J Sci Comput (2018) 77:1030–1054
https://doi.org/10.1007/s10915-018-0736-4

CBS Constants & Their Role in Error Estimation for Stochastic Galerkin Finite Element Methods

Adam J. Crowder[1] · **Catherine E. Powell**[1]

Received: 6 March 2018 / Revised: 30 April 2018 / Accepted: 8 May 2018 / Published online: 19 May 2018

Abstract Stochastic Galerkin finite element methods (SGFEMs) are commonly used to approximate solutions to PDEs with random inputs. However, the study of a posteriori error estimation strategies to drive adaptive enrichment of the associated tensor product spaces is still under development. In this work, we revisit an a posteriori error estimator introduced in Bespalov and Silvester (SIAM J Sci Comput 38(4):A2118–A2140, 2016) for SGFEM approximations of the parametric reformulation of the stochastic diffusion problem. A key issue is that the bound relating the true error to the estimated error involves a CBS (Cauchy–Buniakowskii–Schwarz) constant. If the approximation spaces associated with the parameter domain are orthogonal in a weighted L^2 sense, then this CBS constant only depends on a pair of finite element spaces H_1, H_2 associated with the spatial domain and their compatibility with respect to an inner product associated with a parameter-free problem. For fixed choices of H_1, we investigate non-standard choices of H_2 and the associated CBS constants, with the aim of designing efficient error estimators with effectivity indices close to one. When H_1 and H_2 satisfy certain conditions, we also prove new theoretical estimates for the CBS constant using linear algebra arguments.

Keywords Error estimation for stochastic PDE problems · Stochastic finite element method · Stochastic Galerkin method · Strengthened Cauchy–Schwarz inequality · CBS constants

Mathematics Subject Classification 35R60 · 60H35 · 65N30 · 65F10

✉ Adam J. Crowder
adam.crowder@manchester.ac.uk

Catherine E. Powell
c.powell@manchester.ac.uk

[1] School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, UK

1 Introduction

The motivation for this work is the design of efficient a posteriori error estimators for adaptive Galerkin finite element approximation of solutions to partial differential equations (PDEs). In particular, we are interested in PDEs with random inputs and so-called stochastic Galerkin finite element methods (SGFEMs) (see [4,5,14,22,24,31]). When the inputs are represented by a finite, or countably infinite number of random variables $\xi_m : \Omega \to \mathbb{R}$, $m = 1, 2, \ldots$, where Ω is a sample space, it is conventional to reformulate the stochastic problem of interest as a high-dimensional deterministic one, whose solution depends on a set of *parameters* $y_m = \xi_m(\omega)$. Unlike sampling methods such as Monte Carlo, in the SGFEM approach, an approximation is sought in a tensor product space of the form $H_1 \otimes P$ where H_1 is an appropriate finite element space associated with the spatial domain and P is a set of polynomials associated with the parameter domain. When the number of active parameters is high, the dimension of $H_1 \otimes P$ for standard choices of H_1 and P can become unwieldy. To remedy this, we can either work with standard approximation spaces and deal with the resulting very large discrete systems by using smart linear algebra techniques (see [20,21, 25,26,32,33]), or we can use an adaptive approach, starting in a low-dimensional space $H_1^0 \otimes P^0$, and using a posteriori error estimators to decide whether it is necessary to enrich H_1^0 or P^0, or both. This allows us to build up a tailored sequence of approximation spaces $H_1^\ell \otimes P^\ell$, $\ell = 0, 1, \ldots$ incrementally, so that the dimension of the final space is balanced against an error tolerance for a quantity of interest.

We consider the steady-state stochastic diffusion problem. Let $D \subset \mathbb{R}^{2,3}$ be a bounded spatial domain and let $\mathbf{y} = [y_1, y_2, \ldots]$ be a vector-valued parameter which takes values in $\Gamma = \prod_{m=1}^{\infty} \Gamma_m$ (the parameter domain). We want to approximate the function $u : D \times \Gamma \to \mathbb{R}$ that satisfies

$$-\nabla \cdot (a(\mathbf{x}, \mathbf{y}) \nabla u(\mathbf{x}, \mathbf{y})) = f(\mathbf{x}), \qquad \mathbf{x} \in D,\ \mathbf{y} \in \Gamma, \tag{1a}$$

$$u(\mathbf{x}, \mathbf{y}) = 0, \qquad \mathbf{x} \in \partial D,\ \mathbf{y} \in \Gamma. \tag{1b}$$

