

A posteriori error estimation and adaptivity in stochastic Galerkin FEM for parametric elliptic PDEs: beyond the affine case*

Alex Bespalov[†] Feng Xu[†]

Abstract

We consider a linear elliptic partial differential equation (PDE) with a generic uniformly bounded parametric coefficient. The solution to this PDE problem is approximated in the framework of stochastic Galerkin finite element methods. We perform a posteriori error analysis of Galerkin approximations and derive a reliable and efficient estimate for the energy error in these approximations. Practical versions of this error estimate are discussed and tested numerically for a model problem with non-affine parametric representation of the coefficient. Furthermore, we use the error reduction indicators derived from spatial and parametric error estimators to guide an adaptive solution algorithm for the given parametric PDE problem. The performance of the adaptive algorithm is tested numerically for model problems with two different non-affine parametric representations of the coefficient.

Key words. stochastic Galerkin methods, stochastic finite element methods, parametric PDEs, a posteriori error estimation, adaptive methods, sparse polynomial approximation, generalized polynomial chaos expansion

AMS subject classifications. 35R60, 65C20, 65N30, 65N15

1 Introduction

Partial differential equations (PDEs) with uncertain or parameter-dependent inputs arise in mathematical models of many physical phenomena as well as in engineering applications. Stochastic Galerkin finite element method (sGFEM) is commonly used for solving such PDE problems numerically, in particular, when the input data and solutions are

*This work was supported by the EPSRC under grant EP/P013791/1 and by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

[†]School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT (a.bespalov@bham.ac.uk, f.xu.2@bham.ac.uk).

sufficiently smooth functions of parameters. The sGFEM solution is sought in the tensor product of a finite element space defined on the physical domain and a multivariable polynomial space on the parameter domain. Even if a moderate number of parameters is used to represent the problem inputs, the cost associated with computing high-fidelity sGFEM approximations quickly becomes prohibitive, due to fast growth of the dimension of the tensor product space. An adaptive approach to constructing approximation spaces provides a remedy to this computational bottleneck. Based on rigorous a posteriori error analysis of computed solutions, adaptive solution techniques build spatial and parametric components of approximations incrementally in the course of numerical computation, leading to accelerated convergence and reduced computational cost.

For elliptic PDE problems with *affine-parametric* coefficients, several adaptive sGFEM algorithms have been recently proposed and analyzed, see, e.g., [16, 8, 7, 5, 10, 4, 3]. A range of the underlying a posteriori error estimation techniques is used in these and other works in order to guide adaptive refinement (e.g., residual-based, local equilibration, and hierarchical a posteriori error estimators and error indicators to name but a few). By contrast, the sGFEM-based numerical schemes for problems with *non-affine* parametric representations of coefficients are significantly less well developed. As far as adaptive stochastic Galerkin approximations are concerned, the only work we are aware of is [9], where the adaptive sGFEM procedure driven by reliable *residual-based* error indicators is developed for linear elliptic PDEs with lognormal coefficients. It is worth noting that, due to unboundedness of coefficients, a well-posed weak formulation of this problem needs to be introduced in problem-dependent weighted spaces, as presented in [19]. Practical feasibility of the adaptive algorithm in [9] is ensured by adaptive discretizations of the lognormal coefficient represented in a hierarchical tensor format, as described in [11], under the assumption that the errors in such discretizations are small.

In this paper, we consider a linear elliptic PDE with a *generic* parametric coefficient. In particular, our analysis is not restricted to any specific form of the parametric coefficient (affine, quadratic, log-uniform, etc.). Assuming uniform boundedness of the coefficient, which is a minimal requirement to ensure well-posedness of the weak formulation in standard Lebesgue–Bochner spaces, we derive a reliable and efficient a posteriori estimate of the energy error in sGFEM approximations. This extends the analysis of hierarchical error estimators presented in [2, 5] and fills a gap in the existing theory. Two practical examples of hierarchical error estimates are considered in detail and studied numerically for the steady-state diffusion problem with non-affine parametric representation of the coefficient. We then present an adaptive algorithm driven by the error reduction indicators derived from hierarchical a posteriori error estimators in the spirit of [5, 4]. The performance of the adaptive algorithm is tested numerically for two non-affine parametric representations of the diffusion coefficient.

The rest of the paper is structured as follows. The model problem is introduced in

section 2; its Galerkin approximation and a posteriori error estimation are presented in section 3. The generalized polynomial chaos (gPC) expansion of the parametric coefficient and the associated practical aspects of the developed error estimation strategy are discussed in section 4, while the results of numerical tests are reported in section 5. The adaptive algorithm is proposed in section 6, and its performance is tested in numerical experiments described in section 7. In Appendix A, we derive explicit formulae for calculating the gPC expansion coefficients for parametric exponential and quadratic functions.

2 Stochastic steady-state diffusion problem

Let $D \subset \mathbb{R}^2$ be a bounded (spatial) domain with a Lipschitz polygonal boundary ∂D , and let $\Gamma := \prod_{m=1}^{\infty} \Gamma_m$ be the parameter domain with bounded intervals $\Gamma_m \subset \mathbb{R}$. Let $H_0^1(D)$ be the usual Sobolev space of functions in $H^1(D)$ vanishing at the boundary ∂D in the sense of traces. We will use the standard norm in $H_0^1(D)$ as $\|v\|_{H_0^1(D)} := \|\nabla v\|_{L^2(D)}$. As an example of model problem, we consider the homogeneous Dirichlet problem for the parametric steady-state diffusion equation

$$\begin{aligned} -\nabla \cdot (T(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) &= f(\mathbf{x}), & \mathbf{x} \in D, \mathbf{y} = (y_1, y_2, \dots) \in \Gamma, \\ u(\mathbf{x}, \mathbf{y}) &= 0, & \mathbf{x} \in \partial D, \mathbf{y} \in \Gamma, \end{aligned} \quad (2.1)$$

where $f \in H^{-1}(D)$ and ∇ denotes differentiation with respect to \mathbf{x} only. We assume that the parameters y_m , $m \in \mathbb{N}$, are the images of *independent* random variables with cumulative distribution function $\pi_m(y_m)$ and probability density function $q_m(y_m) = d\pi_m(y_m)/dy_m$. Then for the multivariate random variable formed by all independent univariate random variables, the joint cumulative distribution function and the joint probability density function are $\pi(\mathbf{y}) := \prod_{m=1}^{\infty} \pi_m(y_m)$ and $q(\mathbf{y}) := \prod_{m=1}^{\infty} q_m(y_m)$, respectively. Since each Γ_m is bounded, we can always rescale the corresponding univariate random variable such that it takes values in $[-1, 1]$. Therefore, without loss of generality, we assume that $\Gamma_m := [-1, 1]$ for all $m \in \mathbb{N}$.

Note that each π_m is a probability measure on $(\Gamma_m, \mathcal{B}(\Gamma_m))$, where $\mathcal{B}(\Gamma_m)$ is the Borel σ -algebra on Γ_m . Accordingly, π is a probability measure on $(\Gamma, \mathcal{B}(\Gamma))$, where $\mathcal{B}(\Gamma)$ is the Borel σ -algebra on Γ . Then $L_{\pi_m}^2(\Gamma_m)$ (resp., $L_{\pi}^2(\Gamma)$) represents the Lebesgue space of equivalence classes of functions $v : \Gamma_m \rightarrow \mathbb{R}$ (resp., $v : \Gamma \rightarrow \mathbb{R}$) that are square integrable on Γ_m (resp., Γ) with respect to the measure π_m (resp., π), and $\langle \cdot, \cdot \rangle_{\pi_m}$ (resp., $\langle \cdot, \cdot \rangle_{\pi}$) denotes the associated inner product: $\langle f, g \rangle_{\pi_m} := \int_{\Gamma_m} q_m(y_m) f(y_m) g(y_m) dy_m$ for $f, g \in L_{\pi_m}^2(\Gamma_m)$ (resp., $\langle f, g \rangle_{\pi} := \int_{\Gamma} q(\mathbf{y}) f(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}$ for $f, g \in L_{\pi}^2(\Gamma)$). For a Hilbert space H of functions on D , we will denote by $L_{\pi}^2(\Gamma; H)$ the space of strongly measurable equivalence classes of functions $v : D \times \Gamma \rightarrow \mathbb{R}$ such that

$$\|v\|_{L_{\pi}^2(\Gamma; H)} := \left(\int_{\Gamma} q(\mathbf{y}) \|v(\cdot, \mathbf{y})\|_H^2 d\mathbf{y} \right)^{1/2} < +\infty.$$

In particular, we will denote $V := L^2_\pi(\Gamma; H^1_0(D))$ and $W := L^2_\pi(\Gamma; L^2(D))$.

The weak formulation of (2.1) reads as follows: find $u \in V$ such that

$$B(u, v) = F(v) \quad \forall v \in V, \quad (2.2)$$

where the symmetric bilinear form $B(\cdot, \cdot)$ and the linear functional $F(\cdot)$ are defined by

$$B(u, v) := \int_\Gamma q(\mathbf{y}) \int_D T(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y}, \quad (2.3)$$

$$F(v) := \int_\Gamma q(\mathbf{y}) \int_D f(\mathbf{x}) v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y}. \quad (2.4)$$

To ensure the well-posedness of (2.2), we make the following assumption on the parametric diffusion coefficient $T \in L^\infty_\pi(\Gamma; L^\infty(D))$: there exist constants α_{\min} and α_{\max} such that

$$0 < \alpha_{\min} \leq T(\mathbf{x}, \mathbf{y}) \leq \alpha_{\max} < \infty \quad \text{a.e. in } D \times \Gamma. \quad (2.5)$$

In particular, this implies that $B(\cdot, \cdot)$ is continuous and elliptic on V . Therefore, $B(\cdot, \cdot)$ defines an inner product in V which induces the norm $\|v\|_B := B(v, v)^{1/2}$ that is equivalent to $\|v\|_V$, i.e.,

$$\alpha_{\min} \|v\|_V^2 \leq \|v\|_B^2 \leq \alpha_{\max} \|v\|_V^2 \quad \forall v \in V. \quad (2.6)$$

3 Galerkin approximation and a posteriori error estimation

3.1 Galerkin approximation

Let us introduce the finite-dimensional approximation of the weak problem (2.2). Problem (2.2) can be discretized by using Galerkin projection onto any finite-dimensional subspace of V . Note that the space $V = L^2_\pi(\Gamma; H^1_0(D))$ is isometrically isomorphic to the tensor product Hilbert space $H^1_0(D) \otimes L^2_\pi(\Gamma)$ (see, e.g., [19, Theorem B.17, Remark C.24]). Hence we can construct the finite-dimensional subspace of V by tensorizing a finite-dimensional subspace of $H^1_0(D)$ and a finite-dimensional subspace of $L^2_\pi(\Gamma)$.

For the finite-dimensional subspace of $H^1_0(D)$, we choose the finite element space $X = \text{span}\{\phi_1, \dots, \phi_{n_X}\}$, where ϕ_i are standard finite element basis functions and $n_X = \dim(X)$.

Let us now introduce the finite-dimensional (polynomial) subspaces of $L^2_\pi(\Gamma)$. To that end, we consider the following set of finitely supported sequences:

$$\mathcal{I} := \{\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots) \in \mathbb{N}_0^{\mathbb{N}}; \max(\text{supp } \boldsymbol{\alpha}) < \infty\},$$

where $\text{supp } \boldsymbol{\alpha} = \{m \in \mathbb{N}; \alpha_m \neq 0\}$. The set \mathcal{I} , as well as any of its subsets, will be called the *index set*, and the elements $\boldsymbol{\alpha} \in \mathcal{I}$ will be called the *(multi-)indices*. For each $m \in \mathbb{N}$,

let $\{p_n^m\}_{n \in \mathbb{N}_0}$ denote the set of univariate polynomials on Γ_m that are orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_{\pi_m}$ in $L_{\pi_m}^2(\Gamma_m)$. Then we can define the following tensor product polynomials:

$$p_{\boldsymbol{\alpha}}(\mathbf{y}) := \prod_{m=1}^{\infty} p_{\alpha_m}^m(y_m) = \prod_{m \in \text{supp } \boldsymbol{\alpha}} p_{\alpha_m}^m(y_m) \quad \forall \boldsymbol{\alpha} \in \mathcal{I}.$$

The countable set $\{p_{\boldsymbol{\alpha}}; \boldsymbol{\alpha} \in \mathcal{I}\}$ forms an orthonormal basis of $L_{\pi}^2(\Gamma)$ (see, e.g., [13, section 3.3]). Given a finite index set $\mathcal{P} \subset \mathcal{I}$, the space of tensor product polynomials $P_{\mathcal{P}} := \text{span}\{p_{\boldsymbol{\alpha}}; \boldsymbol{\alpha} \in \mathcal{P}\}$ defines a finite-dimensional subspace of $L_{\pi}^2(\Gamma)$.

With both spaces $X \subset H_0^1(D)$ and $P_{\mathcal{P}} \subset L_{\pi}^2(\Gamma)$, we can now define the finite-dimensional subspace $V_{X\mathcal{P}} := X \otimes P_{\mathcal{P}} \subset V$ and write the discrete formulation of (2.2) as follows: find $u_{X\mathcal{P}} \in V_{X\mathcal{P}}$ such that

$$B(u_{X\mathcal{P}}, v) = F(v) \quad \forall v \in V_{X\mathcal{P}}. \quad (3.1)$$

Hereafter, we assume that \mathcal{P} always contains the zero-index $\mathbf{0} := (0, 0, \dots)$.

3.2 A posteriori error estimation

The aim of this subsection is to generalize the results of [5] to the case of the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ satisfying only the boundedness assumption (2.5) (which is a minimal assumption that guarantees the well-posedness of the weak formulation (2.2)).

We follow the classical hierarchical a posteriori error estimation strategy as described, e.g., in [1, Chapter 5]. First, let us briefly outline the main ingredients of this strategy emphasizing the specific features pertaining to tensor-product approximations. The starting point is the following equation for the discretization error $e := u - u_{X\mathcal{P}} \in V$:

$$B(e, v) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in V. \quad (3.2)$$

Since e lives in the infinite-dimensional space V , we cannot calculate e by using (3.2) directly. However, one can approximate the error e in a finite-dimensional subspace $V_{X\mathcal{P}}^* \subset V$ in a similar way as the solution u is approximated in the finite-dimensional subspace $V_{X\mathcal{P}} \subset V$. Specifically, we introduce the error estimator $e^* \in V_{X\mathcal{P}}^*$ that satisfies

$$B(e^*, v) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in V_{X\mathcal{P}}^*. \quad (3.3)$$

Note that, due to Galerkin orthogonality

$$B(e, v) = F(v) - B(u_{X\mathcal{P}}, v) = 0 \quad \forall v \in V_{X\mathcal{P}}, \quad (3.4)$$

a meaningful approximation of e is obtained by requiring that $V_{X\mathcal{P}} \subsetneq V_{X\mathcal{P}}^*$.

