g}$ the discrete right-hand side and the boundary values. $\mathbf{A}$ and $\mathbf{f}$ are assumed to be assembled with no boundary conditions applied yet. Typically, FEM codes modify the operator and right-hand side, resulting in a modified matrix $\hat{\mathbf{A}}$ and vector $\hat{\mathbf{f}}$ such that the system

$$\hat{\mathbf{A}}\mathbf{u} = \hat{\mathbf{f}} \tag{10.5}$$

can be solved at once for the discrete solution $\mathbf{u}$ and the boundary conditions are fulfilled on the Dirichlet nodes.

To describe the action of the incorporation of Dirichlet boundary conditions into the discrete equations, let $\mathbf{I}_D$ denote a projection (nodal interpolation) onto the Dirichlet nodes and $\mathbf{I}_I$ a projection onto the inner and Neumann nodes. Note that $\mathbf{I}_D + \mathbf{I}_I = \mathbf{I}$ and that $\mathbf{I}_D\mathbf{u} + \mathbf{I}_I\mathbf{u} = \mathbf{u}$. In many FEM codes, the inclusion of the

[1] Note that $\mathcal{T}_{\tilde{\mu}}$ is a slight abuse of notation since $\tilde{\mu}$ does not indicate a valid multi-index. However, using $\mathcal{T}_\mu$ instead would make the notation for projections too unwieldy.

Dirichlet boundary conditions (*e.g.* `assemble_system` in `FEniCS`) transforms the discrete system into

$$\begin{aligned}\hat{\mathbf{A}} &= \mathbf{I}_I \mathbf{A} \mathbf{I}_I + \mathbf{D}\mathbf{I}_D \\ \hat{\mathbf{f}} &= \mathbf{I}_I \mathbf{f} + \mathbf{D}\mathbf{I}_D \mathbf{g} - \mathbf{I}_I \mathbf{A} \mathbf{I}_D \mathbf{g},\end{aligned} \tag{10.6}$$

where $\mathbf{D}$ is some non-singular, diagonal and positive definite matrix.[2] In `FEniCS`, for example, $[\mathbf{D}]_{ii}$ equals the number of elements that contain node i, due to the implementation of the assembly process. In other FEM codes it may also be the identity. Note that symmetry and positive definiteness of the discrete operator is kept in this transformation.

In the stochastic case the operators and functions take on the form

$$\mathbf{A} = A_0 + \sum_m \mathbf{A}_M, \quad \mathbf{f} = \mathbf{f}_0 P_0(\mathbf{y}), \quad \mathbf{g} = \mathbf{g}_0 P_0(\mathbf{y}) \text{ and } \mathbf{u} = \sum \mathbf{u}_\mu P_\mu(\mathbf{y}). \tag{10.7}$$

However, the transforms in (10.6) need to be applied to the *complete* discrete system, *i.e.* we should have

$$\hat{\mathbf{A}} = \mathbf{I}_I(\mathbf{A}_0 + \sum_m \mathbf{A}_m y_m)\mathbf{I}_I + \mathbf{D}\mathbf{I}_D \tag{10.8}$$

for the discrete operator and

$$\begin{aligned}\hat{\mathbf{f}} &= \mathbf{I}_I \mathbf{f}_0 P_0 + \mathbf{D}\mathbf{I}_D \mathbf{g}_0 P_0 - \mathbf{I}_I(\mathbf{A}_0 + \sum_m \mathbf{A}_m y_m)\mathbf{I}_D \mathbf{g}_0 P_0 \\ &= \mathbf{I}_I \mathbf{f}_0 P_0 + \mathbf{D}\mathbf{I}_D \mathbf{g}_0 P_0 - \mathbf{I}_I \mathbf{A}_0 \mathbf{I}_D \mathbf{g}_0 P_0 - \sum_m \mathbf{I}_I \mathbf{A}_m \mathbf{I}_D \mathbf{g}_0 y_m P_0 \\ &= \mathbf{I}_I \mathbf{f}_0 P_0 + \mathbf{D}\mathbf{I}_D \mathbf{g}_0 P_0 - \mathbf{I}_I \mathbf{A}_0 \mathbf{I}_D \mathbf{g}_0 P_0 - \sum_m \mathbf{I}_I \mathbf{A}_m \mathbf{I}_D \mathbf{g}_0 (\beta_0^m P_{\epsilon_m} + \alpha_0^m P_0)\end{aligned} \tag{10.9}$$

for the right-hand side with included boundary conditions. We denote the matrices assembled by the FEM code with included boundary conditions by a hat. It follows

$$\hat{\mathbf{A}} = \hat{\mathbf{A}}_0 + \sum_m (\hat{\mathbf{A}}_m - \mathbf{D}\mathbf{I}_D) y_m. \tag{10.10}$$