For simplicity, we assume that $f = f(\mathbf{x})$ is independent of $\mathbf{y}$, but the methodologies discussed herein can be extended to accommodate $f = f(\mathbf{x}, \mathbf{y})$. We also assume that the parameters are bounded, with $y_m \in \Gamma_m = [-1, 1]$. We denote by $\pi(\mathbf{y})$ a product measure on $(\Gamma, \mathcal{B}(\Gamma))$, where $\mathcal{B}(\Gamma)$ is the Borel σ-algebra on Γ, so that $\pi(\mathbf{y}) = \prod_{m=1}^{\infty} \pi_m(y_m)$, where $\pi_m(y_m)$ is a measure on $(\Gamma_m, \mathcal{B}(\Gamma_m))$. In addition,

$$\int_{\Gamma_m} y_m \, d\pi_m(y_m) = 0, \quad m \in \mathbb{N}, \tag{2}$$

which is true when y_m is the image of a mean zero random variable and $\pi_m(y_m)$ is the associated probability measure.

Assumption 1 The coefficient $a(\mathbf{x}, \mathbf{y})$ admits the decomposition

$$a(\mathbf{x}, \mathbf{y}) = a_0(\mathbf{x}) + \sum_{m=1}^{\infty} a_m(\mathbf{x}) y_m, \quad \mathbf{x} \in D,\ \mathbf{y} \in \Gamma, \tag{3}$$

with $a_0(\mathbf{x}), a_m(\mathbf{x}) \in L^\infty(D)$. Moreover, there exist real positive constants $a^0_{\min}$ and $a^0_{\max}$ such that

$$0 < a^0_{\min} \le a_0(\mathbf{x}) \le a^0_{\max} < \infty \quad \text{a.e. in } D, \tag{4}$$

and $||a_m||_{L^\infty(D)}$ converges sufficiently quickly to zero as $m \to \infty$ so that

$$\sum_{m=1}^{\infty} ||a_m||_{L^\infty(D)} < a_{\min}^0. \tag{5}$$

The parameter-free function $a_0(\mathbf{x})$ typically represents the mean with each term $a_m(\mathbf{x})y_m$ representing a perturbation away from the mean, while (5) helps ensure the well posedness of the weak formulation of (1a)–(1b).

Next, we recall the classical strengthened Cauchy–Buniakowskii–Schwarz (CBS) inequality (see [1, Theorem 5.4]), a key tool in many areas of numerical analysis.

Theorem 1 *Let H be a Hilbert space equipped with inner product $(\cdot,\cdot)$ and induced norm $||\cdot||$ and let U, V be a pair of finite-dimensional subspaces of H satisfying $U \cap V = \{0\}$. Then, there exists a constant $\gamma \in [0, 1)$, depending only on U and V such that*

$$|(u, v)| \leq \gamma \|u\| \|v\|, \qquad \forall\, u \in U, \quad \forall\, v \in V. \tag{6}$$

The smallest constant $\gamma \in [0, 1)$ satisfying (6) is

$$\gamma_{\min} := \sup_{u \in U,\, v \in V} \frac{|(u, v)|}{\|u\| \|v\|}, \tag{7}$$

and is known as the CBS constant. In particular, CBS constants appear in the analysis of hierarchical preconditioners (see [2,3,30]) and certain types of a posteriori error estimators for Galerkin finite element approximations to PDEs. See, for example, [1,19,27,28].

The design of error estimators for SGFEMs for parameter-dependent PDEs is still under development. However, there have been a few recent works (see [8–12,16–18,29]) for the model problem (1a)–(1b). In [12] and [11], algorithms constructing so-called *sparse* SGFEM approximations are driven by a priori error analysis, where the error associated with each discretisation parameter is balanced against the total number of degrees of freedom. In [16] and [17] a general framework for an explicit residual-based error estimation strategy for SGFEMs is proposed, where the selection of hierarchical approximation spaces is driven by a Dörfler marking strategy [15] on both the spatial and parameter domains. In both works, overestimation up to a factor of 10 of the true error is reported. In [18] a similar approach is taken, where residuals are computed using an equilibrated fluxes strategy and overestimation up to a factor of 5 is reported.