It is well known that the error estimator e^* is linked to the enhanced Galerkin approximation $u_{X\mathcal{P}}^* \in V_{X\mathcal{P}}^*$ as follows: $e^* = u_{X\mathcal{P}}^* - u_{X\mathcal{P}}$. Here, $u_{X\mathcal{P}}^* \in V_{X\mathcal{P}}^*$ satisfies

$$B(u_{X\mathcal{P}}^*, v) = F(v) \quad \forall v \in V_{X\mathcal{P}}^*. \quad (3.5)$$

Furthermore, since $B(\cdot, \cdot)$ is symmetric, we deduce from (2.2), (3.1), (3.5) that

$$\|e\|_B^2 = \|u - u_{X\mathcal{P}}\|_B^2 = F(u) - F(u_{X\mathcal{P}}), \quad \|u - u_{X\mathcal{P}}^*\|_B^2 = F(u) - F(u_{X\mathcal{P}}^*) \quad (3.6)$$

and

$$\|e^*\|_B^2 = \|u_{X\mathcal{P}}^* - u_{X\mathcal{P}}\|_B^2 = F(u_{X\mathcal{P}}^*) - F(u_{X\mathcal{P}}) \stackrel{(3.6)}{=} \|e\|_B^2 - \|u - u_{X\mathcal{P}}^*\|_B^2.$$

This implies that: (i) $\|e^*\|_B \leq \|e\|_B$; (ii) the quantity $\|e^*\|_B$ is the energy error reduction achieved by using the enriched space $V_{X\mathcal{P}}^*$; and (iii) $\|u - u_{X\mathcal{P}}^*\|_B \leq \|u - u_{X\mathcal{P}}\|_B$. In order to establish the equivalence between the true energy error $\|e\|_B$ and the energy error estimate $\|e^*\|_B$, the following stronger property than the one given in (iii) is assumed (this property is usually referred to as the saturation assumption): there exists a constant $\beta \in [0, 1)$ such that

$$\|u - u_{X\mathcal{P}}^*\|_B \leq \beta \|u - u_{X\mathcal{P}}\|_B. \quad (3.7)$$

Then the following inequalities hold (see, e.g., [1, Theorem 5.1]):

$$\|e^*\|_B \leq \|e\|_B \leq \frac{1}{\sqrt{1 - \beta^2}} \|e^*\|_B. \quad (3.8)$$

Motivated by high computational cost involved in computing the error estimator e^* defined by (3.3) (the cost that is comparable to computing the enhanced Galerkin approximation $u_{X\mathcal{P}}^*$), hierarchical a posteriori error estimation techniques seek to approximate e^* by making use of the following two key ingredients: (a) an alternative bilinear form $\tilde{B}(\cdot, \cdot)$ in place of $B(\cdot, \cdot)$ on the left-hand side in (3.3) with the aim to obtain an easier to invert (stiffness) matrix in the associated linear system; (b) an appropriate decomposition of the enhanced finite-dimensional space $V_{X\mathcal{P}}^*$ with the aim to further reduce computational cost by solving (3.3) on the subspace(s) of $V_{X\mathcal{P}}^*$.

The alternative bilinear form $\tilde{B}(\cdot, \cdot)$ is employed to define the modified error estimator $\tilde{e} \in V_{X\mathcal{P}}^*$ satisfying

$$\tilde{B}(\tilde{e}, v) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in V_{X\mathcal{P}}^*. \quad (3.9)$$

For problem (3.9) to be well-posed, the auxiliary bilinear form $\tilde{B}(\cdot, \cdot)$ is assumed to be symmetric, continuous, and elliptic. In this case, \tilde{B} defines an inner product in V which induces the norm $\|v\|_{\tilde{B}} := \tilde{B}(v, v)^{1/2}$ that is equivalent to $\|v\|_B$, i.e., there exist two positive constants λ and Λ such that

$$\lambda \|v\|_B \leq \|v\|_{\tilde{B}} \leq \Lambda \|v\|_B \quad \forall v \in V. \quad (3.10)$$

This leads to the following relation between the error estimators e^* and \tilde{e} (see, e.g., [1, Theorem 5.3]):

$$\lambda \|\tilde{e}\|_{\tilde{B}} \leq \|e^*\|_B \leq \Lambda \|\tilde{e}\|_{\tilde{B}}. \quad (3.11)$$

The discussion of the second ingredient of the hierarchical error estimation strategy is linked to the specific choice of the enriched subspace $V_{X\mathcal{P}}^* \subset V$. In the context of tensor-product approximations, an appropriate choice of $V_{X\mathcal{P}}^*$ is important, as this affects the quality of the final error estimate as well as the computational cost associated with computing that estimate, cf. [2, 5]. In this paper, we follow the idea proposed in [5]. Firstly, we construct an enriched finite element subspace $X^* \subset H_0^1(D)$, which has a direct sum decomposition $X^* := X \oplus Y$, where the finite-dimensional subspace $Y \subset H_0^1(D)$ is called the *detail finite element space*. Secondly, we construct an enriched polynomial space $P_{\mathcal{P}^*} := \text{span}\{p_\alpha; \alpha \in \mathcal{P}^*\}$ associated with a finite index set $\mathcal{P}^* := \mathcal{P} \cup \mathcal{Q}$ for some $\mathcal{Q} \subset \mathcal{I}$ such that $\mathcal{P} \cap \mathcal{Q} = \emptyset$. The set \mathcal{Q} is called the *detail index set* and the corresponding polynomial space $P_{\mathcal{Q}} := \text{span}\{p_\alpha; \alpha \in \mathcal{Q}\}$ is called the *detail polynomial space*. Note that $P_{\mathcal{P}^*}$ has an orthogonal direct sum decomposition with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$ as follows:

$$P_{\mathcal{P}^*} = P_{\mathcal{P}} \oplus P_{\mathcal{Q}}.$$

Finally, the enriched finite-dimensional space $V_{X\mathcal{P}}^*$ is defined as the following direct sum:

$$V_{X\mathcal{P}}^* := V_{X\mathcal{P}} \oplus V_{Y\mathcal{P}} \oplus V_{X\mathcal{Q}},$$

where $V_{Y\mathcal{P}} := Y \otimes P_{\mathcal{P}}$ and $V_{X\mathcal{Q}} := X \otimes P_{\mathcal{Q}}$.

The direct sum structure of $V_{X\mathcal{P}}^*$ motivates the definition of two error estimators $e_{Y\mathcal{P}} \in V_{Y\mathcal{P}}$ and $e_{X\mathcal{Q}} \in V_{X\mathcal{Q}}$ satisfying

$$\tilde{B}(e_{Y\mathcal{P}}, v) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in V_{Y\mathcal{P}}, \quad (3.12)$$

$$\tilde{B}(e_{X\mathcal{Q}}, v) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in V_{X\mathcal{Q}}. \quad (3.13)$$

Combining all ingredients, we define the following error estimate

$$\eta := \sqrt{\|e_{Y\mathcal{P}}\|_{\tilde{B}}^2 + \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2}. \quad (3.14)$$

Clearly, making the right choice of the auxiliary bilinear form $\tilde{B}(\cdot, \cdot)$ is important in the above construction. In particular, if this choice implies \tilde{B} -orthogonality of the subspace decomposition, the following abstract result holds.

Lemma 3.1. *Let $\tilde{B}(\cdot, \cdot)$ be a symmetric bilinear form that is continuous and elliptic on a Hilbert space V , and let $G(\cdot)$ be a continuous linear functional on V . Consider three subspaces $V_1, V_2, V_3 \subset V$ such that $V_3 = V_1 \oplus V_2$. Let $e_i \in V_i$ ($i = 1, 2, 3$) satisfy*

$$\tilde{B}(e_i, v) = G(v) \quad \forall v \in V_i. \quad (3.15)$$

If the direct sum decomposition $V_3 = V_1 \oplus V_2$ is \tilde{B} -orthogonal, i.e.,

$$\tilde{B}(u, v) = 0 \quad \forall u \in V_1, \quad \forall v \in V_2, \quad (3.16)$$

then

$$e_3 = e_1 + e_2 \quad \text{and} \quad \tilde{B}(e_3, e_3) = \tilde{B}(e_1, e_1) + \tilde{B}(e_2, e_2). \quad (3.17)$$

Proof. Since $V_3 = V_1 \oplus V_2$, every $v \in V_3$ has a unique decomposition $v = v_1 + v_2$ with $v_1 \in V_1$, $v_2 \in V_2$. Using the orthogonality relation (3.16), we deduce from (3.15) (with $i = 1, 2$) that

$$\begin{aligned}\tilde{B}(e_1 + e_2, v) &= \tilde{B}(e_1 + e_2, v_1 + v_2) = \tilde{B}(e_1, v_1) + \tilde{B}(e_1, v_2) + \tilde{B}(e_2, v_1) + \tilde{B}(e_2, v_2) \\ &= \tilde{B}(e_1, v_1) + \tilde{B}(e_2, v_2) = G(v_1) + G(v_2) = G(v) \quad \forall v \in V_3.\end{aligned}$$

This implies that $e_3 = e_1 + e_2$, because e_3 is the unique solution of (3.15) with $i = 3$. The second equality in (3.17) then follows due to the orthogonality property (3.16) and the symmetry of the bilinear form $\tilde{B}(\cdot, \cdot)$. \square

A direct application of Lemma 3.1 to the subspace $V_{X\mathcal{Q}} = \bigoplus_{\mu \in \mathcal{Q}} X \otimes P_{\{\mu\}}$ gives the following result on the decomposition of the error estimator $e_{X\mathcal{Q}}$ defined by (3.13).

Corollary 3.1. *Assume that the direct sum decomposition $V_{X\mathcal{Q}} = \bigoplus_{\mu \in \mathcal{Q}} X \otimes P_{\{\mu\}}$ is \tilde{B} -orthogonal, i.e., for any $\mu, \nu \in \mathcal{Q}$ ($\mu \neq \nu$) there holds*

$$\tilde{B}(u, v) = 0 \quad \forall u \in X \otimes P_{\{\mu\}}, \quad \forall v \in X \otimes P_{\{\nu\}}. \quad (3.18)$$

Then the error estimator $e_{X\mathcal{Q}}$ defined by (3.13) and its norm $\|e_{X\mathcal{Q}}\|_{\tilde{B}}$ can be decomposed into the contributions associated with individual indices $\mu \in \mathcal{Q}$ as follows:

$$e_{X\mathcal{Q}} = \sum_{\mu \in \mathcal{Q}} e_{X\mathcal{Q}}^{(\mu)}, \quad \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2 = \sum_{\mu \in \mathcal{Q}} \left\| e_{X\mathcal{Q}}^{(\mu)} \right\|_{\tilde{B}}^2. \quad (3.19)$$

Here, for each index $\mu \in \mathcal{Q}$, the estimator $e_{X\mathcal{Q}}^{(\mu)} \in X \otimes P_{\{\mu\}}$ satisfies

$$\tilde{B}\left(e_{X\mathcal{Q}}^{(\mu)}, v\right) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in X \otimes P_{\{\mu\}}. \quad (3.20)$$

The next step is to connect the error estimates $\|\tilde{e}\|_{\tilde{B}}$ and η . To this end, we employ two strengthened Cauchy–Schwarz inequalities (see, e.g., [1, 12]): there exist two constants $\kappa_1, \kappa_2 \in [0, 1)$ such that

$$|\tilde{B}(u, v)| \leq \kappa_1 \|u\|_{\tilde{B}} \|v\|_{\tilde{B}} \quad \forall u \in V_{X^*\mathcal{P}} := V_{X\mathcal{P}} \oplus V_{Y\mathcal{P}}, \quad \forall v \in V_{X\mathcal{Q}}, \quad (3.21)$$

$$|\tilde{B}(u, v)| \leq \kappa_2 \|u\|_{\tilde{B}} \|v\|_{\tilde{B}} \quad \forall u \in V_{X\mathcal{P}}, \quad \forall v \in V_{Y\mathcal{P}}. \quad (3.22)$$

Lemma 3.2. *Let $\|\tilde{e}\|_{\tilde{B}}$ and η be defined in (3.9) and (3.14), respectively. Then the following inequalities hold*

$$\frac{1}{\sqrt{2}}\eta \leq \|\tilde{e}\|_{\tilde{B}} \leq \frac{1}{\sqrt{(1-\kappa_1)(1-\kappa_2^2)}}\eta, \quad (3.23)$$

Furthermore, if $\kappa_1 = 0$ in (3.21) (that is, $V_{X^*\mathcal{P}}$ and $V_{X\mathcal{Q}}$ are \tilde{B} -orthogonal), then

$$\eta \leq \|\tilde{e}\|_{\tilde{B}} \leq \frac{1}{\sqrt{1-\kappa_2^2}}\eta. \quad (3.24)$$

Proof. We start by defining an auxiliary error estimator $e_{X^*\mathcal{P}} \in V_{X^*\mathcal{P}}$ satisfying

$$\tilde{B}(e_{X^*\mathcal{P}}, v) = F(v) - B(u_{X\mathcal{P}}, v) \quad \forall v \in V_{X^*\mathcal{P}}. \quad (3.25)$$

The proof then consists of four steps.

Step 1. In this step, we will establish the following inequalities:

$$\frac{\|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2}{2} \leq \|\tilde{e}\|_{\tilde{B}}^2 \leq \frac{\|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2}{1 - \kappa_1}. \quad (3.26)$$

Since $V_{X^*\mathcal{P}}$ and $V_{X\mathcal{Q}}$ are subspaces of $V_{X^*\mathcal{P}}^*$, we use (3.9), (3.25), (3.13) and apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned} \|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 &= \tilde{B}(e_{X^*\mathcal{P}}, e_{X^*\mathcal{P}}) = \tilde{B}(\tilde{e}, e_{X^*\mathcal{P}}) \leq \|\tilde{e}\|_{\tilde{B}} \|e_{X^*\mathcal{P}}\|_{\tilde{B}}, \\ \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2 &= \tilde{B}(e_{X\mathcal{Q}}, e_{X\mathcal{Q}}) = \tilde{B}(\tilde{e}, e_{X\mathcal{Q}}) \leq \|\tilde{e}\|_{\tilde{B}} \|e_{X\mathcal{Q}}\|_{\tilde{B}}. \end{aligned}$$

Hence, the left-hand inequality in (3.26) follows.

Let us now prove the right-hand inequality in (3.26). Since $V_{X^*\mathcal{P}}^* = V_{X^*\mathcal{P}} \oplus V_{X\mathcal{Q}}$, the estimator $\tilde{e} \in V_{X^*\mathcal{P}}^*$ has a unique decomposition

$$\tilde{e} = w_{X^*\mathcal{P}} + w_{X\mathcal{Q}} \quad \text{with } w_{X^*\mathcal{P}} \in V_{X^*\mathcal{P}}, \quad w_{X\mathcal{Q}} \in V_{X\mathcal{Q}}.$$

Using this representation of \tilde{e} , we deduce that

$$\begin{aligned} \|\tilde{e}\|_{\tilde{B}}^2 &= \tilde{B}(\tilde{e}, \tilde{e}) = \tilde{B}(\tilde{e}, w_{X^*\mathcal{P}} + w_{X\mathcal{Q}}) = \tilde{B}(\tilde{e}, w_{X^*\mathcal{P}}) + \tilde{B}(\tilde{e}, w_{X\mathcal{Q}}) \\ &= \tilde{B}(e_{X^*\mathcal{P}}, w_{X^*\mathcal{P}}) + \tilde{B}(e_{X\mathcal{Q}}, w_{X\mathcal{Q}}) \leq \|e_{X^*\mathcal{P}}\|_{\tilde{B}} \|w_{X^*\mathcal{P}}\|_{\tilde{B}} + \|e_{X\mathcal{Q}}\|_{\tilde{B}} \|w_{X\mathcal{Q}}\|_{\tilde{B}} \\ &\leq \left(\|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2 \right)^{1/2} \left(\|w_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|w_{X\mathcal{Q}}\|_{\tilde{B}}^2 \right)^{1/2}, \end{aligned} \quad (3.27)$$

where the fourth equality is due to (3.9), (3.25) and (3.13), the first inequality is due to the Cauchy–Schwarz inequality, and the second inequality is due to the algebraic inequality $ab + cd \leq (a^2 + c^2)^{1/2}(b^2 + d^2)^{1/2}$.