Equivalently,

$$\hat{\mathbf{A}} = \hat{\mathbf{A}}_0 + \sum_m \hat{\mathbf{A}}_m \mathbf{I}_I y_m, \tag{10.11}$$

since

$$\begin{aligned}\hat{\mathbf{A}}\mathbf{I}_I &= (\mathbf{I}_I \mathbf{A} \mathbf{I}_I + \mathbf{D}\mathbf{I}_D)\mathbf{I}_I \\ &= \mathbf{I}_I \mathbf{A} \mathbf{I}_I = \hat{\mathbf{A}} - \mathbf{D}\mathbf{I}_D\end{aligned} \tag{10.12}$$

This means that either the diagonal elements corresponding to boundary nodes can be set to zero in $\hat{\mathbf{A}}_m$ or, in order to apply $\hat{\mathbf{A}}_m$ to some vector $\mathbf{v}$, the boundary degrees of freedom in $\mathbf{v}$ are first set to zero and then $\hat{\mathbf{A}}_m$ is applied.

For the right-hand side, denote the action of the FEM code by φ, i.e.,

$$\hat{\mathbf{f}} = \varphi(\mathbf{f}, \mathbf{g}, \mathbf{A}) = \mathbf{I}_I \mathbf{f} + \mathbf{D}\mathbf{I}_D \mathbf{g} - \mathbf{I}_I \mathbf{A} \mathbf{I}_D \mathbf{g}. \tag{10.13}$$

The stochastic right-hand side can then be written as

$$\begin{aligned}\hat{\mathbf{f}} =& \varphi(\mathbf{f}_0, \mathbf{g}_0, \mathbf{A}_0) P_0(\mathbf{y}) + \sum_m \alpha_0^m \mathbf{I}_I \varphi(\mathbf{0}, \mathbf{g}_0, \mathbf{A}_m) P_0(\mathbf{y}) \\ &+ \sum_m \beta_0^m \mathbf{I}_I \varphi(\mathbf{0}, \mathbf{g}_0, \mathbf{A}_m) P_{\epsilon_m}(\mathbf{y})\end{aligned} \tag{10.14}$$

[2] This can be seen from combining the equations $\mathbf{I}_I \mathbf{A} \mathbf{u} = \mathbf{I}_I \mathbf{A}(\mathbf{I}_I \mathbf{u} + \mathbf{I}_D \mathbf{g})\mathbf{I}_I \mathbf{f}$ and $\mathbf{I}_D \mathbf{u} = \mathbf{I}_D \mathbf{g}$ into one system of equations.

Set $\hat{\mathbf{g}}_{0,m} := \mathbf{I}_I \varphi(\mathbf{0}, \mathbf{g}_0, \mathbf{A}_m)$. Collecting the terms for each multi-index $\mu \in \Lambda$ yields

$$\begin{aligned} \hat{\mathbf{f}}_0 &= \varphi(\mathbf{f}_0, \mathbf{g}_0, \mathbf{A}_0) + \sum_m \alpha_0^m \hat{\mathbf{g}}_{0,m}, \\ \hat{\mathbf{f}}_{\epsilon_m} &= \beta_0^m \hat{\mathbf{g}}_{0,m} \quad \text{and} \quad \hat{\mathbf{f}}_\mu = 0 \text{ for } |\mu| \geq 2. \end{aligned} \tag{10.15}$$

One concern in this presentation is that it does not explicitly treat the case that the solutions live on different meshes depending on the multi-index μ. This however, does not pose a real problem since the operators $\mathbf{I}_I$ and $\mathbf{I}_D$ can be considered to act on *the mesh at hand*. For some expression like $\mathbf{I}_I \mathbf{v}$ this means

$$\mathbf{I}_I \mathbf{v} = \mathbf{I}_I \sum_\mu \mathbf{v}_\mu P_\mu(\mathbf{y}) \sum_\mu = \mathbf{I}_{I_\mu} \mathbf{v}_\mu P_\mu(\mathbf{y}) \tag{10.16}$$

where I_μ indicates the interior nodes of mesh $\mathcal{T}_\mu$.

10.4. **Computation of the error estimator.** The evaluation of the residual error estimator consists of different components according to (7.1), (7.4) and (7.11). In the actual computation, we ensure that for each $\mu \in \Lambda$ the mesh $\mathcal{T}_\mu$ is sufficiently fine to resolve the oscillations of the coefficients in the operator. Moreover, we assume availability of an in general higher-order projection $\hat{\Pi}_{\mathcal{T}_{\mu_2}}^{\mu_1} : V(\mathcal{T}_{\mu_1}) \to V_k(\mathcal{T}_{\mu_2})$ where $k \geq 1$ indicates the polynomial degree. For the numerical examples we choose $\mathcal{T}_{\mu_2} \subseteq \mathcal{T}_{\mu_1}$ and $k = 2$. The implemented computation is depicted in Algorithm 4.