We focus on the (implicit) approach taken in [8–10] which is based on solving local subproblems for the error over a 'detail' space. The bound for the effectivity index of the resulting error estimators depends on a CBS constant. In [8–10] no insight into which choices of detail space result in a sharp error bound (effectivity indices close to one) is given. In this paper we provide that analysis. Specifically, we provide detailed information about the CBS constant and derive theoretical bounds for it, for certain choices of SGFEM solution and detail spaces. Due to the way that the spaces associated with the parameter domain are chosen, the CBS constants needed to analyse the estimators in [8–10] depend only on a pair of finite element spaces on the spatial domain. Hence, our results are also applicable to the design of adaptive finite element schemes for deterministic PDEs. We investigate which choices of detail FEM space result in CBS constants close to zero, to help ensure a sharp error bound. In particular, due to cost restrictions imposed to avoid high-dimensional detail spaces, we investigate non-standard choices that aren't typically considered in the deterministic setting. The error estimation strategy in [29] also relies on a CBS constant, but in a different setting. Enrichment of the finite element space is not considered.

1.1 Outline

In Sect. 2, for the benefit of readers who are not familiar with the area, we review classical results from [1,6] concerning a posteriori error estimation for Galerkin approximation. In Sect. 3 we demonstrate how those results are applied to SGFEM approximations of (1a)–(1b), leading to a simplified analysis of the error estimator introduced in [10] and the associated error bound. In particular, we show how the bound depends on a CBS constant associated with two finite element spaces H_1 and H_2. In Sect. 4 we first remind the reader how to compute CBS constants numerically. We then study some (non-standard) pairs of H_1 and H_2 and compute the associated CBS constants. In Sect. 5 we demonstrate that if H_1 is the space of piecewise bilinear ($\mathbb{Q}_1$) functions, theoretical estimates for the CBS constant can be obtained using a novel linear algebra approach for several choices of H_2. Finally, in Sect. 6 we present numerical results demonstrating the quality of the aforementioned error estimator, and the vital importance of choosing the right detail spaces.

2 Classical a Posteriori Error Estimation

The following classical results from [1,6], along with Theorem 1, form the foundations of our main investigation in Sects. 3, 4, 5, 6. Let V be a Hilbert space with norm $||\cdot||_V$ and let $B : V \times V \to \mathbb{R}$ and $F : V \to \mathbb{R}$ denote a bounded and coercive bilinear form and linear functional, respectively. Consider the problem:

$$\text{find } u \in V : \quad B(u, v) = F(v), \quad \text{for all } v \in V. \tag{8}$$

We assume that $B(\cdot, \cdot)$ is an inner product on V with the 'energy' norm

$$||\cdot||_B = B(\cdot, \cdot)^{1/2}$$

and that (8) is uniquely solvable. We now seek a Galerkin approximation to $u \in V$. Let X be an N_X–dimensional subspace of V. We then solve:

$$\text{find } u_X \in X : \quad B(u_X, v) = F(v), \quad \text{for all } v \in X. \tag{9}$$

Computing the error $e = u - u_X \in V$ is a non-trivial task. Our goal is to estimate the energy error $||e||_B$. Clearly, e satisfies the problem:

$$\text{find } e \in V : \quad B(e, v) = F(v) - B(u_X, v), \quad \text{for all } v \in V, \tag{10}$$

where $u_X \in X$ is the Galerkin approximation satisfying (9).

Now suppose we choose a second subspace $W \subset V$ of dimension N_W, where $X \subset W$ (i.e., W is richer than X) and consider the following problem:

$$\text{find } u_W \in W : \quad B(u_W, v) = F(v), \quad \text{for all } v \in W. \tag{11}$$

By letting $e_W = u_W - u_X$ we see that

$$B(e_W, v) = B(u_W, v) - B(u_X, v) = F(v) - B(u_X, v), \quad \text{for all } v \in W. \tag{12}$$

Note that (12) is simply a restatement of (10) over W. We deduce then that the function $e_W \in W$ satisfying (12) estimates the true error $e \in V$ satisfying (10). Whilst we do not compute u_W, it is clear that the quality of that Galerkin approximation (and hence the choice of W) determines the quality of the estimator e_W. To analyse this, we require the following assumption.