On the other hand, we can estimate $\|\tilde{e}\|_{\tilde{B}}^2$ from below as follows:

$$\begin{aligned} \|\tilde{e}\|_{\tilde{B}}^2 &= \tilde{B}(\tilde{e}, \tilde{e}) = \tilde{B}(w_{X^*\mathcal{P}} + w_{X\mathcal{Q}}, w_{X^*\mathcal{P}} + w_{X\mathcal{Q}}) \\ &= \tilde{B}(w_{X^*\mathcal{P}}, w_{X^*\mathcal{P}}) + 2\tilde{B}(w_{X^*\mathcal{P}}, w_{X\mathcal{Q}}) + \tilde{B}(w_{X\mathcal{Q}}, w_{X\mathcal{Q}}) \\ &\stackrel{(3.21)}{\geq} \|w_{X^*\mathcal{P}}\|_{\tilde{B}}^2 - 2\kappa_1 \|w_{X^*\mathcal{P}}\|_{\tilde{B}} \|w_{X\mathcal{Q}}\|_{\tilde{B}} + \|w_{X\mathcal{Q}}\|_{\tilde{B}}^2 \\ &\geq \|w_{X^*\mathcal{P}}\|_{\tilde{B}}^2 - \kappa_1 \left(\|w_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|w_{X\mathcal{Q}}\|_{\tilde{B}}^2 \right) + \|w_{X\mathcal{Q}}\|_{\tilde{B}}^2 \\ &= (1 - \kappa_1) \left(\|w_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|w_{X\mathcal{Q}}\|_{\tilde{B}}^2 \right). \end{aligned} \quad (3.28)$$

Combining (3.27) with (3.28) gives the right-hand inequality in (3.26).

Step 2. In the second step, we will establish the following inequalities:

$$\|e_{Y\mathcal{P}}\|_{\tilde{B}}^2 \leq \|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 \leq \frac{\|e_{Y\mathcal{P}}\|_{\tilde{B}}^2}{1 - \kappa_2^2}. \quad (3.29)$$

Since $V_{Y\mathcal{P}} \subset V_{X^*\mathcal{P}}$, we use (3.25), (3.12) and the Cauchy–Schwarz inequality to obtain

$$\|e_{Y\mathcal{P}}\|_{\tilde{B}}^2 = \tilde{B}(e_{Y\mathcal{P}}, e_{Y\mathcal{P}}) = \tilde{B}(e_{X^*\mathcal{P}}, e_{Y\mathcal{P}}) \leq \|e_{X^*\mathcal{P}}\|_{\tilde{B}} \|e_{Y\mathcal{P}}\|_{\tilde{B}}.$$

Hence, the left-hand inequality in (3.29) follows.

Using similar arguments as in Step 1, the proof for the right-hand inequality in (3.29) first makes use of the decomposition

$$V_{X^*\mathcal{P}} \ni e_{X^*\mathcal{P}} = w_{X\mathcal{P}} + w_{Y\mathcal{P}} \quad \text{with } w_{X\mathcal{P}} \in V_{X\mathcal{P}}, \quad w_{Y\mathcal{P}} \in V_{Y\mathcal{P}}$$

and the Cauchy–Schwarz inequality to estimate

$$\begin{aligned} \|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 &= \tilde{B}(e_{X^*\mathcal{P}}, e_{X^*\mathcal{P}}) = \tilde{B}(e_{X^*\mathcal{P}}, w_{X\mathcal{P}} + w_{Y\mathcal{P}}) \\ &= \tilde{B}(e_{X^*\mathcal{P}}, w_{Y\mathcal{P}}) = \tilde{B}(e_{Y\mathcal{P}}, w_{Y\mathcal{P}}) \leq \|e_{Y\mathcal{P}}\|_{\tilde{B}} \|w_{Y\mathcal{P}}\|_{\tilde{B}}; \end{aligned} \quad (3.30)$$

here, the third equality is due to $\tilde{B}(e_{X^*\mathcal{P}}, w_{X\mathcal{P}}) = 0$ as follows from (3.1) and (3.25), and the fourth equality is due to (3.25) and (3.12). On the other hand, applying the strengthened Cauchy–Schwarz inequality (3.22) and the algebraic inequality $2\kappa_2 ab \leq a^2 + \kappa_2^2 b^2$, we obtain the lower bound for $\|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2$:

$$\begin{aligned} \|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 &= \tilde{B}(e_{X^*\mathcal{P}}, e_{X^*\mathcal{P}}) = \tilde{B}(w_{X\mathcal{P}} + w_{Y\mathcal{P}}, w_{X\mathcal{P}} + w_{Y\mathcal{P}}) \\ &\geq \|w_{X\mathcal{P}}\|_{\tilde{B}}^2 - 2\kappa_2 \|w_{X\mathcal{P}}\|_{\tilde{B}} \|w_{Y\mathcal{P}}\|_{\tilde{B}} + \|w_{Y\mathcal{P}}\|_{\tilde{B}}^2 \\ &\geq \|w_{X\mathcal{P}}\|_{\tilde{B}}^2 - \|w_{X\mathcal{P}}\|_{\tilde{B}}^2 - \kappa_2^2 \|w_{Y\mathcal{P}}\|_{\tilde{B}}^2 + \|w_{Y\mathcal{P}}\|_{\tilde{B}}^2 = (1 - \kappa_2^2) \|w_{Y\mathcal{P}}\|_{\tilde{B}}^2. \end{aligned} \quad (3.31)$$

Combining (3.30) with (3.31) gives the right-hand inequality in (3.29).

Step 3. Combining (3.26) with (3.29) and recalling the definition of η gives (3.23).

Step 4 ($\kappa_1 = 0$). A tighter lower bound in (3.23) can be proved in this case. Indeed, using the \tilde{B} -orthogonality of the decomposition $V_{X^*\mathcal{P}} = V_{X^*\mathcal{P}} \oplus V_{X\mathcal{Q}}$ and applying Lemma 3.1 we conclude that $\|\tilde{e}\|_{\tilde{B}}^2 = \|e_{X^*\mathcal{P}}\|_{\tilde{B}}^2 + \|e_{X\mathcal{Q}}\|_{\tilde{B}}^2$. Combining this equality with the estimates (3.29) from Step 2 and recalling the definition of η we obtain (3.24). \square

Putting together (3.8), (3.11), (3.23) and (3.24), the following theorem gives two-sided bounds for the energy norm (i.e., B -norm) of the true discretization error $e = u - u_{X\mathcal{P}}$ in terms of the estimate η .

Theorem 3.1. *Let $u \in V$ be the solution of (2.2) and let $u_{X\mathcal{P}} \in V_{X\mathcal{P}}$ be the Galerkin approximation satisfying (3.1). Suppose that the saturation assumption (3.7) and the norm equivalence (3.10) hold. Then the a posteriori error estimate η defined by (3.14) satisfies*

$$\frac{\lambda}{\sqrt{2}} \eta \leq \|u - u_{X\mathcal{P}}\|_B \leq \frac{\Lambda}{\sqrt{1 - \beta^2} \sqrt{(1 - \kappa_1)(1 - \kappa_2^2)}} \eta, \quad (3.32)$$

where $\beta \in [0, 1)$ is the constant in (3.7), λ and Λ are the constants in (3.10), and $\kappa_1, \kappa_2 \in [0, 1)$ are the constants in the strengthened Cauchy–Schwarz inequalities (3.21), (3.22).

Furthermore, if $\kappa_1 = 0$ in (3.21) (that is, $V_{X^*\mathcal{P}}$ and $V_{X\mathcal{Q}}$ are \tilde{B} -orthogonal), then

$$\lambda \eta \leq \|u - u_{X\mathcal{P}}\|_B \leq \frac{\Lambda}{\sqrt{1 - \beta^2} \sqrt{1 - \kappa_2^2}} \eta. \quad (3.33)$$

Remark 3.1. While error estimates (3.32) are new, the estimates in (3.33) have been proved in [5, Theorem 4.1] for the model problem (2.1) with the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ that has affine dependence on random parameters. In that framework, the auxiliary bilinear form $\tilde{B}(\cdot, \cdot)$ is associated with the parameter-free part of the representation for $T(\mathbf{x}, \mathbf{y})$ and yields the orthogonality of the decomposition $V_{X\mathcal{P}}^* = V_{X^*\mathcal{P}} \oplus V_{X\mathcal{Q}}$. Thus, Theorem 3.1 generalizes the results of [5] to the case of a more general diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ that is only assumed to be bounded (the assumption that ensures the well-posedness of (2.2)). In fact, our result is not limited to the diffusion problem (2.1). Theorem 3.1 applies to tensor-product Galerkin approximations of the solution to a general variational problem of the type (2.2) with symmetric bilinear form B that is continuous and elliptic on a Bochner-type space V .

Recalling that $e^* = u_{X\mathcal{P}}^* - u_{X\mathcal{P}}$ and putting together (3.11), (3.23) and (3.24), the following theorem gives two-sided bounds for the error reduction $\|u_{X\mathcal{P}}^* - u_{X\mathcal{P}}\|_B$ in terms of the estimate η .

Theorem 3.2. Let $u_{X\mathcal{P}} \in V_{X\mathcal{P}}$ be the Galerkin approximation satisfying (3.1), and let $u_{X\mathcal{P}}^* \in V_{X\mathcal{P}}^*$ be the enhanced Galerkin approximation satisfying (3.5). Suppose that the norm equivalence (3.10) holds. Then the following estimates for the error reduction hold:

$$\frac{\lambda}{\sqrt{2}}\eta \leq \|u_{X\mathcal{P}}^* - u_{X\mathcal{P}}\|_B \leq \frac{\Lambda}{\sqrt{(1-\kappa_1)(1-\kappa_2^2)}}\eta, \quad (3.34)$$

where λ and Λ are the constants in (3.10) and $\kappa_1, \kappa_2 \in [0, 1)$ are the constants in the strengthened Cauchy–Schwarz inequalities (3.21), (3.22).

Furthermore, if $\kappa_1 = 0$ in (3.21) (that is, $V_{X^*\mathcal{P}}$ and $V_{X\mathcal{Q}}$ are \tilde{B} -orthogonal), then

$$\lambda\eta \leq \|u_{X\mathcal{P}}^* - u_{X\mathcal{P}}\|_B \leq \frac{\Lambda}{\sqrt{1-\kappa_2^2}}\eta. \quad (3.35)$$

Remark 3.2. Theorem 3.2 states that η provides an estimate for the error reduction $\|u_{X\mathcal{P}}^* - u_{X\mathcal{P}}\|_B$. We distinguish the following two important cases of enriching the approximation space $V_{X\mathcal{P}}$:

- (1) If only the finite element space is enriched, that is, $V_{X\mathcal{P}}^* = V_{X^*\mathcal{P}} = V_{X\mathcal{P}} \oplus V_{Y\mathcal{P}}$, and $u_{X^*\mathcal{P}} \in V_{X^*\mathcal{P}}$ denotes the enhanced Galerkin solution, then $\kappa_1 = 0$ and therefore $\eta = \|e_{Y\mathcal{P}}\|_{\tilde{B}}$ provides an effective estimate for the error reduction $\|u_{X^*\mathcal{P}} - u_{X\mathcal{P}}\|_B$, i.e.,

$$\lambda\|e_{Y\mathcal{P}}\|_{\tilde{B}} \leq \|u_{X^*\mathcal{P}} - u_{X\mathcal{P}}\|_B \leq \frac{\Lambda}{\sqrt{1-\kappa_2^2}}\|e_{Y\mathcal{P}}\|_{\tilde{B}}. \quad (3.36)$$

- (2) If only the polynomial space on Γ is enriched, that is, $V_{X\mathcal{P}}^* = V_{X\mathcal{P}^*} := V_{X\mathcal{P}} \oplus V_{X\mathcal{Q}}$, and $u_{X\mathcal{P}^*} \in V_{X\mathcal{P}^*}$ denotes the corresponding enhanced Galerkin solution, then $\kappa_2 = 0$

and therefore $\eta = \|e_{X\mathcal{Q}}\|_{\tilde{B}}$ provides an effective estimate for the error reduction $\|u_{X\mathcal{P}^*} - u_{X\mathcal{P}}\|_B$, i.e.,

$$\frac{\lambda}{\sqrt{2}} \|e_{X\mathcal{Q}}\|_{\tilde{B}} \leq \|u_{X\mathcal{P}^*} - u_{X\mathcal{P}}\|_B \leq \frac{\Lambda}{\sqrt{1 - \kappa_1}} \|e_{X\mathcal{Q}}\|_{\tilde{B}},$$

when $\kappa_1 \neq 0$, and

$$\lambda \|e_{X\mathcal{Q}}\|_{\tilde{B}} \leq \|u_{X\mathcal{P}^*} - u_{X\mathcal{P}}\|_B \leq \Lambda \|e_{X\mathcal{Q}}\|_{\tilde{B}}, \quad (3.37)$$

when $\kappa_1 = 0$.

Similar to Remark 3.1, we emphasize that Theorem 3.2 generalizes the results of [2, 5], where the error reduction estimates (3.36), (3.37) have been proved for the model problem (2.1) with the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ that has affine dependence on random parameters.

4 Galerkin approximations for the model problem with coefficient in the gPC expansion form

While the results of section 3 hold for a general variational problem of type (2.2) (see, e.g., Remark 3.1), we now focus on the steady-state diffusion problem (2.1). For this problem, we use the generalized polynomial chaos expansion of the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ and specify main ingredients of computing stochastic Galerkin approximations and the associated error estimators. Here, and in the rest of the paper, we assume that $T(\mathbf{x}, \mathbf{y})$ depends on *finite* number of parameters y_m ($m = 1, \dots, M$, $M \in \mathbb{N}$). As before, we suppose that $T(\mathbf{x}, \mathbf{y})$ satisfies the boundedness assumption (2.5). Then $T(\mathbf{x}, \mathbf{y}) \in W$ can be represented using the gPC expansion as follows (see, e.g., [22] or [13, Theorem 3.6]):

$$T(\mathbf{x}, \mathbf{y}) = \sum_{\gamma \in \mathbb{N}_0^M} t_\gamma(\mathbf{x}) p_\gamma(\mathbf{y}), \quad (4.1)$$

where the orthonormality of the polynomial basis $\{p_\gamma\}_{\gamma \in \mathbb{N}_0^M}$ gives

$$t_\gamma(\mathbf{x}) = \langle T, p_\gamma \rangle_\pi = \int_\Gamma T(\mathbf{x}, \mathbf{y}) p_\gamma(\mathbf{y}) q(\mathbf{y}) \, d\mathbf{y} \quad \forall \gamma \in \mathbb{N}_0^M. \quad (4.2)$$

4.1 Discrete formulation revisited

Recalling that $X = \text{span}\{\phi_1, \phi_2, \dots, \phi_{n_X}\}$ and $P_{\mathcal{P}} = \text{span}\{p_\alpha; \alpha \in \mathcal{P} \subset \mathbb{N}_0^M\}$, we can write any $u \in V_{X\mathcal{P}} = X \otimes P_{\mathcal{P}}$ as

$$u(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n_X} \sum_{\alpha \in \mathcal{P}} u_{i,\alpha} \phi_i(\mathbf{x}) p_\alpha(\mathbf{y}), \quad u_{i,\alpha} \in \mathbb{R}. \quad (4.3)$$

We note that given multi-indices $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{N}_0^M$, the orthogonality of the polynomial basis (with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$) yields the following property:

$$\int_{\Gamma} q(\mathbf{y}) p_{\boldsymbol{\alpha}}(\mathbf{y}) p_{\boldsymbol{\beta}}(\mathbf{y}) p_{\boldsymbol{\gamma}}(\mathbf{y}) d\mathbf{y} = \prod_{m=1}^M \int_{\Gamma_m} q_m(y_m) p_{\alpha_m}^m(y_m) p_{\beta_m}^m(y_m) p_{\gamma_m}^m(y_m) dy_m = 0 \quad (4.4)$$

if there exists $m \in \{1, 2, \dots, M\}$ such that the sum of any two of α_m, β_m and γ_m is less than the third one. Therefore, with two *finite* index sets $\mathcal{P}, \mathcal{Q} \subset \mathbb{N}_0^M$, we obtain by using (4.1) in the definition (2.3) of the bilinear form $B(\cdot, \cdot)$

$$\begin{aligned} B(u, v) &= \sum_{\boldsymbol{\gamma} \in \mathbb{N}_0^M} \int_{\Gamma} q p_{\boldsymbol{\gamma}} \int_D t_{\boldsymbol{\gamma}} \nabla u \cdot \nabla v d\mathbf{x} d\mathbf{y} \\ &= \sum_{\boldsymbol{\gamma} \in \mathcal{N}(\mathcal{P}, \mathcal{Q})} \int_{\Gamma} q p_{\boldsymbol{\gamma}} \int_D t_{\boldsymbol{\gamma}} \nabla u \cdot \nabla v d\mathbf{x} d\mathbf{y} \quad \forall u \in V_{X\mathcal{P}}, \quad \forall v \in V_{X\mathcal{Q}}, \end{aligned} \quad (4.5)$$

where

$$\begin{aligned} \mathcal{N}(\mathcal{P}, \mathcal{Q}) &:= \{\boldsymbol{\gamma} \in \mathbb{N}_0^M; \exists \boldsymbol{\alpha} \in \mathcal{P}, \exists \boldsymbol{\beta} \in \mathcal{Q}, \text{ such that} \\ &\quad |\alpha_m - \beta_m| \leq \gamma_m \leq \alpha_m + \beta_m, \quad \forall m = 1, \dots, M\}. \end{aligned} \quad (4.6)$$