Algorithm 4: `Error`$[w_N^i, \zeta_N^i] \mapsto \xi_N^i, \hat{\eta}_N^i, \hat{\zeta}_N^i, \hat{\delta}_N^i$

initialize empty sets $\hat{\eta}_N^i,\ \hat{\zeta}_N^i,\ \hat{\delta}_N^i$
for $\mu \in \Lambda$ **do**
  `// residual part` (7.1)
  $\hat{\eta}_\mu(w_N^i) \coloneqq \{(\hat{\eta}_{\mu,S}, S)\,;\ S \in S_\mu\}$
  $\hat{\eta}_N^i := \hat{\eta}_N^i \cup \{(\mu, \hat{\eta}_\mu(w_N^i))\}$
  `// projection part` (7.4)
  for $m \in \mathbb{N}$ **do**
    refine $\mathcal{T}_\mu$ to obtain $\mathcal{T}_{\tilde{\mu}}$ with $\mathcal{T}_{\tilde{\mu}} \subseteq \mathcal{T}_{\mu \mp \epsilon_m}$
    for $T \in \mathcal{T}_{\tilde{\mu}}$ **do**
      $\zeta_{\mu,T,m}^{\mu \mp \epsilon_m}(w_N) \coloneqq$
      $\|\frac{a_m}{\bar{a}}\|_{L^\infty(D)} \beta_{\mu_m + \delta_{1 \mp 1}}^m |\hat{\Pi}_{\mathcal{T}_{\tilde{\mu}}}^{\mu} \Pi_\mu^{\mu \mp \epsilon_m} w_{N,\mu \mp \epsilon_m} - \hat{\Pi}_{\mathcal{T}_{\tilde{\mu}}}^{\mu \mp \epsilon_m} w_{N,\mu \mp \epsilon_m}|_{V,\Sigma_m}$
    $\hat{\zeta}_\mu(w_N^i) \coloneqq \left\{ \left(\sum_{\substack{\hat{T} \subseteq T \\ \hat{T} \in \mathcal{T}_{\tilde{\mu}}}} \hat{\zeta}_{\mu,\hat{T}}(w_N^i), T\right)\,;\ T \in \mathcal{T}_\mu \right\}$
  $\hat{\zeta}_N^i := \hat{\zeta}_N^i \cup \{(\mu, \hat{\zeta}_\mu(w_N^i))\}$
  `// new multi-indices` $\mu \in \mathcal{F} \setminus \Lambda$ (7.11)
  $\delta_\mu \coloneqq \{(\delta_{\mu \mp \epsilon_m}(w_N^i), \mu \mp \epsilon_m)\,;\ m \in \mathbb{N}\}$
$\hat{\delta} \coloneqq \left\{ \left(\sum_{(\tilde{\delta},\tilde{\mu}) \in \{(\delta^*,\mu^*) \in \delta_\mu\,;\ \mu^* = \tilde{\mu},\ \mu \in \Lambda\}} \tilde{\delta}, \tilde{\mu}\right)\,;\ \tilde{\mu} \in \mathcal{F} \setminus \Lambda \right\}$
`// total error` (8.5)
$\xi_N^i \leftarrow \text{evaluate}(\hat{\eta}_N^i, \hat{\zeta}_N^i, \hat{\delta}_N^i)$

11. Numerical Experiments

This section is devoted to several benchmark problems which illustrate the performance of the residual error estimator. With the implementation described in Section 10, numerical experiments for the Poisson model problem (2.1) and for the

Navier–Lamé equations (9.8) of linearized elasticity in a plane, polygonal domain $D \subset \mathbb{R}^2$ are examined. Recall from Section 2.1 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 (2.2).

The expansion coefficients of the stochastic field (2.2) are chosen to be

$$a_m(x) \coloneqq \alpha_m \cos(2\pi\beta_1(m)x_1)\cos(2\pi\beta_2(m)x_2) \tag{11.1}$$

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

$$\beta_1(m) = m - k(m)(k(m)+1)/2 \quad \text{and} \quad \beta_2(m) = k(m) - \beta_1(m) \tag{11.2}$$

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 σ in (11.1) to either 2 or 4. An overview of the activated multi-indices with the dimensions of the respective discrete spaces is depicted in Table 1.