Assumption 2 Let the functions u, u_X and u_W satisfy (8), (9) and (11) respectively. There exists a constant $\beta \in [0, 1)$ (the saturation constant) such that

$$||u - u_W||_B \leq \beta ||u - u_X||_B. \tag{13}$$

In many applications, Assumption 2 is reasonable (see [1, p. 88]). The relationship between $||e||_B$ and $||e_W||_B$ is summarised in the next result.

Theorem 2 (See [1, p. 89]) *Let Assumption 2 hold and let $e \in V$ and $e_W \in W$ satisfy (10) and (12) respectively, then*

$$||e_W||_B \leq ||e||_B \leq \frac{1}{\sqrt{1-\beta^2}} ||e_W||_B, \tag{14}$$

where $\beta \in [0, 1)$ is the saturation constant satisfying (13).

The interpretation of (14) is as follows; $||e_W||_B$ will never overestimate the true error $||e||_B$, but could underestimate it by a factor of $(1 - \beta^2)^{-1/2}$.

Problem (12) leads to a linear system of N_W equations which may be too expensive to solve. This is the case for the problem considered in Sect. 3. Suppose then that $B_0 : V \times V \to \mathbb{R}$ is an inner product with induced norm $||\cdot||_{B_0} = B_0(\cdot, \cdot)^{1/2}$, whose matrix representation on W is more convenient to work with. We may then consider the alternative problem:

$$\text{find } e_0 \in W : \quad B_0(e_0, v) = F(v) - B(u_X, v), \quad \text{for all } v \in W. \tag{15}$$

The next result summarises the relationship between $||e_W||_B$ and $||e_0||_{B_0}$.

Theorem 3 (See [1, Theorem 5.3]) *Let $e_W \in W$ and $e_0 \in W$ satisfy (12) and (15) respectively and suppose that there exist $\lambda, \Lambda \in \mathbb{R}^+$ such that*

$$\lambda ||v||_B^2 \leq ||v||_{B_0}^2 \leq \Lambda ||v||_B^2, \quad \textit{for all } v \in V, \tag{16}$$

(the norms are equivalent) then

$$\sqrt{\lambda} ||e_0||_{B_0} \leq ||e_W||_B \leq \sqrt{\Lambda} ||e_0||_{B_0}. \tag{17}$$

Even after replacing $B(\cdot, \cdot)$ with $B_0(\cdot, \cdot)$, since $W \supset X$, it may be more expensive to compute $e_0 \in W$ satisfying (15) than $u_X \in X$ satisfying (9). To reduce the cost further, we insist that

$$W = X \oplus Y, \quad X \cap Y = \{0\}, \tag{18}$$

where the 'detail' space $Y \subset V$ has dimension N_Y and consider the lower-dimensional problem:

$$\text{find } e_Y \in Y : \quad B_0(e_Y, v) = F(v) - B(u_X, v), \quad \text{for all } v \in Y. \tag{19}$$

Does $||e_Y||_{B_0}$ provide a good estimate for $||e||_B$? To answer this, we require Theorem 1. Since $X \cap Y = \{0\}$, there exists a constant $\gamma \in [0, 1)$ such that

$$|B_0(u, v)| \leq \gamma ||u||_{B_0} ||v||_{B_0}, \quad \text{for all } u \in X, \quad \text{for all } v \in Y. \tag{20}$$

The estimates $||e_Y||_{B_0}$ and $||e_0||_{B_0}$ are then related by the following theorem.

Theorem 4 (See [1, Theorem 5.2]) *Let $e_0 \in W$ and $e_Y \in Y$ satisfy (15) and (19) respectively and suppose that (18) holds. Then*

$$||e_Y||_{B_0} \leq ||e_0||_{B_0} \leq \frac{1}{\sqrt{1-\gamma^2}} ||e_Y||_{B_0}, \tag{21}$$

where $\gamma \in [0, 1)$ satisfies (20).

If X and Y are orthogonal with respect to the inner product $B_0(\cdot,\cdot)$ then $\gamma = 0$ and $||e_Y||_{B_0} = ||e_0||_{B_0}$. Consolidating Theorems 2–4 yields the final result.

Theorem 5 *Let $e \in V$ and $e_Y \in Y$ satisfy (10) and (19) respectively, where (18) holds. If Assumption 2 holds, and there exist $\lambda, \Lambda \in \mathbb{R}^+$ such that (16) holds, then*

$$\sqrt{\lambda}\,||e_Y||_{B_0} \leq ||e||_B \leq \frac{\sqrt{\Lambda}}{\sqrt{1-\beta^2}\sqrt{1-\gamma^2}}\,||e_Y||_{B_0}, \tag{22}$$

where $\gamma \in [0,1)$ satisfies (20) and $\beta \in [0,1)$ satisfies (13).