Thus, by using the Galerkin projection (3.1) onto the finite-dimensional subspace $V_{X\mathcal{P}}$, the infinite sum in the expansion (4.1) of $T(\mathbf{x}, \mathbf{y})$ is effectively truncated to the finite sum over the indices $\boldsymbol{\gamma} \in \mathcal{N}(\mathcal{P}, \mathcal{P})$ ¹. In particular, using the representation (4.3) for the Galerkin approximation $u_{X\mathcal{P}} \in V_{X\mathcal{P}}$ and setting $v = \phi_j p_{\boldsymbol{\beta}}$ in (3.1), we obtain for all $j = 1, \dots, n_X$ and $\boldsymbol{\beta} \in \mathcal{P}$

$$\sum_{\boldsymbol{\gamma} \in \mathcal{N}(\mathcal{P}, \mathcal{P})} \sum_{i=1}^{n_X} \sum_{\boldsymbol{\alpha} \in \mathcal{P}} u_{i, \boldsymbol{\alpha}} \int_D t_{\boldsymbol{\gamma}} \nabla \phi_i \cdot \nabla \phi_j d\mathbf{x} \int_{\Gamma} q p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}} p_{\boldsymbol{\gamma}} d\mathbf{y} = \int_D f \phi_j d\mathbf{x} \int_{\Gamma} q p_{\boldsymbol{\beta}} d\mathbf{y}. \quad (4.7)$$

Hence, the discrete formulation (3.1) results in the linear system $A\mathbf{u} = \mathbf{b}$ with the matrix A and the right-hand side vector \mathbf{b} being defined as follows:

$$\begin{aligned} A &:= \sum_{\boldsymbol{\gamma} \in \mathcal{N}(\mathcal{P}, \mathcal{P})} G_{\boldsymbol{\gamma}} \otimes K_{\boldsymbol{\gamma}}, \quad \mathbf{b} := \mathbf{g} \otimes \mathbf{f}, \\ [G_{\boldsymbol{\gamma}}]_{\iota(\boldsymbol{\alpha})\iota(\boldsymbol{\beta})} &:= \int_{\Gamma} q p_{\boldsymbol{\alpha}} p_{\boldsymbol{\beta}} p_{\boldsymbol{\gamma}} d\mathbf{y}, \quad \iota(\boldsymbol{\alpha}), \iota(\boldsymbol{\beta}) = 1, \dots, \#\mathcal{P}, \\ [K_{\boldsymbol{\gamma}}]_{ij} &:= \int_D t_{\boldsymbol{\gamma}} \nabla \phi_i \cdot \nabla \phi_j d\mathbf{x}, \quad i, j = 1, \dots, n_X, \\ [\mathbf{g}]_{\iota(\boldsymbol{\beta})} &:= \int_{\Gamma} q p_{\boldsymbol{\beta}} d\mathbf{y}, \quad [\mathbf{f}]_j := \int_D f \phi_j d\mathbf{x}, \end{aligned}$$

where $\iota : \mathcal{P} \rightarrow \{1, \dots, \#\mathcal{P}\}$ is a bijection. Thus, the $[i + (\iota(\boldsymbol{\alpha}) - 1)n_X]$ -th entry of the solution vector \mathbf{u} is given by $u_{i, \boldsymbol{\alpha}}$.

¹Note that if $P_{\mathcal{P}}$ is a set of complete polynomials of total degree $\leq d$, then $P_{\mathcal{N}(\mathcal{P}, \mathcal{P})}$ is a set of complete polynomials of total degree $\leq 2d$.

4.2 Auxiliary bilinear forms

An important ingredient of the error estimation strategy described in section 3 is the auxiliary bilinear form $\tilde{B}(\cdot, \cdot)$. In this subsection, we consider two choices of $\tilde{B}(\cdot, \cdot)$, which both exploit the gPC expansion (4.1) of the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$.

The first auxiliary bilinear form employs the parameter-free part $t_0(\mathbf{x})$ in the expansion (4.1) of $T(\mathbf{x}, \mathbf{y})$:

$$B_0(u, v) := \int_{\Gamma} q(\mathbf{y}) \int_D t_0(\mathbf{x}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y}. \quad (4.8)$$

The auxiliary bilinear form of this type has been used in the a posteriori error analysis of the sGFEM for problem (2.1) with the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ having affine dependence on y_m (see, e.g., [2, 5, 4]).

Since $\int_{\Gamma} q(\mathbf{y}) \, d\mathbf{y} = 1$ and $T(\mathbf{x}, \mathbf{y})$ is bounded (see (2.5)), we deduce from (4.2) that

$$\alpha_{\min} \leq t_0(\mathbf{x}) \leq \alpha_{\max} \quad \forall \mathbf{x} \in D. \quad (4.9)$$

Hence, the symmetric bilinear form $B_0(\cdot, \cdot)$ is continuous and elliptic on V . Therefore, it defines an inner product in V which induces the norm $\|v\|_{B_0} := B_0(v, v)^{1/2}$ that is equivalent to $\|v\|_V$. Specifically, using (4.9), we obtain

$$\alpha_{\min} \|v\|_V^2 \leq \|v\|_{B_0}^2 \leq \alpha_{\max} \|v\|_V^2 \quad \forall v \in V. \quad (4.10)$$

Furthermore, using (4.10) together with (2.6), we show that the norm equivalence in (3.10) holds with $\tilde{B} = B_0$, $\lambda = \sqrt{\frac{\alpha_{\min}}{\alpha_{\max}}}$ and $\Lambda = \sqrt{\frac{\alpha_{\max}}{\alpha_{\min}}}$.

Turning now to the error estimators $e_{Y\mathcal{P}}$ and $e_{X\mathcal{Q}}$ that are defined in (3.12) and (3.13) by employing the bilinear form $\tilde{B} = B_0$, we use the same arguments as in §4.1 (see (4.4)–(4.6)) to rewrite (3.12) and (3.13) as follows:

$$B_0(e_{Y\mathcal{P}}, v) = F(v) - \int_{\Gamma} q \int_D \left(\sum_{\gamma \in \mathcal{N}(\mathcal{P}, \mathcal{P})} t_{\gamma} p_{\gamma} \right) \nabla u_{X\mathcal{P}} \cdot \nabla v \, d\mathbf{x} \, d\mathbf{y} \quad \forall v \in V_{Y\mathcal{P}}, \quad (4.11)$$

$$B_0(e_{X\mathcal{Q}}, v) = F(v) - \int_{\Gamma} q \int_D \left(\sum_{\gamma \in \mathcal{N}(\mathcal{P}, \mathcal{Q})} t_{\gamma} p_{\gamma} \right) \nabla u_{X\mathcal{P}} \cdot \nabla v \, d\mathbf{x} \, d\mathbf{y} \quad \forall v \in V_{X\mathcal{Q}}. \quad (4.12)$$

Furthermore, the definition of the bilinear form $B_0(\cdot, \cdot)$ in (4.8) and the orthogonality of the polynomial basis $\{p_{\gamma}\}_{\gamma \in \mathcal{Q}}$ (with respect to the inner product $\langle \cdot, \cdot \rangle_{\pi}$) imply the B_0 -orthogonality of the direct sum decomposition $V_{X\mathcal{Q}} = \bigoplus_{\mu \in \mathcal{Q}} X \otimes P_{\{\mu\}}$ (cf. (3.18)). Therefore, by Corollary 3.1, the error estimator $e_{X\mathcal{Q}}$ and its norm $\|e_{X\mathcal{Q}}\|_{B_0}$ can be decomposed into the contributions associated with individual indices $\mu \in \mathcal{Q}$, see (3.19) and (3.20) with $\tilde{B} = B_0$.

The construction of the auxiliary bilinear form $\tilde{B}(\cdot, \cdot)$ can be linked to designing a preconditioner for the coefficient matrix associated with the bilinear form $B(\cdot, \cdot)$. Indeed, the coefficient matrix associated with the auxiliary bilinear form $B_0(\cdot, \cdot)$ has been used in

many works as a preconditioner (called the *mean-based* preconditioner) for linear systems resulting from sGFEM formulations of parametric PDE problems (see, e.g., [17, 18]). Conversely, if there exists a good preconditioner for the coefficient matrix associated with bilinear form $B(\cdot, \cdot)$, then one can try to design the auxiliary bilinear form by mimicking the structure of that preconditioner. The above reasoning motivates our second choice of the auxiliary bilinear form $\tilde{B}(\cdot, \cdot)$. Specifically, motivated by the Kronecker product structure of the preconditioner proposed in [21], we construct the following bilinear form:

$$B_1(u, v) := \sum_{\gamma \in \mathbb{N}_0^M} \int_{\Gamma} q(\mathbf{y}) p_{\gamma}(\mathbf{y}) \int_D C_{\gamma} t_0(\mathbf{x}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y}, \quad (4.13)$$

where $C_{\gamma} \in \mathbb{R}$ are chosen to minimize the quantity $\mathcal{S} := \left\| \sum_{\gamma \in \mathbb{N}_0^M} C_{\gamma} t_0 p_{\gamma} - T \right\|_W^2$. Using the expansion (4.1) of T , we rewrite \mathcal{S} as follows:

$$\mathcal{S} = \left\| \sum_{\gamma \in \mathbb{N}_0^M} (C_{\gamma} t_0 - t_{\gamma}) p_{\gamma} \right\|_W^2 = \int_{\Gamma} q(\mathbf{y}) \int_D \left(\sum_{\gamma \in \mathbb{N}_0^M} (C_{\gamma} t_0(\mathbf{x}) - t_{\gamma}(\mathbf{x})) p_{\gamma}(\mathbf{y}) \right)^2 \, d\mathbf{x} \, d\mathbf{y}.$$

Hence, the values of C_{γ} can be found from the following equation:

$$\begin{aligned} \frac{\partial \mathcal{S}}{\partial C_{\gamma}} &= \frac{\partial \left(\int_{\Gamma} q(\mathbf{y}) \int_D \left(\sum_{\gamma' \in \mathbb{N}_0^M} (C_{\gamma'} t_0(\mathbf{x}) - t_{\gamma'}(\mathbf{x})) p_{\gamma'}(\mathbf{y}) \right)^2 \, d\mathbf{x} \, d\mathbf{y} \right)}{\partial C_{\gamma}} \\ &= 2 \sum_{\gamma' \in \mathbb{N}_0^M} \int_{\Gamma} q(\mathbf{y}) p_{\gamma'}(\mathbf{y}) p_{\gamma}(\mathbf{y}) \, d\mathbf{y} \int_D (C_{\gamma'} t_0(\mathbf{x}) - t_{\gamma'}(\mathbf{x})) t_0(\mathbf{x}) \, d\mathbf{x} \\ &= 2 \int_D (C_{\gamma} t_0(\mathbf{x}) - t_{\gamma}(\mathbf{x})) t_0(\mathbf{x}) \, d\mathbf{x} = 0 \quad \forall \gamma \in \mathbb{N}_0^M. \end{aligned}$$

As a result, we have

$$C_{\gamma} = \frac{\int_D t_{\gamma}(\mathbf{x}) t_0(\mathbf{x}) \, d\mathbf{x}}{\|t_0\|_{L^2(D)}^2} \quad \forall \gamma \in \mathbb{N}_0^M \quad (4.14)$$

(note that with these values of C_{γ} , one has $\left\| \sum_{\gamma \in \mathbb{N}_0^M} C_{\gamma} t_0 p_{\gamma} \right\|_W \leq \|T\|_W < +\infty$).

Substituting (4.14) into (4.13) and using (4.1), we rewrite $B_1(u, v)$ as follows:

$$\begin{aligned} B_1(u, v) &= \sum_{\gamma \in \mathbb{N}_0^M} \int_{\Gamma} q(\mathbf{y}) p_{\gamma}(\mathbf{y}) \int_D \frac{\int_D t_{\gamma}(\mathbf{x}') t_0(\mathbf{x}') \, d\mathbf{x}'}{\|t_0\|_{L^2(D)}^2} t_0(\mathbf{x}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \\ &= \frac{\int_{\Gamma} q(\mathbf{y}) \int_D \int_D t_0(\mathbf{x}') t_0(\mathbf{x}) T(\mathbf{x}', \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y}) \cdot \nabla v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x}' \, d\mathbf{x} \, d\mathbf{y}}{\|t_0\|_{L^2(D)}^2}. \end{aligned} \quad (4.15)$$

Using this representation of $B_1(\cdot, \cdot)$ as well as the boundedness of $T(\mathbf{x}, \mathbf{y})$ and $t_0(\mathbf{x})$ (see (2.5) and (4.9), resp.), we conclude that $B_1(\cdot, \cdot)$ defines an inner product in V which induces the norm $\|v\|_{B_1} := B_1(v, v)^{1/2}$ that is equivalent to $\|v\|_V$. In particular, there holds

$$\frac{\alpha_{\min}^3}{\alpha_{\max}^2} \|v\|_V^2 \leq \|v\|_{B_1}^2 \leq \frac{\alpha_{\max}^3}{\alpha_{\min}^2} \|v\|_V^2 \quad \forall v \in V. \quad (4.16)$$

Furthermore, using (4.16) together with (2.6) shows that the norm equivalence in (3.10) holds with $\tilde{B} = B_1$, $\lambda = \left(\frac{\alpha_{\min}}{\alpha_{\max}}\right)^{3/2}$ and $\Lambda = \left(\frac{\alpha_{\max}}{\alpha_{\min}}\right)^{3/2}$.

If the bilinear form $B_1(\cdot, \cdot)$ is employed to define the error estimators $e_{Y\mathcal{P}} \in V_{Y\mathcal{P}}$ and $e_{X\mathcal{Q}} \in V_{X\mathcal{Q}}$, then the associated discrete formulations (3.12) and (3.13) can be rewritten as follows (here, we use the same arguments as in §4.1):

$$\begin{aligned} & \int_{\Gamma} q \cdot \left(\sum_{\gamma \in \mathcal{N}(\mathcal{P}, \mathcal{P})} C_{\gamma p_{\gamma}} \right) \int_D t_0 \nabla e_{Y\mathcal{P}} \cdot \nabla v \, d\mathbf{x} \, d\mathbf{y} \\ &= F(v) - \int_{\Gamma} q \int_D \left(\sum_{\gamma \in \mathcal{N}(\mathcal{P}, \mathcal{P})} t_{\gamma p_{\gamma}} \right) \nabla u_{X\mathcal{P}} \cdot \nabla v \, d\mathbf{x} \, d\mathbf{y} \quad \forall v \in V_{Y\mathcal{P}}, \end{aligned} \quad (4.17)$$

$$\begin{aligned} & \int_{\Gamma} q \cdot \left(\sum_{\gamma \in \mathcal{N}(\mathcal{Q}, \mathcal{Q})} C_{\gamma p_{\gamma}} \right) \int_D t_0 \nabla e_{X\mathcal{Q}} \cdot \nabla v \, d\mathbf{x} \, d\mathbf{y} \\ &= F(v) - \int_{\Gamma} q \int_D \left(\sum_{\gamma \in \mathcal{N}(\mathcal{P}, \mathcal{Q})} t_{\gamma p_{\gamma}} \right) \nabla u_{X\mathcal{P}} \cdot \nabla v \, d\mathbf{x} \, d\mathbf{y} \quad \forall v \in V_{X\mathcal{Q}}. \end{aligned} \quad (4.18)$$

Comparing the left-hand sides in (4.17), (4.18) with those in (4.11), (4.12), respectively, it is easy to see that the computational cost associated with assembling linear systems for computing the error estimators $e_{Y\mathcal{P}}$ and $e_{X\mathcal{Q}}$ will be significantly lower if the bilinear form B_0 is employed to define these estimators.