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, identically distributed 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

$$\begin{aligned} \|e\|_V^2 &= \int_\Omega \|e(\cdot,\omega)\|_V^2 dP(\omega) \\ &= \int_\Omega \|A^{-1}(y(\omega))f - u_N(y(\omega))\|_V^2 dP(\omega) \\ &\approx \frac{1}{M}\sum_{i=1}^M \|A^{-1}(y^{(i)})f - u_N(y^{(i)})\|_V^2 \,. \end{aligned} \tag{11.3}$$

Here, the samples $y^{(i)} \in \Gamma$ of parameter sequences are assumed to be statistically independent, and identically distributed with law P in Γ. Note that the sampled solutions $A^{-1}(y^{(i)})f$ are only computed approximately since the operator is discretized on a mesh which is a uniform refinement of the joint mesh generated from the SGFEM discretization of the final iteration. Moreover, the expansion (2.2) of the random field $a(y,x)$ is truncated to the same length as for the approximate parametric solution.

The ensuing numerical experiments are based on the following standard parameter choices for the adaptive algorithm of Section 8.2,

$$\bar{c}_Q = 1, \quad \bar{c}_\eta = 1, \quad \vartheta_\eta = 2/5, \quad \vartheta_\zeta = 10 \quad \text{and} \quad \vartheta_\delta = 1\,. \tag{11.4}$$

The initial mesh $\mathring{\mathcal{T}}$ for activated $\mu \in \mathcal{F} \setminus \Lambda$ is sufficiently fine to approximate the solution with respect to the oscillating coefficient a_m with good accuracy.

11.1. **Poisson model problem.**

11.1.1. *Square domain.* The first example is the Poisson model problem (2.1) 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 8.2 for a slow decay of the coefficients with $\sigma = 2$ and a fast decay with $\sigma = 4$ are shown in Figure 1. The amplitude A in (11.1) was chosen as $\gamma/\zeta(\sigma)$ with $\gamma = 0.9$, resulting in $A \approx 0.547$ for $\sigma = 2$ and $A \approx 0.832$ for $\sigma = 4$. Depicted is the residual estimator, the reference error obtained by Monte Carlo sampling, the efficiency of the estimator

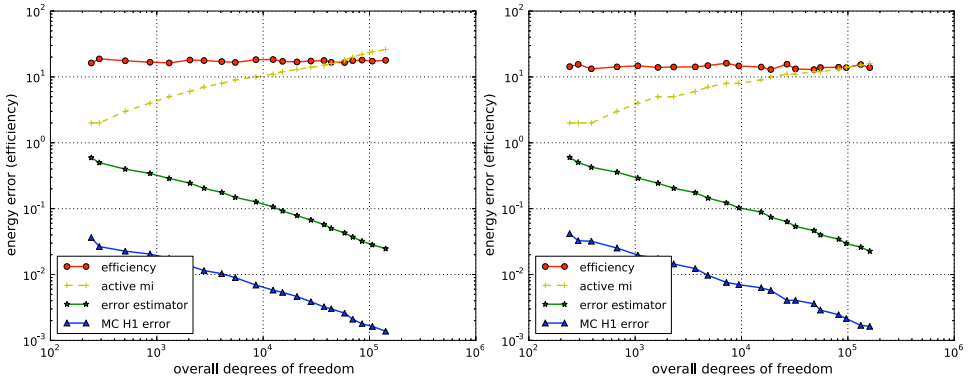

FIGURE 1. Convergence of the error estimator and the MC error for the Poisson model problem with homogeneous Dirichlet boundary conditions in the energy norm for slow ($\sigma = 2$, left) and fast ($\sigma = 4$, right) decay. Number of activated multi-indices and efficiency of the error estimator with respect to the MC reference error.

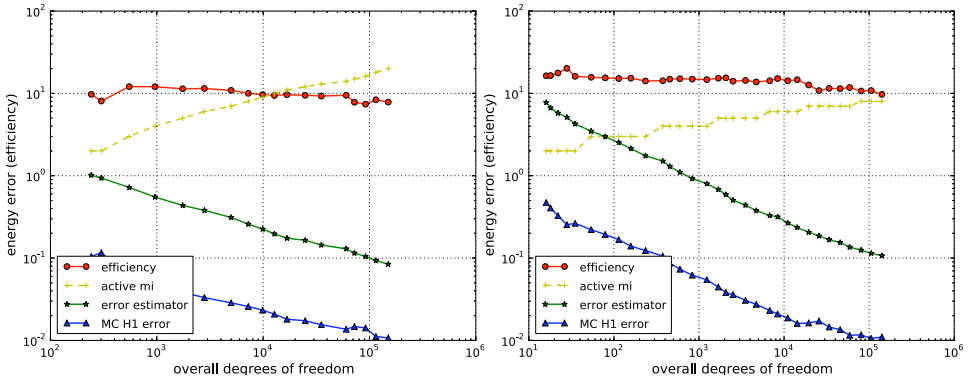

FIGURE 2. Convergence of the estimated and the MC error for the Poisson model problem with Neumann boundaries on the square (left) and with homogeneous Dirichlet boundary condition on the L-shaped domain (right). Number of activated multi-indices and efficiency of the error estimator.

and the number of active multi-indices. The observed convergence rate of 1/2 with respect to the total number of degrees of freedom, which is the convergence rate for a single deterministic problem, coincides with the approximation rates predicted by [9, 17].