In summary, the quality of the energy error estimate $||e_Y||_{B_0}$ is determined by two constants β and γ, which both depend on X and Y. Ideally, we want $\sqrt{1-\beta^2}\sqrt{1-\gamma^2} \approx 1$. Given a fixed initial approximation space X, what is the best choice of detail space Y, from the point of view of obtaining the best possible error estimate? This is the essence of our investigation.

3 The Parametric Diffusion Problem

The variational formulation of (1a)–(1b) is:

$$\text{find } u \in V = H_0^1(D) \otimes L_\pi^2(\Gamma): \quad B(u,v) = F(v), \quad \text{for all } v \in V, \tag{23}$$

where $H_0^1(D)$ is the usual Hilbert space and $L_\pi^2(\Gamma)$ is given by

$$L_\pi^2(\Gamma) = \left\{ v(\mathbf{y}) \;\middle|\; \langle v, v\rangle_{L_\pi^2(\Gamma)} := \int_\Gamma v(\mathbf{y})^2\, d\pi(\mathbf{y}) < \infty \right\}.$$

V is equipped with the norm $||\cdot||_V$, where $||v||_V^2 = \int_\Gamma ||v(\cdot,\mathbf{y})||_{H_0^1(D)}^2\, d\pi(\mathbf{y})$ and the bilinear form $B : V \times V \to \mathbb{R}$ and the linear functional $F : V \to \mathbb{R}$ are given by

$$B(u,v) = \int_\Gamma \int_D a(\mathbf{x},\mathbf{y}) \nabla u(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\pi(\mathbf{y}), \tag{24}$$

$$F(v) = \int_\Gamma \int_D f(\mathbf{x}) v(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\pi(\mathbf{y}). \tag{25}$$

To ensure that (23) is well-posed, we make the following assumption.

Assumption 3 There exist real positive constants $a_{\min}$ and $a_{\max}$ such that

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

If Assumption 3 holds, the bilinear form $B(\cdot,\cdot)$ defined in (24) induces a norm $||\cdot||_B = B(\cdot,\cdot)^{1/2}$ (the energy norm). Note that due to (3), we have the decomposition

$$B(u,v) = B_0(u,v) + \sum_{m=1}^{\infty} B_m(u,v), \tag{26}$$

for all $u, v \in V$, where the component bilinear forms are defined by

$$B_0(u,v) = \int_\Gamma \int_D a_0(\mathbf{x}) \nabla u(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\pi(\mathbf{y}), \tag{27a}$$

$$B_m(u,v) = \int_\Gamma \int_D a_m(\mathbf{x}) y_m \nabla u(\mathbf{x},\mathbf{y}) \cdot \nabla v(\mathbf{x},\mathbf{y})\, d\mathbf{x}\, d\pi(\mathbf{y}). \tag{27b}$$

If (4) in Assumption 1 holds, then $B_0(\cdot,\cdot)$ in (27a) also induces a norm $||\cdot||_{B_0} = B_0(\cdot,\cdot)^{1/2}$. Moreover, the following norm equivalence holds:

$$\lambda||v||_B^2 \leq ||v||_{B_0}^2 \leq \Lambda||v||_B^2, \quad \text{for all } v \in V, \tag{28}$$

where $0 < \lambda < 1 < \Lambda < \infty$ and

$$\lambda := \frac{a_{\min}^0}{a_{\max}}, \quad \Lambda := \frac{a_{\max}^0}{a_{\min}}. \tag{29}$$

3.1 SGFEM Approximation

We now seek a Galerkin approximation to $u \in V$ satisfying (23). As in Sect. 2, we denote by X an N_X–dimensional subspace of V. Here, we exploit the tensor product structure of V and choose $X := H_1 \otimes P$, where $H_1 \subset H_0^1(D)$ and $P \subset L_\pi^2(\Gamma)$. We then look for $u_X \in X$ satisfying (9).