4.3 Detail index set

We now discuss the construction of the detail index set \mathcal{Q} for computing the error estimator $e_{X\mathcal{Q}}$ defined by (3.13) in the case when the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$ is given by its gPC expansion (4.1). Let $\mathcal{J} \subset \mathbb{N}_0^M$ denote the index set such that all non-zero terms in expansion (4.1) are indexed by $\gamma \in \mathcal{J}$. We will distinguish between two cases: (i) \mathcal{J} is a *finite* index set; and (ii) \mathcal{J} is an *infinite* (countable) set.

If the auxiliary bilinear form \tilde{B} satisfies (3.18) (which is the case when $\tilde{B} = B_0$), then by Corollary 3.1, the estimator $e_{X\mathcal{Q}}$ is the sum of individual estimators $e_{X\mathcal{Q}}^{(\boldsymbol{\mu})}$ ($\boldsymbol{\mu} \in \mathcal{Q}$) satisfying (3.20). In this case, for a given $\boldsymbol{\mu} \in \mathcal{Q}$, $e_{X\mathcal{Q}}^{(\boldsymbol{\mu})} = 0$ if and only if the right-hand side of (3.20) is equal to zero for all $v \in X \otimes P_{\{\boldsymbol{\mu}\}}$, which is equivalent to $B(u_{X\mathcal{P}}, v) = 0$ for all $v \in X \otimes P_{\{\boldsymbol{\mu}\}}$ (note that $F(v) = 0$ for all $v \in X \otimes P_{\{\boldsymbol{\mu}\}}$, since $\mathbf{0} \notin \mathcal{Q}$ and hence $\boldsymbol{\mu} \neq \mathbf{0}$). Assume that \mathcal{J} is a finite index set. Then, recalling the definition of $\mathcal{N}(\cdot, \cdot)$ in (4.6) and the orthogonality property (4.4), we conclude that $e_{X\mathcal{Q}}^{(\boldsymbol{\mu})} = 0$ for any $\boldsymbol{\mu} \in \mathbb{N}_0^M \setminus \mathcal{N}(\mathcal{P}, \mathcal{J})$. Therefore, for a finite index set \mathcal{J} and an auxiliary bilinear form \tilde{B} satisfying (3.18), a natural choice of the detail index set is $\mathcal{Q} := \mathcal{N}(\mathcal{P}, \mathcal{J}) \setminus \mathcal{P}$.

If \tilde{B} does not satisfy (3.18) (which is the case when $\tilde{B} = B_1$) or \mathcal{J} is an infinite index set, then, in general, we can only build the *finite* detail index set \mathcal{Q} heuristically.

5 Numerical experiments: error estimation

The aim of this section is to test the error estimation strategy from §3 for the model problem (2.1) with a non-affine parametric representation of the diffusion coefficient. To that end, we set $f(\mathbf{x}) = 1$ and $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$, where $a(\mathbf{x}, \mathbf{y})$ is represented as follows:

$$a(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \sum_{m=1}^M a_m(\mathbf{x})y_m, \quad \mathbf{x} \in D, \mathbf{y} \in \Gamma. \quad (5.1)$$

Here, we assume that y_m are the images of *independent and identically distributed* random variables that follow the same truncated Gaussian probability density function

$$q_m(y_m) = \frac{\exp(-y_m^2/2\sigma_0^2)}{\sigma_0\sqrt{2\pi} \operatorname{erf}(1/\sqrt{2}\sigma_0)}, \quad (5.2)$$

where $\operatorname{erf}(\cdot)$ is the error function and σ_0 is a parameter of the truncated Gaussian distribution measuring the standard deviation.

Note that for $T = \exp(a)$ and a given by (5.1), the gPC expansion (4.1) has infinite number of non-zero terms; the formulae for calculating the expansion coefficients t_γ in this case are given in Appendix A. The following two examples of decompositions of $a(\mathbf{x}, \mathbf{y})$ are considered in our experiments.

Example 5.1. Let $D = (-1, 1)^2$. We assume that $a(\mathbf{x}, \mathbf{y})$ is represented by a truncated Karhunen–Loève expansion of a second-order random field with the mean $\mathbb{E}[a] = 1$ and the covariance function given by

$$\operatorname{Cov}[a](\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{|x_1 - x'_1|}{\ell_1} - \frac{|x_2 - x'_2|}{\ell_2}\right), \quad (5.3)$$

where σ is the standard deviation and ℓ_1, ℓ_2 are correlation lengths (we set $\ell_1 = \ell_2 = 1$).

Thus, in (5.1), we have: $a_0 = 1$ and $a_m(\mathbf{x}) = \sqrt{\lambda_m}\varphi_m(\mathbf{x})$ ($m = 1, \dots, M$), where $\{(\lambda_m, \varphi_m)\}_{m=1}^\infty$ are the eigenpairs of the integral operator $\int_D \operatorname{Cov}[a](\mathbf{x}, \mathbf{x}')\varphi(\mathbf{x}') d\mathbf{x}'$ (see, e.g. [15, pp. 28–29]).

Example 5.2. Let $D = (0, 1)^2$, $a_0 = 1$ and choose the spatial coefficient functions $a_m(\mathbf{x})$ ($m = 1, \dots, M$) in (5.1) as those introduced in [8, section 11]:

$$a_m(\mathbf{x}) = \bar{\alpha}m^{-\bar{\sigma}} \cos(2\pi\bar{\beta}_1(m)x_1) \cos(2\pi\bar{\beta}_2(m)x_2), \quad \mathbf{x} = (x_1, x_2) \in D. \quad (5.4)$$

Here, $\bar{\sigma} > 1$ characterizes the decay rate of the amplitudes $\bar{\alpha}m^{-\bar{\sigma}}$ of these coefficients (we set $\bar{\sigma} = 2$ in our experiments), $\bar{\alpha} > 0$, and $\bar{\beta}_1, \bar{\beta}_2$ are defined as

$$\bar{\beta}_1(m) = m - \bar{k}(m)(\bar{k}(m) + 1)/2 \quad \text{and} \quad \bar{\beta}_2(m) = \bar{k}(m) - \bar{\beta}_1(m)$$

with $\bar{k}(m) = \lfloor -1/2 + \sqrt{1/4 + 2m} \rfloor$.

All experiments in this section and in section 7 were performed using the open source MATLAB toolbox S-IFISS [20]. In our computations, we use the finite element space $X = X(h)$ of bilinear (Q_1) approximations on uniform grids \square_h of square elements with edge length h . In this case, the detail finite element space $Y = Y(h)$ is the span of the set of bilinear bubble functions corresponding to edge midpoints and element centroids of the grid. For the polynomial approximation on Γ , we first construct a polynomial basis in $L^2_\pi(\Gamma)$ by tensorizing univariate orthonormal polynomials generated by the probability density function (5.2) (these polynomials are known in the literature as Rys polynomials, see, e.g., [14, Example 1.11]); then we employ the set $P_{M,d}$ of complete polynomials of degree $\leq d$ in M variables, $P_{M,d} := \text{span} \{p_\alpha; \alpha \in \mathcal{P}_{M,d}\}$, where

$$\mathcal{P}_{M,d} := \left\{ \alpha = (\alpha_1, \dots, \alpha_M) \in \mathbb{N}_0^M; \alpha_1 + \dots + \alpha_M \leq d \right\}.$$

Thus, given h , M and d , we compute the Galerkin approximation $u_{XP} \in X(h) \otimes P_{M,d}$ satisfying (3.1).

The spatial error estimator e_{YP} satisfying (3.12) is computed approximately by using a standard element residual technique (see, e.g., [1]). Specifically, we solve the following local residual problems associated with (3.12): find $e_{YP}|_S \in Y(h)|_S \otimes P_{\mathcal{P}}$ satisfying

$$\begin{aligned} \tilde{B}_S(e_{YP}|_S, v) &= F_S(v) + \int_\Gamma q(\mathbf{y}) \int_S \nabla \cdot (T(\mathbf{x}, \mathbf{y}) \nabla u_{XP}(\mathbf{x}, \mathbf{y})) v(\mathbf{x}, \mathbf{y}) \, d\mathbf{x} \, d\mathbf{y} \\ &\quad - \frac{1}{2} \int_\Gamma q(\mathbf{y}) \int_{\partial S \setminus \partial D} T(s, \mathbf{y}) \left[\left[\frac{\partial u_{XP}}{\partial n} \right] \right] v(s, \mathbf{y}) \, ds \, d\mathbf{y} \end{aligned} \quad (5.5)$$

for any $v \in Y(h)|_S \otimes P_{\mathcal{P}}$. Here, \tilde{B}_S and F_S denote the elementwise auxiliary bilinear form and linear functional, respectively; $Y(h)|_S$ is the restriction of $Y(h)$ to the element $S \in \square_h$; and $\left[\left[\frac{\partial u_{XP}}{\partial n} \right] \right]$ denotes the flux jump in the approximate solution u_{XP} across interelement edges. The parametric error estimator e_{XQ} is computed by solving (3.13) (see also (3.19)–(3.20) in the case $\tilde{B} = B_0$). Then two total error estimates are computed as follows (see (3.14) with $\tilde{B} = B_0$ and $\tilde{B} = B_1$, resp.):

$$\eta_0 = \left(\sum_{S \in \square_h} \|e_{YP}|_S\|_{B_0,S}^2 + \sum_{\mu \in \mathcal{Q}} \|e_{XQ}^{(\mu)}\|_{B_0}^2 \right)^{1/2}, \quad \eta_1 = \left(\sum_{S \in \square_h} \|e_{YP}|_S\|_{B_1,S}^2 + \|e_{XQ}\|_{B_1}^2 \right)^{1/2}. \quad (5.6)$$

In the experiments below, we will examine the quality of the error estimates η_0 and η_1 by computing the corresponding effectivity indices

$$\Theta_i := \frac{\eta_i}{\sqrt{\|u_{\text{ref}}\|_B^2 - \|u_{XP}\|_B^2}}, \quad i = 0, 1, \quad (5.7)$$

where $u_{\text{ref}} \in X(h_{\text{ref}}) \otimes P_{M,d_{\text{ref}}}$ is an accurate (reference) solution computed using bi-quadratic (Q_2) approximations on a uniform grid $\square_{h_{\text{ref}}}$ with $h_{\text{ref}} < h$ and an enriched polynomial space $P_{M,d_{\text{ref}}}$ with $d_{\text{ref}} > d$.

h	$\sigma = 0.2$		$\sigma = 0.4$		$\sigma = 0.6$		$\sigma = 0.8$	
	Θ_0	Θ_1	Θ_0	Θ_1	Θ_0	Θ_1	Θ_0	Θ_1
2^{-1}	1.3275	1.3238	1.3331	1.3186	1.3434	1.3125	1.3599	1.3093
2^{-2}	1.1370	1.1331	1.1531	1.1382	1.1807	1.1496	1.2206	1.1714
2^{-3}	1.0198	1.0162	1.0389	1.0254	1.0715	1.0439	1.1179	1.0754
2^{-4}	0.9513	0.9480	0.9708	0.9586	1.0042	0.9797	1.0515	1.0143
2^{-5}	0.9134	0.9104	0.9329	0.9215	0.9668	0.9441	1.0142	0.9793

Table 1: The effectivity indices for Galerkin approximations $u_{X\mathcal{P}} \in X(h) \otimes P_{3,2}$ for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.1 with $\ell_1 = \ell_2 = 1$. The fixed detail index set $\mathcal{Q} = \mathcal{P}_{3,6} \setminus \mathcal{P}_{3,2}$ is employed to compute the underlying error estimates.

In the experiments below, we set $\sigma_0 = 1$ in (5.2) and fix $M = 3$, $d = 2$.

In the first set of experiments, we consider two model problems described above and vary the parameters that characterize the magnitude of the spatial coefficient functions a_m in (5.1) (i.e., the parameters σ and $\bar{\alpha}$ in Examples 5.1 and 5.2, respectively). Specifically, we choose $\sigma, \bar{\alpha} \in \{0.2, 0.4, 0.6, 0.8\}$. For each problem, we use spatial grids of decreasing mesh size $h = 2^{-j} \sqrt{|D|}$ ($j = 2, \dots, 6$) to compute a sequence of Galerkin approximations $u_{X\mathcal{P}}$ and the corresponding error estimates η_0, η_1 defined in (5.6). In particular, the parametric error estimators $e_{X\mathcal{Q}}$ are computed with the detail index set $\mathcal{Q} = \mathcal{P}_{3,6} \setminus \mathcal{P}_{3,2}$. For each computed error estimate, we calculate the effectivity index via (5.7). Here, we use the reference solutions $u_{\text{ref}} \in X(h_{\text{ref}}) \otimes P_{3,4}$, where we choose $h_{\text{ref}} = 2^{-7} \sqrt{|D|}$. The results of these computations are presented in Table 1 (for the decomposition of $a(\mathbf{x}, \mathbf{y})$ in Example 5.1) and in Table 2 (for the decomposition in Example 5.2).

From Tables 1 and 2 we find that both effectivity indices Θ_0 and Θ_1 are close to unity and decrease as the spatial grid is refined or the corresponding coefficient parameter (σ or $\bar{\alpha}$) decreases. We also observe that $\Theta_0 > \Theta_1$ in each case, and the difference between Θ_0 and Θ_1 grows as σ and $\bar{\alpha}$ increase.

In the second set of experiments, we consider the same model problems as in the first set of experiments but choose larger problem parameters, namely $\sigma, \bar{\alpha} \in \{1, 3, 5\}$. In each case, we compute the Galerkin approximation $u_{X\mathcal{P}} \in X(h) \otimes P_{3,2}$ with fixed $h = 2^{-5} \sqrt{|D|}$. For each Galerkin approximation, two sequences of error estimates $\{\eta_0\}$ and $\{\eta_1\}$ are computed with different detail index sets; specifically, we use $\mathcal{Q} = \mathcal{P}_{3,\bar{d}} \setminus \mathcal{P}_{3,2}$ with $\bar{d} \in \{3, 4, \dots, 7\}$. Then, the effectivity index is calculated for each error estimate; here, we again use the corresponding reference solutions $u_{\text{ref}} \in X(h_{\text{ref}}) \otimes P_{3,4}$ with $h_{\text{ref}} = 2^{-7} \sqrt{|D|}$. The effectivity indices are reported in Table 3 (for the decomposition of $a(\mathbf{x}, \mathbf{y})$ in Example 5.1) and in Table 4 (for the decomposition in Example 5.2).

From Tables 3 and 4 we again observe that $\Theta_0 > \Theta_1$ for each fixed σ (resp., $\bar{\alpha}$) and

h	$\bar{\alpha} = 0.2$		$\bar{\alpha} = 0.4$		$\bar{\alpha} = 0.6$		$\bar{\alpha} = 0.8$	
	Θ_0	Θ_1	Θ_0	Θ_1	Θ_0	Θ_1	Θ_0	Θ_1
2^{-2}	1.3785	1.3783	1.5197	1.5177	1.7139	1.7055	1.9284	1.9061
2^{-3}	1.1806	1.1805	1.3144	1.3126	1.5056	1.4975	1.7288	1.7066
2^{-4}	1.0674	1.0673	1.2133	1.2113	1.4193	1.4108	1.6575	1.6344
2^{-5}	1.0029	0.0028	1.1583	1.1563	1.3742	1.3655	1.6177	1.5941
2^{-6}	0.9678	0.9676	1.1289	1.1268	1.3486	1.3397	1.5808	1.5571

Table 2: The effectivity indices for Galerkin approximations $u_{XP} \in X(h) \otimes P_{3,2}$ for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.2 with $\bar{\sigma} = 2$. The fixed detail index set $\mathcal{Q} = \mathcal{P}_{3,6} \setminus \mathcal{P}_{3,2}$ is employed to compute the underlying error estimates.