In addition to the homogeneous Dirichlet problem, we also consider the Poisson model problem with homogeneous Neumann boundary conditions on the three sides $x_2 = 0,\ x_2 = 1,\ x_1 = 1$ of the unit square and a homogeneous Dirichlet boundary condition on the side $x_1 = 0$ as before. The convergence graphs for slow coefficient decay in the coefficient expansion (2.2) are presented in Figure 2 (left). We observe that the estimator is slightly more accurate in this setting than with complete Dirichlet boundary conditions.

The number of active stochastic modes in the set Λ after a fixed number of iterations is significantly larger in the case of a slower decay rate of the coefficient amplitude due to the influence of higher modes on the solution. This can also be

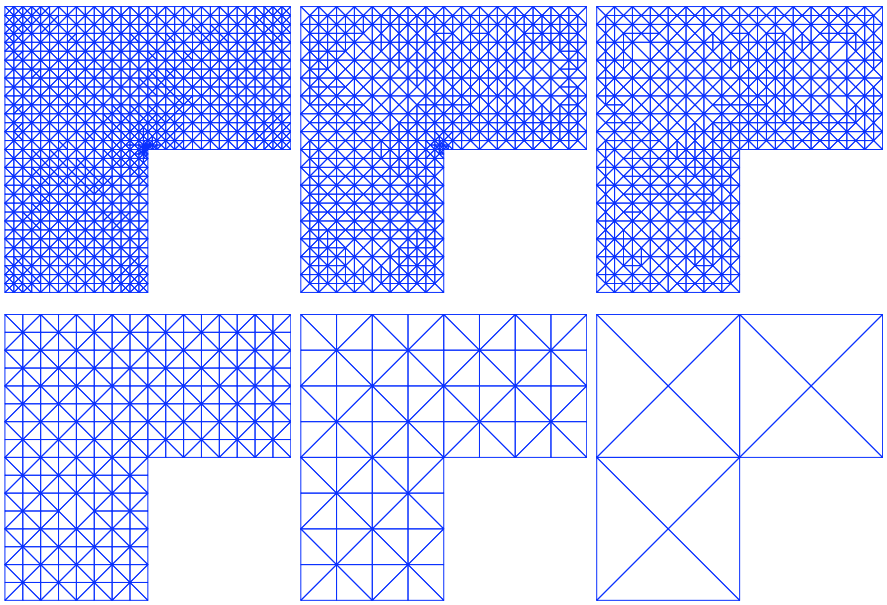

FIGURE 3. Adapted meshes for the Poisson model problem on the L-shaped domain of Section 11.1.2. Deterministic mesh (top left) and meshes for some active stochastic modes after 15 iterations (from top left to bottom right): (1), (0 0 0 1), (2), (0 0 0 0 0 1), (0 0 0 0 0 0 0 1).

seen in the data of Table 1. In particular, modes of higher (polynomial) degree are favoured in case of fast decay. As is common for residual error estimators in deterministic FEM, the overestimation of the error amounts to approximately 10. This could be considered quite large, but is, in part, due to the fact that the residual error bounds are uniform with respect to the stochastic parameters. It is to be expected that with other a posteriori error bounds better reliability constants could be achieved.

11.1.2. *L-shaped domain.* A standard benchmark problem for deterministic a posteriori error estimators is the Poisson problem (2.1) on the L-shaped domain $D = (-1,1)^2 \setminus (0,1) \times (-1,0)$. The solution exhibits a well-known singularity at the reentrant corner at $(0,0)$ which is resolved by a pronounced mesh refinement in its vicinity. In this example, the adaptive algorithm is thus assumed to also show this behavior for the deterministic part and the most significant stochastic modes which is illustrated with the meshes in Figures 3. The convergence results are depicted in Figure 2 (right). Since the singularity is the main contribution to the error estimator as long as it is not resolved adequately, the adaptive algorithm first focuses on the residual before also refining with regard to the stochastic dimensions. Thus, the number of active multi-indices is smaller than in the problem of the previous section. More details can be found in Table 1.