We choose $H_1 = \text{span}\{\phi_i(\mathbf{x})\}_{i=1}^n$ to be a space of finite element functions associated with a mesh on D and $P = \text{span}\{\varphi_i(\mathbf{y})\}_{i=1}^s$ to be a space of global (multivariate) polynomials on Γ, so that $N_X = ns$. We choose the basis functions for P to be orthonormal with respect to $\langle\cdot,\cdot\rangle_{L_\pi^2(\Gamma)}$. To this end, we introduce the set of finitely supported multi-indices; $J := \{\mu = (\mu_1, \mu_2, \dots) \in \mathbb{N}_0^{\mathbb{N}};\ \#\,\text{supp}(\mu) < \infty\}$, where $\text{supp}(\mu) := \{m \in \mathbb{N};\ \mu_m \neq 0\}$. For a given multi-index $\mu \in J$ we then construct

$$\varphi_\mu(\mathbf{y}) = \prod_{m=1}^{\infty} \varphi_{\mu_m}(y_m), \tag{30}$$

where the families of univariate polynomials $\{\varphi_{\mu_m}(y_m),\ \mu_m = 0, 1, 2\dots\}$, for $m = 1, \dots, \infty$, are chosen to be orthonormal with respect to the inner product associated with $\pi_m(y_m)$. We also assume that $\varphi_0(y_m) = 1$ so that $\varphi_\mu(\mathbf{y}) = \prod_{\mu_m \neq 0} \varphi_{\mu_m}(y_m)$ for any $\mu \in J$. Clearly, choosing the subspace P is equivalent to choosing a set of multi-indices $J_P \subset J$ with cardinality $\text{card}(J_P) = s$.

To compute a Galerkin approximation $u_X \in X$ satisfying (9), it is essential that the sum in (26) has a finite number of nonzero terms. It is not necessary to truncate the diffusion coefficient a priori. We need only assume that P contains polynomials in which a *finite* number of parameters y_m are 'active'. If we assume that the first M parameters are active, then, provided (2) holds, $B_m(u_X, v) = 0$ for $u_X, v \in X$ for all $m > M$ (e.g., see [8]). In other words, the projection onto $X = H_1 \otimes P$ truncates the sum after M terms.

3.2 A Posteriori Error Estimation

Suppose we now choose a second SGFEM space $W \subset V = H_0^1(D) \otimes L_\pi^2(\Gamma)$ such that $W \supset X := H_1 \otimes P$ and solve (12) to obtain an estimator $e_W \in W$ for the error $e = u - u_X$. If Assumption 2 holds for the chosen spaces X and W, then (14) also holds, where $||\cdot||_B$ is the energy norm induced by the bilinear form defined in (24). In addition, due to the norm equivalence (28), the bound (17) also holds, where $e_0 \in W$ satisfies (15) and $||\cdot||_{B_0}$ is the norm induced by the bilinear form defined in (27a).

There are several possible ways to construct W. Following [10], we choose

$$\begin{aligned} &W := (H_1 \otimes P) \oplus ((H_2 \otimes P) \oplus (H_1 \otimes Q)) =: X \oplus Y, \\ &H_1 \cap H_2 = \{0\},\ \ P \cap Q = \{0\}, \end{aligned} \tag{31}$$

with $H_2 \subset H_0^1(D)$ and $Q \subset L^2_\pi(\Gamma)$. Consequently $Y_1 \cap Y_2 = \{0\}$ for the spaces $Y_1 := H_2 \otimes P$ and $Y_2 := H_1 \otimes Q$. Let J_Q denote the set of finitely supported multi-indices which correspond to the subspace Q. If $J_P \cap J_Q = \emptyset$, then we have $P \cap Q = \{0\}$ as required. In this case, P and Q are mutually orthogonal with respect to $\langle \cdot, \cdot \rangle_{L^2_\pi(\Gamma)}$ since

$$\begin{aligned} \langle \varphi_\mu(\mathbf{y}), \varphi_\nu(\mathbf{y}) \rangle_{L^2_\pi(\Gamma)} &= \prod_{m=1}^{\infty} \int_{\Gamma_m} \varphi_{\mu_m}(y_m) \varphi_{\nu_m}(y_m)\, d\pi_m(y_m) \\ &= \prod_{m=1}^{\infty} \delta_{\mu_m \nu_m} = \delta_{\mu\nu} = 0, \end{aligned} \tag{32}$$

for all $\mu \in J_P$ and $\nu \in J_Q$. Furthermore, due to the tensor product structure of Y_1 and Y_2 and the fact that $P \cap Q = \{0\}$, it can be shown that

$$B_0(u, v) = 0 \quad \text{for all } u \in Y_1,\ v \in Y_2. \tag{33}$$

To see this, expand $u \in Y_1$ and $v \in Y_2$ in the chosen bases and use (32). Given Y, we can then compute the error estimate $\eta := ||e_Y||_{B_0}$ by solving (19) and the bound (21) holds. Combining all these results yields the result of Theorem 5. For completeness, we restate this for our parametric diffusion problem.