σ	$\mathcal{P}_{3,3} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,4} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,5} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,6} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,7} \setminus \mathcal{P}_{3,2}$	
	Θ_0	Θ_1								
1	1.1027	1.0591	1.1075	1.0601	1.1078	1.0601	1.1078	1.0601	1.1078	1.0601
3	0.7658	0.6889	1.0578	0.7796	1.1469	0.7849	1.1634	0.7862	1.1654	0.7863
5	0.4398	0.3543	0.8163	0.2236	1.0517	0.5773	1.1512	0.5877	1.1813	0.5900

Table 3: The effectivity indices for Galerkin approximations $u_{XP} \in X(h) \otimes P_{3,2}$ with $h = 2^{-4}$ for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.1 with $\ell_1 = \ell_2 = 1$. The sequence of expanded index sets $\mathcal{Q} = \mathcal{P}_{3,\bar{d}} \setminus \mathcal{P}_{3,2}$ with $\bar{d} \in \{3, 4, \dots, 7\}$ is employed to compute the underlying error estimates.

for each detail index set. Furthermore, for fixed σ and $\bar{\alpha}$, the effectivity indices in each sequence $\{\Theta_0\}$ and $\{\Theta_1\}$ approach their limiting values as the detail index set expands. This convergence to limiting values is faster for smaller values of σ and $\bar{\alpha}$. For all σ in Table 3, the limiting values of Θ_0 stay close to unity, whereas the limiting values of Θ_1 decrease rapidly away from unity as σ increases (see the last two columns in Table 3). This shows a robustness of the error estimate η_0 with respect to the ‘roughness’ of the parametric coefficient in Example 5.1. This difference between the limiting values of Θ_0 and Θ_1 is less pronounced for the parametric coefficient in Example 5.2 for given values of $\bar{\alpha}$ (see the last two columns in Table 4). We can see, however, a faster decay of Θ_1 as $\bar{\alpha}$ increases, which indicates a deterioration of quality of the error estimate η_1 for larger values of $\bar{\alpha}$.

Based on the numerical results reported in this section, we conclude that the bilinear form $\tilde{B} = B_0$ is preferable to the bilinear form $\tilde{B} = B_1$ for estimating the energy errors in sGFEM approximations for problems with non-affine parametric representations of coefficients. Indeed, it follows from the numerical comparison of the associated effectivity

$\bar{\alpha}$	$\mathcal{P}_{3,3} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,4} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,5} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,6} \setminus \mathcal{P}_{3,2}$		$\mathcal{P}_{3,7} \setminus \mathcal{P}_{3,2}$	
	Θ_0	Θ_1								
1	1.8582	1.8090	1.8589	1.8097	1.8589	1.8097	1.8589	1.8097	1.8589	1.8097
3	1.6786	1.3352	1.7779	1.4150	1.7996	1.4276	1.8038	1.4331	1.8042	1.4341
5	0.9080	0.7911	1.0825	0.8825	1.1786	0.9253	1.2218	0.9413	1.2353	0.9517

Table 4: The effectivity indices for Galerkin approximations $u_{X\mathcal{P}} \in X(h) \otimes P_{3,2}$ with $h = 2^{-5}$ for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.2 with $\bar{\sigma} = 2$. The sequence of expanded index sets $\mathcal{Q} = \mathcal{P}_{3,\bar{d}} \setminus \mathcal{P}_{3,2}$ with $\bar{d} \in \{3, 4, \dots, 7\}$ is employed to compute the underlying error estimates.

indices that the quality of the error estimate η_0 that employs B_0 is, in general, not worse than that of the error estimate η_1 employing B_1 . Furthermore, as emphasized in §4.2, using the bilinear form B_0 is also preferable from the computational cost point of view. In addition to that, the B_0 -orthogonality of the direct sum $\bigoplus_{\mu \in \mathcal{Q}} X \otimes P_{\{\mu\}}$ gives immediate access to individual parametric estimators $e_{X\mathcal{Q}}^{(\mu)}$ ($\mu \in \mathcal{Q}$), which is critical for building adaptive polynomial approximations on the parameter domain. All this motivates the choice of the auxiliary bilinear form \tilde{B} and the associated energy error estimators in the adaptive algorithm presented in the next section.

6 Adaptive algorithm

In this section, we present an adaptive solution algorithm for the model problem (2.1). Our focus here is on effective enrichment of the polynomial space in the parameter domain. We follow the ideas developed in [5] but use Dörfler marking for enriching polynomial approximations on Γ and employ the error reduction estimates for *marked* polynomial basis functions in order to choose the refinement type (spatial *vs.* parametric) at each iteration step (cf. [4, section 5]). The choice of Dörfler marking is motivated by the fact that it facilitates convergence analysis of adaptive algorithms (cf. [6, 7, 3]); in particular, linear convergence of adaptive stochastic Galerkin approximations is only proved in the case of Dörfler marking; see [7, Theorem 7.2] and [3, Theorem 8].

Starting with a coarse grid of edge length h_0 and an initial index set $\mathcal{P}_0 \supseteq \mathcal{P}_{M,1}$, the adaptive algorithm generates a sequence of finite element spaces

$$X(h_0) \subseteq X(h_1) \subseteq X(h_2) \subseteq \dots \subseteq X(h_K) \subset H_0^1(D),$$

a sequence of polynomial spaces

$$P_{\mathcal{P}_0} \subseteq P_{\mathcal{P}_1} \subseteq P_{\mathcal{P}_2} \subseteq \dots \subseteq P_{\mathcal{P}_K} \subset L_\pi^2(\Gamma),$$

and a sequence of Galerkin solutions $u^{(k)} \in V_{X\mathcal{P}}^k := X(h_k) \otimes P_{\mathcal{P}_k}$.

At each iteration step k , the Galerkin solution $u^{(k)}$ satisfying (3.1) is computed by the subroutine SOLVE as follows:

$$u^{(k)} = \text{SOLVE}(T, f, h_k, \mathcal{P}_k),$$

where T and f are the problem data (see (2.1)).

At the error estimation step, we choose $\tilde{B} = B_0$. With this choice of \tilde{B} , we use (5.5) to compute the local (spatial) estimators $\{e_{Y\mathcal{P}}|_S\}_{S \in \square_{h_k}}$ and employ (4.12) to compute the parametric estimator $e_{X\mathcal{Q}}$; the latter gives access to individual parametric estimators $\{e_{X\mathcal{Q}}^{(\mu)}\}_{\mu \in \mathcal{Q}_k}$ due to (3.19). All estimators are computed by the subroutine ESTIMATE:

$$\left[\{e_{Y\mathcal{P}}|_S; S \in \square_{h_k}\}, \{e_{X\mathcal{Q}}^{(\mu)}; \mu \in \mathcal{Q}_k\} \right] = \text{ESTIMATE}(T, f, h_k, \mathcal{P}_k, \mathcal{Q}_k, u^{(k)}).$$

Here, as discussed in section 4.3, the detail index set is built as follows: if \mathcal{J} is finite, then the natural choice of \mathcal{Q}_k is $\mathcal{Q}_k = \mathcal{N}(\mathcal{P}_k, \mathcal{J}) \setminus \mathcal{P}_k$; if \mathcal{J} is infinite, then we build \mathcal{Q}_k heuristically as $\mathcal{Q}_k = \mathcal{N}(\mathcal{P}_k, \mathcal{N}(\mathcal{P}_k, \mathcal{P}_k)) \setminus \mathcal{P}_k$. Then we calculate the total error estimate $\eta^{(k)}$ via the first equation in (5.6).

Algorithm 6.1: Adaptive stochastic Galerkin finite element algorithm

Input: data T, f ; initial edge length h_0 , initial index set $\mathcal{P}_0 \supseteq \mathcal{P}_{M,1}$;

marking threshold $\theta_{\mathcal{P}}$; tolerance ϵ

Output: final Galerkin solution $u^{(K)}$, final error estimate $\eta^{(K)}$

for $k = 0, 1, 2, \dots$ **do**

$u^{(k)} = \text{SOLVE}(T, f, h_k, \mathcal{P}_k);$

$\left[\{e_{Y\mathcal{P}}|_S; S \in \square_{h_k}\}, \{e_{X\mathcal{Q}}^{(\mu)}; \mu \in \mathcal{Q}_k\} \right] = \text{ESTIMATE}(T, f, h_k, \mathcal{P}_k, \mathcal{Q}_k, u^{(k)});$

$\eta^{(k)} = \left(\sum_{S \in \square_{h_k}} \|e_{Y\mathcal{P}}|_S\|_{B_{0,S}}^2 + \sum_{\mu \in \mathcal{Q}_k} \|e_{X\mathcal{Q}}^{(\mu)}\|_{B_0}^2 \right)^{1/2};$

if $\eta^{(k)} < \epsilon$ **then**

$K := k; \text{break};$

else

$\mathcal{M}_k = \text{MARK}\left(\left\{ \|e_{X\mathcal{Q}}^{(\mu)}\|_{B_0}; \mu \in \mathcal{Q}_k \right\}, \theta_{\mathcal{P}}\right);$

if $\sum_{S \in \square_{h_k}} \|e_{Y\mathcal{P}}|_S\|_{B_{0,S}}^2 \geq \sum_{\mu \in \mathcal{M}_k} \|e_{X\mathcal{Q}}^{(\mu)}\|_{B_0}^2$ **then**

$h_{k+1} = h_k/2; \mathcal{P}_{k+1} = \mathcal{P}_k;$

else

$h_{k+1} = h_k; \mathcal{P}_{k+1} = \mathcal{P}_k \cup \mathcal{M}_k;$

end

end

end

If the error estimate $\eta^{(k)}$ exceeds the prescribed tolerance ϵ , then an enriched finite-dimensional space $V_{X\mathcal{P}}^{k+1} \supset V_{X\mathcal{P}}^k$ must be constructed. Before doing this, we identify those

σ	0.4	0.6	0.8	1
t , sec	1.8181e+02	5.4331e+02	4.1346e+03	5.7141e+03
K	4	5	6	6
$\eta^{(K)}$	1.8628e-02	1.6762e-02	1.1463e-02	1.4294e-02
h_K	2^{-4}	2^{-5}	2^{-5}	2^{-5}
$\#\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K)$	125	125	999	1,339
N_K	6,534	25,350	71,825	80,275
\mathcal{P}	$k = 0$ (0 0 0 0 0)	$k = 0$ (0 0 0 0 0)	$k = 0$ (0 0 0 0 0)	$k = 0$ (0 0 0 0 0)
	(0 0 0 0 1)	(0 0 0 0 1)	(0 0 0 0 1)	(0 0 0 0 1)
	(0 0 0 1 0)	(0 0 0 1 0)	(0 0 0 1 0)	(0 0 0 1 0)
	(0 0 1 0 0)	(0 0 1 0 0)	(0 0 1 0 0)	(0 0 1 0 0)
	(0 1 0 0 0)	(0 1 0 0 0)	(0 1 0 0 0)	(0 1 0 0 0)
	(1 0 0 0 0)	(1 0 0 0 0)	(1 0 0 0 0)	(1 0 0 0 0)
			$k = 5$ (2 0 0 0 0)	$k = 5$ (2 0 0 0 0)
			(1 1 0 0 0)	(1 1 0 0 0)
			(1 0 1 0 0)	(1 0 1 0 0)
			(1 0 0 1 0)	(1 0 0 1 0)
			(1 0 0 0 1)	(1 0 0 0 1)
			(0 1 1 0 0)	(0 1 1 0 0)
			(0 1 0 1 0)	(0 1 0 1 0)
			(0 0 1 0 1)	(0 0 1 0 1)
			(0 2 0 0 0)	(2 0 1 0 0)
			(0 0 2 0 0)	(2 1 0 0 0)
			(0 1 0 0 1)	(0 2 0 0 0)
				(0 0 2 0 0)
				(1 1 1 0 0)

Table 5: The results of running Algorithm 6.1 for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.1 with $\ell_1 = \ell_2 = 1$.

indices $\mu \in \mathcal{Q}_k$ that yield larger contributing estimators $e_{X\mathcal{Q}}^{(\mu)}$. To that end, we employ the Dörfler marking strategy [6]. Specifically, we fix a threshold parameter $\theta_{\mathcal{P}} \in (0, 1]$ and build a minimal subset $\mathcal{M}_k \subseteq \mathcal{Q}_k$ such that

$$\sum_{\mu \in \mathcal{M}_k} \|e_{X\mathcal{Q}}^{(\mu)}\|_{B_0}^2 \geq \theta_{\mathcal{P}} \sum_{\mu \in \mathcal{Q}_k} \|e_{X\mathcal{Q}}^{(\mu)}\|_{B_0}^2. \quad (6.1)$$

The marked index set is generated by the subroutine **MARK**:

$$\mathcal{M}_k = \text{MARK}\left(\left\{\|e_{X\mathcal{Q}}^{(\mu)}\|_{B_0}; \mu \in \mathcal{Q}_k\right\}, \theta_{\mathcal{P}}\right).$$

In order to construct the enriched approximation space, we either enrich the finite element space by uniformly refining the mesh (in this case, we define $V_{X\mathcal{P}}^{k+1,1} := X(h_{k+1}) \otimes P_{\mathcal{P}_k}$ with $h_{k+1} = h_k/2$), or enrich the polynomial space by including the (marked) indices from $\mathcal{M}_k \subseteq \mathcal{Q}_k$ (i.e., we set $V_{X\mathcal{P}}^{k+1,2} := X(h_k) \otimes P_{\mathcal{P}_{k+1}}$ with $\mathcal{P}_{k+1} = \mathcal{P}_k \cup \mathcal{M}_k$). Let

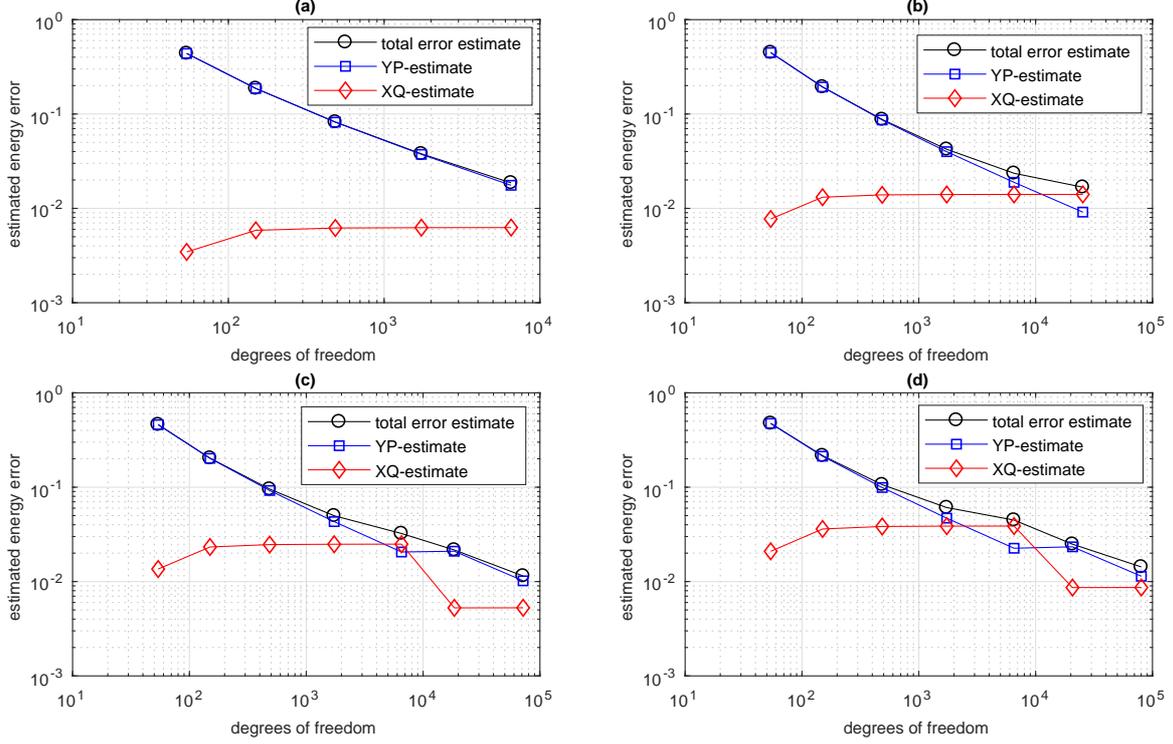


Figure 1: Energy error estimates at each step of the adaptive algorithm for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.1 with $\ell_1 = \ell_2 = 1$: (a) $\sigma = 0.4$; (b) $\sigma = 0.6$; (c) $\sigma = 0.8$; (d) $\sigma = 1$.