11.1.3. *Cook's membrane example.* This common benchmark problem for bending dominated elastic response defines the tapered panel D which is clamped at the side $x_1 = 0$ and subjected to a shearing load $g = (0,1)^\top$on the opposite side $x_1 = 48$ with vanishing volume force $f = (0,0)^\top$. The geometry is defined by $D = \mathrm{conv}\{(0,0),(48,44),(48,60),(0,44)\}$, see Figure 5 (top left). We assume the

	square		L-shape		Cook	
multi-index	$\sigma = 2$	$\sigma = 4$	$\sigma = 2$	$\sigma = 4$	$\sigma = 2$	$\sigma = 4$
(0)	93599	102397	116262	104456	25262	27222
(1)	15539	24667	10982	17454	56848	98886
(2)	2701	10919	405	4105	∗	2020
(3)	121	3073	∗	∗	∗	∗
(4)	∗	1039	∗	∗	∗	∗
(0 1)	7009	5963	5028	3462	25622	19522
(0 2)	531	121	∗	∗	∗	∗
(1 1)	2397	4255	21	819	∗	∗
(2 1)	391	2057	∗	∗	∗	∗
(3 1)	∗	819	∗	∗	∗	∗
(0 0 1)	2561	2025	4759	21	17990	∗
(1 0 1)	1415	1277	∗	∗	∗	∗
(1 1 1)	749	∗	∗	∗	∗	∗
(0 0 0 1)	1929	867	1667	∗	3246	∗
(1 0 0 1)	841	∗	∗	∗	∗	∗
(0 0 0 0 1)	1579	351	2164	∗	486	∗
(1 0 0 0 1)	677	∗	∗	∗	∗	∗
(0 0 0 0 0 1)	873	∗	∗	∗	∗	∗
(1 0 0 0 0 1)	221	∗	∗	∗	∗	∗
(0 0 0 0 0 0 1)	783	∗	∗	∗	∗	∗
(0 0 0 0 0 0 0 1)	749	∗	∗	∗	∗	∗
(0 0 0 0 0 0 0 0 1)	505	∗	∗	∗	∗	∗
(0 0 0 0 0 0 0 0 0 1)	437	∗	∗	∗	∗	∗
(0 0 0 0 0 0 0 0 0 0 1)	333	∗	∗	∗	∗	∗
other indices (dofs)	2 (342)	∗	∗	∗	∗	∗
cardΛ	26	16	8	6	6	4
overall dofs	**141824**	**160584**	**141288**	**130317**	**129454**	**147650**

TABLE 1. Activated multi-indices and dimensions of discrete spaces for benchmark problems. The number of iterations for the two different decay rates is fixed per experiment.

nominal Young modulus $E = 2900$ and nominal Poisson ratio $\nu = 0.4$ which corresponds to plexiglass. These values determine the mean field of the Lamé parameters $\bar{\mu} \approx 4142.9$ and $\bar{\lambda} \approx 1035.7$. As noted in Section 9.3, the parameter λ is modeled as spatially heterogeneous random field according to (2.2). For the coefficient functions a_m the same model as in (11.1) has been chosen with $A = \bar{\lambda}\gamma/\zeta(\sigma)$ resulting in $A \approx 2266.7$ for $\sigma = 2$ and $A \approx 3445.0$ for $\sigma = 4$, given that $\gamma = 0.9$.

The convergence graphs for the error estimator and for the MC reference error in the energy norm for slow and fast decay are depicted in Figure 4. The efficiency of the estimator again is in the expected range with an overestimation by a factor of approximately 10 in both examples. Again, the activation rate is significantly lower for faster decay of the coefficient weights. More details about the activated

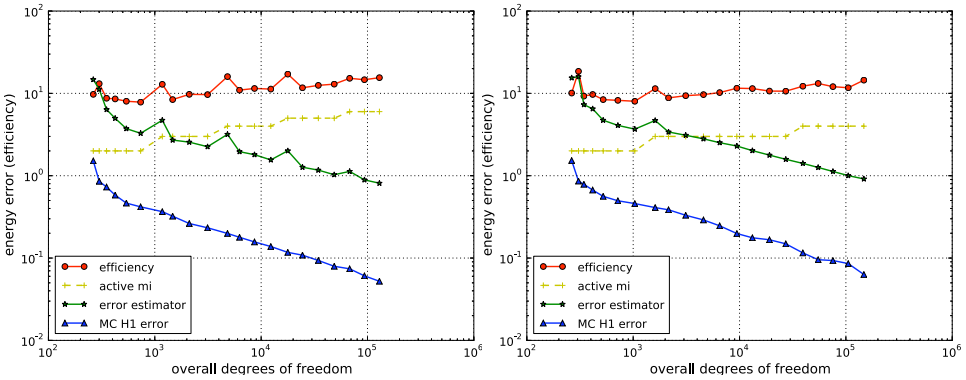

FIGURE 4. Convergence of the error estimator and the MC error in the energy norm for Cook's membrane of Section 11.1.3. Number of activated multi-indices and efficiency of the error estimator for slow ($\sigma = 2$, left) and fast ($\sigma = 4$, right) decay.

multi-indices and the dimension of the corresponding discrete spaces can be found in Table 1. Some adaptively refined meshes for the deterministic part and some stochastic modes are pictured in Figure 5.