Theorem 6 *Let $u \in V = H_0^1(D) \otimes L^2_\pi(\Gamma)$ satisfy the variational problem (8) associated with the parametric diffusion problem (1a)–(1b) and let $u_X \in X := H_1 \otimes P$ satisfy (9). Choose H_2, Q and Y as in (31). Let $e_Y \in Y$ satisfy (19). If Assumptions 1–3 hold, then $\eta := ||e_Y||_{B_0}$ satisfies*

$$\sqrt{\lambda}\, \eta \leq ||u - u_X||_B \leq \frac{\sqrt{\Lambda}}{\sqrt{1-\gamma^2}\sqrt{1-\beta^2}}\, \eta, \tag{34}$$

where λ and Λ are defined in (29), $\gamma \in [0, 1)$ satisfies (20), and $\beta \in [0, 1)$ satisfies (13).

When Y is chosen as in (31), problem (19) decouples. Since $Y_1 \cap Y_2 = \{0\}$, $e_Y = e_{Y_1} + e_{Y_2}$ for some $e_{Y_1} \in Y_1, e_{Y_2} \in Y_2$ and thus $B_0(e_Y, v) = B_0(e_{Y_1} + e_{Y_2}, v) = B_0(e_{Y_1}, v) + B_0(e_{Y_2}, v)$ for all $v \in Y$. By choosing test functions $v \in Y_1$ and $v \in Y_2$ in (19) and considering the identity (33), we find that $e_Y \in Y$ satisfying (19) can be determined by solving the lower-dimensional problems

$$\text{find } e_{Y_1} \in Y_1: \quad B_0(e_{Y_1}, v) = F(v) - B(u_X, v), \quad \text{for all } v \in Y_1, \tag{35}$$

$$\text{find } e_{Y_2} \in Y_2: \quad B_0(e_{Y_2}, v) = F(v) - B(u_X, v), \quad \text{for all } v \in Y_2. \tag{36}$$

Moreover, since $B_0(e_{Y_1}, e_{Y_2}) = 0$, we have

$$\eta = ||e_Y||_{B_0} = \left(||e_{Y_1}||^2_{B_0} + ||e_{Y_2}||^2_{B_0} \right)^{\frac{1}{2}}. \tag{37}$$

This is precisely the estimator considered in [10]. In [10] however, (31) is rearranged as $W = ((H_1 \oplus H_2) \otimes P) \oplus (H_2 \otimes Q)$. The analysis in that work relies on the orthogonality of P and Q, and the decoupling of (15) into two smaller problems over $((H_1 \oplus H_2) \otimes P)$ and $(H_2 \otimes Q)$. A CBS constant is introduced into the analysis by splitting the former into $H_1 \otimes P$ and $H_2 \otimes P$. Our approach is subtly different. We introduce a CBS constant by splitting the augmented space W into X and Y, as would be done for the analogous deterministic problem (for which $X = H_1$ and $Y = H_2$).

If (4) holds, then $H_0^1(D)$ is a Hilbert space with respect to the inner product

$$\langle a_0 u, v\rangle = \int_D a_0(\mathbf{x})\nabla u(\mathbf{x})\cdot\nabla v(\mathbf{x})\, d\mathbf{x}, \quad u, v \in H_0^1(D), \tag{38}$$

and since $H_1 \cap H_2 = \{0\}$, by Theorem 1, there exists a $\gamma \in [0, 1)$ such that

$$|\langle a_0 u, v\rangle| \leq \gamma \langle a_0 u, u\rangle^{1/2}\langle a_0 v, v\rangle^{1/2}, \quad \text{for all } u \in H_1, \quad \text{for all } v \in H_2. \tag{39}$$

In [8, Lemma 3.1], it is shown that the γ that features in (39) also satisfies

$$|B_0(u, v)| \leq \gamma \|u\|_{B_0}\|v\|_{B_0}, \quad \text{for all } u \in X, \quad \text{for all } v \in Y_1. \tag{40}$$