$u^{(k+1,l)} \in V_{XP}^{k+1,l}$ ($l = 1, 2$) denote the corresponding enhanced Galerkin approximations (note that none of these approximations is computed at this stage). In order to determine the refinement type (spatial or parametric), we recall that Theorem 3.2 implies that $\left(\sum_{S \in \square_h} \|e_{YP}|_S\|_{B_{0,S}}^2\right)^{1/2}$ and $\left(\sum_{\mu \in \mathcal{M}_k} \|e_{XQ}^{(\mu)}\|_{B_0}^2\right)^{1/2}$ provide effective estimates for the error reductions $\|u^{(k+1,1)} - u^{(k)}\|_B$ and $\|u^{(k+1,2)} - u^{(k)}\|_B$, respectively. Therefore, if $\left(\sum_{S \in \square_h} \|e_{YP}|_S\|_{B_{0,S}}^2\right)^{1/2}$ is greater than or equal to $\left(\sum_{\mu \in \mathcal{M}_k} \|e_{XQ}^{(\mu)}\|_{B_0}^2\right)^{1/2}$, we define $V_{XP}^{k+1} := V_{XP}^{k+1,1}$, leading to spatial refinement; otherwise, we set $V_{XP}^{k+1} := V_{XP}^{k+1,2}$, leading to parametric refinement. Then a more accurate Galerkin solution $u^{(k+1)} \in V_{XP}^{k+1}$ is computed. The process is then repeated until the tolerance is met.

The complete adaptive algorithm is listed in Algorithm 6.1.

7 Numerical experiments: adaptivity

In this section, we test the performance of Algorithm 6.1 for the model problem (2.1) with non-affine parametric representations of the diffusion coefficient. As in section 5, numerical results are presented for bilinear (Q1) spatial approximations on uniform grids \square_h of square elements with edge length h . In all experiments, we set the marking parameter $\theta_P = 0.9$ in (6.1) and run the adaptive algorithm with the stopping tolerance $\epsilon = 2 \times 10^{-2}$.

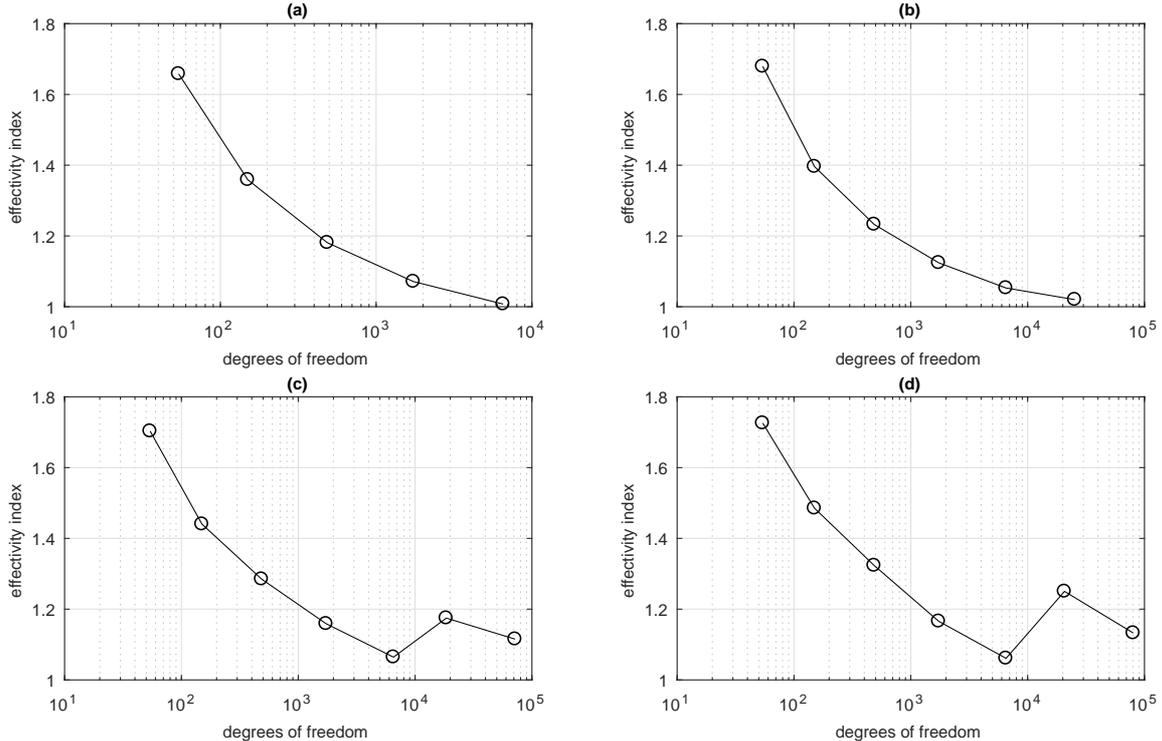


Figure 2: The effectivity indices for the Galerkin solutions at each iteration of the adaptive algorithm for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.1 with $\ell_1 = \ell_2 = 1$: (a) $\sigma = 0.4$; (b) $\sigma = 0.6$; (c) $\sigma = 0.8$; (d) $\sigma = 1$.

In our first set of experiments in this section, we consider the model problem (2.1) on the domain $D = (-1, 1)^2$ and we set $f(\mathbf{x}) = 1$, $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$, where $a(\mathbf{x}, \mathbf{y})$ is represented as in (5.1) by using the truncated Karhunen–Loève expansion in Example 5.1. As in section 5, we assume that y_m are the images of independent and identically distributed random variables that follow the truncated Gaussian probability density function in (5.2) with $\sigma_0 = 1$. We fix $M = 5$ in (5.1) and for each $\sigma \in \{0.4, 0.6, 0.8, 1\}$ in (5.3) we run the adaptive algorithm. The results of these computations are presented in Table 5 and Figures 1 and 2.

In Table 5, for each computation we report the overall computational time t (in seconds), the number of iterations K needed to reach the prescribed tolerance, the final error estimate $\eta^{(K)}$, the edge length h_K of the final mesh, the cardinality of the index set $\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K)$ that is used in calculating the estimator $e_{X\mathcal{Q}}$ (see (4.12)), the final number of degrees of freedom $N_K := \dim(V_{X\mathcal{P}}^K)$, and the evolution of the index set \mathcal{P} .

From Table 5, we find that in the experiments with larger values of σ , the tolerance is met by the final Galerkin solution calculated on a more refined spatial grid \square_{h_K} and with a larger index set \mathcal{P}_K . This leads to significant increase in computational times and is due to a dramatic expansion of the index set $\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K)$ as σ increases. For example, as σ increases from 0.6 to 0.8, the cardinality of $\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K)$ increases by approximately a factor

$\bar{\alpha}$	0.4	0.6	0.8	1
t , sec	1.5620e+01	1.8369e+01	3.2806e+01	1.1202e+02
K	4	5	6	8
$\eta^{(K)}$	1.4633e-02	1.8236e-02	1.6919e-02	1.8534e-02
h_K	2^{-5}	2^{-5}	2^{-6}	2^{-7}
$\#(\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K) \cap \mathcal{P}_{5,2})$	20	20	20	20
N_K	6,534	9,801	38,025	249,615
\mathcal{P}	$k = 0$ (0 0 0 0 0) (0 0 0 0 1) (0 0 0 1 0) (0 0 1 0 0) (0 1 0 0 0) (1 0 0 0 0)	$k = 0$ (0 0 0 0 0) (0 0 0 0 1) (0 0 0 1 0) (0 0 1 0 0) (0 1 0 0 0) (1 0 0 0 0)	$k = 0$ (0 0 0 0 0) (0 0 0 0 1) (0 0 0 1 0) (0 0 1 0 0) (0 1 0 0 0) (1 0 0 0 0)	$k = 0$ (0 0 0 0 0) (0 0 0 0 1) (0 0 0 1 0) (0 0 1 0 0) (0 1 0 0 0) (1 0 0 0 0)
		$k = 5$ (2 0 0 0 0) (1 1 0 0 0) (3 0 0 0 0)	$k = 5$ (2 0 0 0 0) (1 1 0 0 0) (3 0 0 0 0)	$k = 5$ (2 0 0 0 0) (1 1 0 0 0) (3 0 0 0 0)
				$k = 7$ (2 1 0 0 0) (4 0 0 0 0) (1 0 1 0 0) (5 0 0 0 0) (3 1 0 0 0) (1 0 0 1 0)

Table 6: The results of running Algorithm 6.1 for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = a^2(\mathbf{x}, \mathbf{y})$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.2 with $\bar{\sigma} = 2$.

of 8, while the cardinality of \mathcal{P}_K only increases by approximately a factor of 3. Greater cardinality of $\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K)$ means longer computational time for finding $e_{X\mathcal{Q}}$ via (4.12), taking a significant share of the overall computational time.

In Figure 1, we plot the error estimates η , $\|e_{Y\mathcal{P}}\|_{B_0}$ and $\|e_{X\mathcal{Q}}\|_{B_0}$ at each iteration of the adaptive loop. In Figure 1(a) (i.e., for $\sigma = 0.4$), we observe that $\|e_{Y\mathcal{P}}\|_{B_0}$ is greater than $\|e_{X\mathcal{Q}}\|_{B_0}$ throughout the computation, that is why no parametric refinement is performed in this case before the tolerance is met. In Figure 1(b) ($\sigma = 0.6$), we find that $\|e_{Y\mathcal{P}}\|_{B_0}$ is only smaller than $\|e_{X\mathcal{Q}}\|_{B_0}$ at the final iteration, when the total error estimate is below the tolerance; thus no parametric refinement is performed in this case either. In the experiments with $\sigma = 0.8$ and $\sigma = 1$, one parametric refinement is needed before the tolerance is met (see Figures 1(c) and 1(d)). Note that more indices were activated in the case of $\sigma = 1$.

In Figure 2, we plot the effectivity indices computed via (5.7) with $i = 0$ at each iteration of the algorithm. Here, the reference solution u_{ref} in each experiment is computed using biquadratic (Q2) spatial approximations on a fine grid $\square_{h_{\text{ref}}}$ with $h_{\text{ref}} = h_K/2$ and

employing the polynomial space P_{M,d_K+1} with d_K being the highest (total) degree of the polynomials in $P_{\mathcal{P}_K}$. We can see that for all experiments the effectivity indices are within the interval $(1, 2)$ throughout all iterations.

In the final set of experiments, we consider the model problem (2.1) on the domain $D = (0, 1)^2$ and we set $f(\mathbf{x}) = 1$, $T(\mathbf{x}, \mathbf{y}) = a^2(\mathbf{x}, \mathbf{y})$, where $a(\mathbf{x}, \mathbf{y})$ is represented as in (5.1) with the coefficient functions a_m ($m = 0, \dots, M$) chosen as in Example 5.2. We again assume that y_m are the images of independent and identically distributed random variables that follow the truncated Gaussian probability density function in (5.2) with $\sigma_0 = 1$. Note that for $T = a^2$ and a given by (5.1), the gPC expansion (4.1) has a finite number of non-zero terms; the formulae for calculating the expansion coefficients t_γ in this case are given in Appendix A. We fix $M = 5$ in (5.1) and for each $\bar{\alpha} \in \{0.4, 0.6, 0.8, 1\}$ in (5.4) we run the adaptive algorithm. The results of computations are presented in Table 6 and Figures 3 and 4.

From Table 6 and Figure 3, we find that no parametric refinement is performed in the experiment with $\bar{\alpha} = 0.4$; one parametric refinement is performed in the experiments with $\bar{\alpha} = 0.6$ and $\bar{\alpha} = 0.8$; and two parametric refinements are performed in the experiment with $\bar{\alpha} = 1$. Since for the model problem in this set of experiments, the gPC expansion (4.1) of the diffusion coefficient $T = a^2$ reduces to a finite sum over the index set $\mathcal{J} \subset \mathcal{P}_{5,2}$ (see Appendix A), the sum in (4.12) is over the set $\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K) \cap \mathcal{P}_{5,2}$. We observe from Table 6 that $\#\mathcal{N}(\mathcal{P}_K, \mathcal{Q}_K) \cap \mathcal{P}_{5,2}$ does not change throughout this set of experiments. This partly explains the reason why the overall computational times for larger values of coefficient parameter (i.e., the parameter $\bar{\alpha}$ in this set of experiments) do not increase as significantly as they do in the first set of experiments in this section.

In Figure 4, we plot the effectivity indices for the error estimate at each iteration of the algorithm (here, the reference Galerkin solution is computed similarly to other experiments). We can see that for all experiments in this set, the effectivity indices are within the interval $(0.5, 2.5)$ throughout all iterations.

8 Concluding remarks

Adaptivity is a critical ingredient of effective algorithms for numerical solution of PDE problems with parametric or uncertain inputs. In this paper, we consider a linear elliptic PDE with a generic parametric coefficient satisfying minimal assumptions that guarantee well-posedness of the weak formulation in standard Lebesgue–Bochner spaces. Building on earlier works for PDEs with affine-parametric representation of input data, we have performed a posteriori error analysis of Galerkin approximations and designed an adaptive solution algorithm for the considered problem. An important contribution of this work is that it opens the possibility of solving elliptic PDE problems with *non-affine* parametric representations of input data using Galerkin approximations with rigorous error control,

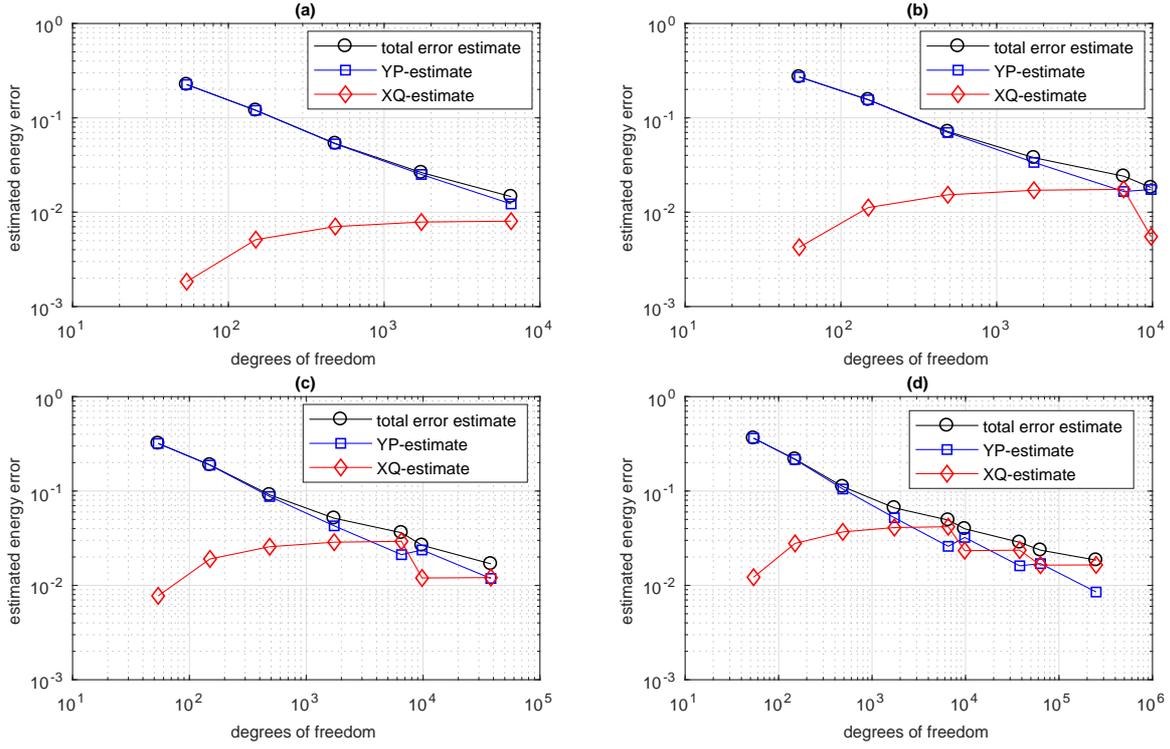


Figure 3: Energy error estimates at each step of the adaptive algorithm for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = a^2(\mathbf{x}, \mathbf{y})$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.2 with $\bar{\sigma} = 2$: (a) $\bar{\alpha} = 0.4$; (b) $\bar{\alpha} = 0.6$; (c) $\bar{\alpha} = 0.8$; (d) $\bar{\alpha} = 1$.