REFERENCES

[1] M. AINSWORTH AND J. T. ODEN, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000.

[2] I. BABUSKA, F. NOBILE, AND R. TEMPONE, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM Rev., 52 (2010), pp. 317–355.

[3] S. BARTELS, C. CARSTENSEN, AND G. DOLZMANN, *Inhomogeneous Dirichlet conditions in a priori and a posteriori finite element error analysis*, Numer. Math., 99 (2004), pp. 1–24.

[4] M. BIERI, *A sparse composite collocation finite element method for elliptic SPDEs*, SIAM J. Numer. Anal., 49 (2011), pp. 2277–2301.

[5] M. BIERI, R. ANDREEV, AND C. SCHWAB, *Sparse tensor discretization of elliptic SPDEs*, SIAM J. Sci. Comput., 31 (2009/10), pp. 4281–4304.

[6] S. C. BRENNER AND L. R. SCOTT, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer-Verlag, New York, second ed., 2002.

[7] C. CARSTENSEN, M. EIGEL, R. H. HOPPE, AND C. LÖBHARD, *A review of unified a posteriori finite element error control*, Numer. Math. Theor. Meth. Appl., 5 (2012), pp. 509–558.

[8] 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), pp. 2524–2550.

[9] A. COHEN, R. DEVORE, AND C. SCHWAB, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's*, Anal. Appl. (Singap.), 9 (2011), pp. 11–47.

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

[11] M. EIGEL AND E. ZANDER, *SpUQ - A Python Framework for Spectral Methods for Uncertainty Quantification*.

[12] O. G. ERNST, A. MUGLER, H.-J. STARKLOFF, AND E. ULLMANN, *On the convergence of generalized polynomial chaos expansions*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 317–339.

[13] W. GAUTSCHI, *Orthogonal polynomials: computation and approximation*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2004. Oxford Science Publications.

[14] C. J. GITTELSON, *An adaptive stochastic Galerkin method for random elliptic operators*. To appear in Math. Comp.

[15] ———, *Adaptive Galerkin Methods for Parametric and Stochastic Operator Equations*, PhD thesis, ETH Zürich, 2011. ETH Dissertation No. 19533.

[16] ———, *Stochastic Galerkin approximation of operator equations with infinite dimensional noise*, Tech. Rep. 2011-10, Seminar for Applied Mathematics, ETH Zürich, 2011.
[17] ———, *Convergence rates of multilevel and sparse tensor approximations for a random elliptic PDE*. submitted, 2012.
[18] W. Hackbusch, *Iterative Lösung großer schwachbesetzter Gleichungssysteme*, vol. 69 of Leitfäden der Angewandten Mathematik und Mechanik [Guides to Applied Mathematics and Mechanics], B. G. Teubner, Stuttgart, 1991. Teubner Studienbücher Mathematik. [Teubner Mathematical Textbooks].
[19] A. Logg, K.-A. Mardal, G. N. Wells, et al., *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012.
[20] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488 (electronic).
[21] V. Nistor and C. Schwab, *High order Galerkin approximations for parametric second order elliptic partial differential equations*, Tech. Rep. 2012-21, Seminar for Applied Mathematics, ETH Zürich, 2012.
[22] C. Schwab and C. J. Gittelson, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, Acta Numer., 20 (2011), pp. 291–467.
[23] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Teubner Verlag and J. Wiley, Stuttgart, 1996.

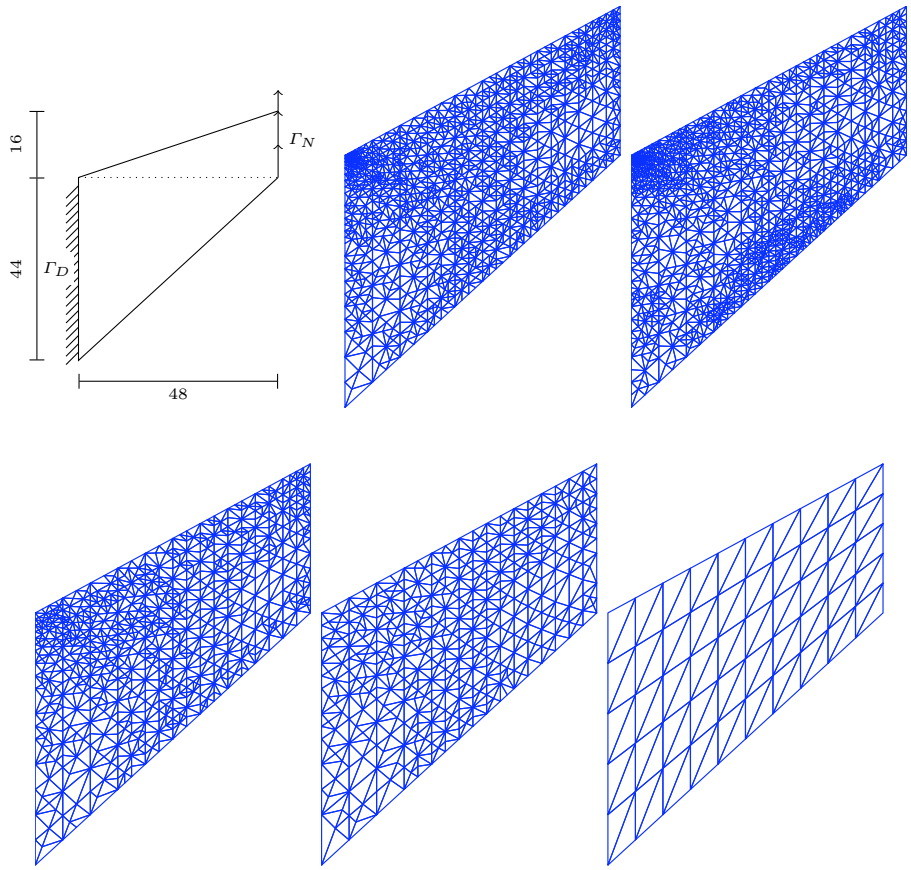

FIGURE 5. Adaptively refined meshes (deterministic and stochastic) for Cook's membrane example of Section 11.1.3. Depicted meshes are for active multi-indices (top left to bottom right): (0), (1), (0 0 0 1), (0 0 0 0 0 1), (0 0 0 0 0 0 0 0 1).

Humboldt Universität, Rudower Chaussee 25, D-12489 Berlin, Germany
E-mail address: eigel@mathematik.hu-berlin.de

Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA
E-mail address: cgittels@purdue.edu

Seminar for Applied Mathematics, ETH Zürich, Rämistrasse 101, CH-8092 Zürich, Switzerland
E-mail address: schwab@sam.math.ethz.ch

Institute of Scientific Computing, Technical University Braunschweig, D-38092 Braunschweig, Germany
E-mail address: e.zander@tu-bs.de

Research Reports

No.	Authors/Title
13-01	*M. Eigel, C.J. Gittelson, Ch. Schwab and E. Zander* Adaptive stochastic Galerkin FEM
12-42	*J. Waldvogel* Jost Bürgi and the discovery of the logarithms
12-41	*M. Hansen* n-term approximation rates and Besov regularity for elliptic PDEs on polyhedral domains
12-40	*D. Schötzau, Ch. Schwab, T. Wihler and M. Wirz* Exponential convergence of *hp*-DGFEM for elliptic problems in polyhedral domains
12-39	*A. Buffa, G. Sangalli and Ch. Schwab* Exponential convergence of the *hp* version of isogeometric analysis in 1D
12-38	*R. Hiptmair, A. Moiola, I. Perugia and Ch. Schwab* Approximation by harmonic polynomials in star-shaped domains and exponential convergence of Trefftz *hp*-DGFEM
12-37	*Cl. Schillings and Ch. Schwab* Sparse, adaptive Smolyak algorithms for Bayesian inverse problems
12-36	*R. Hiptmair and L. Kielhorn* BETL A generic boundary element template library
12-35	*S. Mishra, N.H. Risebro, Ch. Schwab and S. Tokareva* Numerical solution of scalar conservation laws with random flux functions
12-34	*R. Hiptmair, Ch. Schwab and C. Jerez-Hanckes* Sparse tensor edge elements
12-33	*R. Hiptmair, C. Jerez-Hanckes and S. Mao* Extension by zero in discrete trace spaces: Inverse estimates
12-32	*A. Lang, S. Larsson and Ch. Schwab* Covariance structure of parabolic stochastic partial differential equations
12-31	*A. Madrane, U.S. Fjordholm, S. Mishra and E. Tadmor* Entropy conservative and entropy stable finite volume schemes for multidimensional conservation laws on unstructured meshes
12-30	*G.M. Coclite, L. Di Ruvo, J. Ernest and S. Mishra* Convergence of vanishing capillarity approximations for scalar conservation laws with discontinuous fluxes