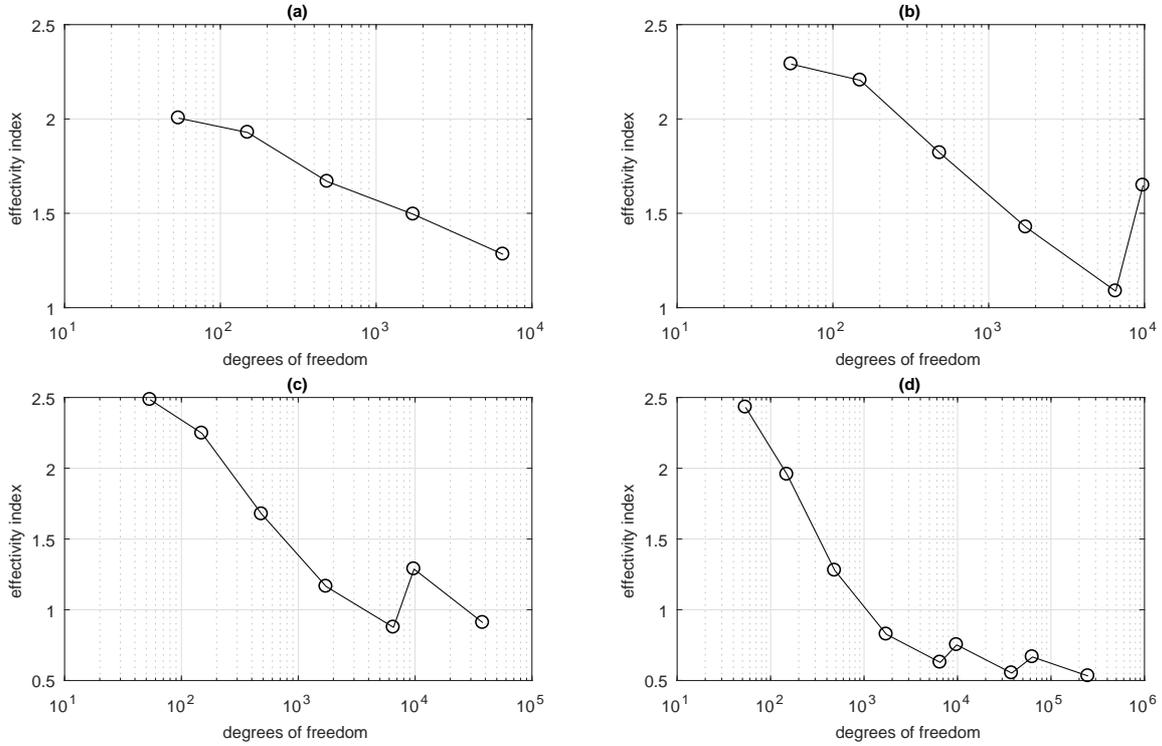


Figure 4: The effectivity indices for the Galerkin solutions at each iteration of the adaptive algorithm for the model problem (2.1) with $T(\mathbf{x}, \mathbf{y}) = a^2(\mathbf{x}, \mathbf{y})$ and the decomposition of $a(\mathbf{x}, \mathbf{y})$ as in Example 5.2 with $\bar{\sigma} = 2$: (a) $\bar{\alpha} = 0.4$; (b) $\bar{\alpha} = 0.6$; (c) $\bar{\alpha} = 0.8$; (d) $\bar{\alpha} = 1$.

thus, providing an effective alternative to traditional sampling techniques for such problems. Furthermore, our proof of concept implementation and extensive numerical tests demonstrate the effectiveness of our error estimation strategy and the practicality of the developed adaptive algorithm for this class of parametric PDE problems.

Appendix A gPC expansion coefficients for parametric exponential and quadratic functions

In this paper, we work with two forms of the diffusion coefficient $T(\mathbf{x}, \mathbf{y})$: $T(\mathbf{x}, \mathbf{y}) = \exp(a(\mathbf{x}, \mathbf{y}))$ and $T(\mathbf{x}, \mathbf{y}) = a^2(\mathbf{x}, \mathbf{y})$, where $a(\mathbf{x}, \mathbf{y})$ is given by (5.1). For $T = \exp(a)$, we are able to separate the variables y_m . Specifically, for $M > 1$, the integral in (4.2) can be expressed as a product of 1D integrals as follows:

$$t_\gamma(\mathbf{x}) = \exp(a_0(\mathbf{x})) \prod_{m=1}^M \int_{\Gamma_m} \exp(a_m(\mathbf{x})y_m) p_{\gamma_m}^m(y_m) q_m(y_m) dy_m. \quad (\text{A.1})$$

The 1D integrals with respect to y_m in (A.1) can be approximated numerically by using Gaussian quadrature.

For $T = a^2$, the infinite sum in (4.1) is naturally truncated to a finite sum of terms indexed by $\gamma \in \mathcal{P}_{M,2}$. Indeed, we have

$$T = a_0^2 + 2a_0 \sum_{m=1}^M a_m y_m + \sum_{m=1}^M a_m^2 y_m^2 + 2 \sum_{m=2}^M \sum_{n=1}^{m-1} a_m a_n y_m y_n$$

and

$$\begin{aligned} t_\gamma(\mathbf{x}) &= a_0^2 \prod_{i=1}^M \int_{\Gamma_i} p_{\gamma_i}^i q_i dy_i + 2a_0 \sum_{m=1}^M a_m \left(\prod_{\substack{i=1 \\ i \neq m}}^M \int_{\Gamma_i} p_{\gamma_i}^i q_i dy_i \right) \int_{\Gamma_m} y_m p_{\gamma_m}^m q_m dy_m \\ &\quad + \sum_{m=1}^M a_m^2 \left(\prod_{\substack{i=1 \\ i \neq m}}^M \int_{\Gamma_i} p_{\gamma_i}^i q_i dy_i \right) \int_{\Gamma_m} y_m^2 p_{\gamma_m}^m q_m dy_m \\ &\quad + 2 \sum_{m=2}^M \sum_{n=1}^{m-1} a_m a_n \left(\prod_{\substack{i=1 \\ i \neq m, n}}^M \int_{\Gamma_i} p_{\gamma_i}^i q_i dy_i \right) \int_{\Gamma_m} y_m p_{\gamma_m}^m q_m dy_m \int_{\Gamma_n} y_n p_{\gamma_n}^n q_n dy_n. \end{aligned}$$

The orthogonality of $\{p_n^m\}_{n \in \mathbb{N}_0}$ gives the following conditions:

$$\begin{aligned} \int_{\Gamma_i} p_{\gamma_i}^i q_i dy_i &= \begin{cases} 1 & \text{for } \gamma_i = 0, \\ 0 & \text{for } \gamma_i \neq 0, \end{cases} \\ \int_{\Gamma_i} y_i p_{\gamma_i}^i q_i dy_i &= 0 \text{ for } \gamma_i > 1, \quad \int_{\Gamma_i} y_i^2 p_{\gamma_i}^i q_i dy_i = 0 \text{ for } \gamma_i > 2; \end{aligned}$$

this implies that $t_\gamma = 0$ for $\gamma \notin \mathcal{P}_{M,2}$. If $M = 1$, then there are only three non-zero terms in (4.1) that are indexed by $\gamma \in \mathcal{P}_{1,2} \subset \mathbb{N}_0$ with the spatial expansion coefficients as follows:

$$t_\gamma(\mathbf{x}) = \begin{cases} a_0^2 + 2a_0a_1 \int_{\Gamma_1} y_1 q_1(y_1) dy_1 + a_1^2 \int_{\Gamma_1} y_1^2 q_1(y_1) dy_1 & \text{for } \gamma = (0), \\ 2a_0a_1 \int_{\Gamma_1} y_1 p_1^1(y_1) q_1(y_1) dy_1 + a_1^2 \int_{\Gamma_1} y_1^2 p_1^1(y_1) q_1(y_1) dy_1 & \text{for } \gamma = (1), \\ a_1^2 \int_{\Gamma_1} y_1^2 p_2^1(y_1) q_1(y_1) dy_1 & \text{for } \gamma = (2). \end{cases}$$

If $M > 1$, then there are four types of non-zero items in (4.1) that are indexed by $\gamma \in \mathcal{P}_{M,2}$ with the following spatial expansion coefficients:

(i) if $\gamma = \mathbf{0}$, then

$$t_\gamma(\mathbf{x}) = a_0^2 + 2a_0 \sum_{m=1}^M a_m \int_{\Gamma_m} y_m q_m dy_m + \sum_{m=1}^M a_m^2 \int_{\Gamma_m} y_m^2 q_m dy_m \\ + 2 \sum_{m=2}^M \sum_{n=1}^{m-1} a_m a_n \int_{\Gamma_m} y_m q_m dy_m \int_{\Gamma_n} y_n q_n dy_n;$$

(ii) if γ has only one non-zero element, $\gamma_j = 1$, $1 \leq j \leq M$, then

$$t_\gamma(\mathbf{x}) = 2a_0 a_j \int_{\Gamma_j} y_j p_1^j q_j dy_j + a_j^2 \int_{\Gamma_j} y_j^2 p_1^j q_j dy_j \\ + 2 \sum_{\substack{n=1 \\ n \neq j}}^M a_j a_n \int_{\Gamma_j} y_j p_1^j q_j dy_j \int_{\Gamma_n} y_n q_n dy_n;$$

(iii) if γ has only one non-zero element, $\gamma_j = 2$, $1 \leq j \leq M$, then

$$t_\gamma(\mathbf{x}) = a_j^2 \int_{\Gamma_j} y_j^2 p_2^j q_j dy_j;$$

(iv) if γ has only two non-zero elements, $\gamma_i = \gamma_j = 1$, $1 \leq i < j \leq M$, then

$$t_\gamma(\mathbf{x}) = 2a_i a_j \int_{\Gamma_i} y_i p_1^i q_i dy_i \int_{\Gamma_j} y_j p_1^j q_j dy_j.$$

Thus, the index sets $\mathcal{N}(\mathcal{P}, \mathcal{P})$ in (4.7), (4.11), (4.17), $\mathcal{N}(\mathcal{P}, \mathcal{Q})$ in (4.12), (4.18), and $\mathcal{N}(\mathcal{Q}, \mathcal{Q})$ in (4.18) are replaced by $\mathcal{N}(\mathcal{P}, \mathcal{P}) \cap \mathcal{P}_{M,2}$, $\mathcal{N}(\mathcal{P}, \mathcal{Q}) \cap \mathcal{P}_{M,2}$, and $\mathcal{N}(\mathcal{Q}, \mathcal{Q}) \cap \mathcal{P}_{M,2}$, respectively.

References

- [1] M. AINSWORTH AND J. T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, 2000.

- [2] A. BESPALOV, C. E. POWELL, AND D. SILVESTER, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM Journal on Scientific Computing, 36 (2014), pp. A339–A363.
- [3] A. BESPALOV, D. PRAETORIUS, L. ROCCHI, AND M. RUGGERI, *Convergence of adaptive stochastic Galerkin FEM*, SIAM Journal on Numerical Analysis, 57 (2019), pp. 2359–2382.
- [4] A. BESPALOV AND L. ROCCHI, *Efficient adaptive algorithms for elliptic PDEs with random data*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 243–272.
- [5] A. BESPALOV AND D. SILVESTER, *Efficient adaptive stochastic Galerkin methods for parametric operator equations*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2118–A2140.
- [6] W. DÖRFLER, *A convergent adaptive algorithm for Poisson’s equation*, SIAM Journal on Numerical Analysis, 33 (1996), pp. 1106–1124.
- [7] M. EIGEL, C. GITTELSON, C. SCHWAB, AND E. ZANDER, *A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes*, ESAIM Mathematical Modelling and Numerical Analysis, 49 (2015), pp. 1367–1398.
- [8] M. EIGEL, C. J. GITTELSON, C. SCHWAB, AND E. ZANDER, *Adaptive stochastic Galerkin FEM*, Computer Methods in Applied Mechanics and Engineering, 270 (2014), pp. 247–269.
- [9] M. EIGEL, M. MARSCHALL, M. PFEFFER, AND R. SCHNEIDER, *Adaptive stochastic Galerkin FEM with for lognormal coefficients in hierarchical tensor representations*, Preprint 2515, WIAS, 2018. <http://dx.doi.org/10.20347/WIAS.PREPRINT.2515>.
- [10] M. EIGEL AND C. MERDON, *Local equilibration error estimators for guaranteed error control in adaptive stochastic higher-order Galerkin finite element methods*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 1372–1397.
- [11] M. EIGEL, M. PFEFFER, AND R. SCHNEIDER, *Adaptive stochastic Galerkin FEM with hierarchical tensor representations*, Numerische Mathematik, 136 (2017), pp. 765–803.
- [12] V. EIJKHOUT AND P. VASSILEVSKI, *The role of the strengthened Cauchy-Buniakowski-Schwarz inequality in multilevel methods*, SIAM Rev., 33 (1991), pp. 405–419.

- [13] O. G. ERNST, A. MUGLER, H.-J. STARKLOFF, AND E. ULLMANN, *On the convergence of generalized polynomial chaos expansions*, ESAIM: Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 317–339.
- [14] W. GAUTSCHI, *Orthogonal Polynomials: Computation and Approximation*, Oxford University Press, Oxford, 2004.
- [15] R. G. GHANEM AND P. D. SPANOS, *Stochastic Finite Elements: a Spectral Approach*, Springer-Verlag, New York, 1991.
- [16] C. J. GITTELSON, *An adaptive stochastic Galerkin method for random elliptic operators*, Mathematics of Computation, 82 (2013), pp. 1515–1541.
- [17] M. F. PELLISSETTI AND R. G. GHANEM, *Iterative solution of systems of linear equations arising in the context of stochastic finite elements*, Advances in Engineering Software, 31 (2000), pp. 607–616.
- [18] C. E. POWELL AND H. C. ELMAN, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA Journal of Numerical Analysis, 29 (2009), pp. 350–375.
- [19] C. SCHWAB AND C. J. GITTELSON, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, Acta Numerica, 20 (2011), pp. 291–467.
- [20] D. J. SILVESTER, A. BESPALOV, AND C. E. POWELL, *Stochastic IFISS (S-IFISS), version 1.04*, June 2017. Available online at <http://www.manchester.ac.uk/ifiss/sifiss.html>.
- [21] E. ULLMANN, *A Kronecker product preconditioner for stochastic Galerkin finite element discretizations*, SIAM Journal on Scientific Computing, 32 (2010), pp. 923–946.
- [22] D. XIU AND G. E. KARNIADAKIS, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM Journal on Scientific Computing, 24 (2002), pp. 619–644.