Now consider (20). For each $v \in Y$, we have $v = v_1 + v_2$, with $v_1 \in Y_1$ and $v_2 \in Y_2$. Since P and Q are orthogonal with respect to $\langle\cdot,\cdot\rangle_{L^2_\pi}$, Y_1 and Y_2 are orthogonal with respect to $B_0(\cdot,\cdot)$ and so $B_0(u, v) = B_0(u, v_1)$ for all $u \in X$ and $v \in Y$. Hence, using $||v_1||^2_{B_0} = ||v||^2_{B_0} - ||v_2||^2_{B_0}$, we have $|B_0(u, v)| \leq \gamma||u||_{B_0}||v_1||_{B_0} \leq \gamma||u||_{B_0}||v||_{B_0}$, for all $u \in X$, and $v \in Y$, where $\gamma \in [0, 1)$ is the *same* constant satisfying (39). Consequently, γ in (34) can be determined by analyzing the spaces H_1 and H_2. P and Q do not play a role. They do, of course, affect the saturation constant β, and this will be discussed in Sect. 6.

3.3 Estimated Error Reductions

The constant γ also plays an important role in adaptivity. Given $u_X \in X$, how do we choose an enriched space $X^* \supset X$ in which to compute a new approximation u^* which yields a reduced energy error? Consider the problems;

$$\text{find } u_{W_1} \in W_1: \quad B(u_{W_1}, v) = F(v), \quad \text{for all } v \in W_1, \tag{41}$$

$$\text{find } u_{W_2} \in W_2: \quad B(u_{W_2}, v) = F(v), \quad \text{for all } v \in W_2, \tag{42}$$

where $W_1 := (H_1 \oplus H_2)\otimes P$ and $W_2 := H_1 \otimes (P \oplus Q)$. Let $e_{W_1} = u - u_{W_1}$ denote the error corresponding to the enhanced approximation $u_{W_1} \in W_1$. Due to Galerkin orthogonality we find;

$$||e_{W_1}||^2_B = ||e||^2_B - ||u_{W_1} - u_X||^2_B. \tag{43}$$

Hence, $||u_{W_1} - u_X||_B$ characterises the energy error reduction that would be achieved by enriching only the subspace $H_1 \subset H_0^1(D)$. Similarly, $||u_{W_2} - u_X||_B$ characterises the energy error reduction that would be achieved by enriching only the subspace $P \subset L^2_\pi(\Gamma)$. Fortunately, the two components $||e_{Y_1}||_{B_0}$ and $||e_{Y_2}||_{B_0}$ of our error estimator provide estimates of these error reductions, see [8, Theorem 5.1].

Theorem 7 *Let $u_X \in X = H_1 \otimes P$ be the Galerkin approximation satisfying (9) and let $u_{W_1} \in W_1$ and $u_{W_2} \in W_2$ satisfy (41) and (42). Then,*

$$\sqrt{\lambda}||e_{Y_1}||_{B_0} \leq ||u_{W_1} - u_X||_B \leq \frac{\sqrt{\Lambda}}{\sqrt{1-\gamma^2}}||e_{Y_1}||_{B_0}, \tag{44}$$

$$\sqrt{\lambda}||e_{Y_2}||_{B_0} \leq ||u_{W_2} - u_X||_B \leq \sqrt{\Lambda}||e_{Y_2}||_{B_0}, \tag{45}$$

where e_{Y_1} and e_{Y_2} satisfy (35) and (36), respectively, λ and Λ are the constants in (28), and $\gamma \in [0, 1)$ is the constant satisfying (39).

Given H_2 and Q, Theorem 7 allows us to assess whether enrichment of H_1 (with functions in H_2) is more beneficial than enrichment of P (with functions in Q). We may choose $X^* = W_1$ or $X^* = W_2$. Our choice is determined by which space offers the greatest estimated error reduction per additional degree of freedom. Note that the bound (44) is independent of the saturation constant β, and choosing H_2 in (31) so that the constant γ in (39) is small tightens the bound (44). That is, if the CBS constant is small, we can have more confidence in our decisions when performing adaptivity. We now study this constant for various choices of H_1 and H_2.

4 Numerical Estimates of CBS Constants

The constant $\gamma \in [0, 1)$ in (34) and (44), which is equivalent to the constant $\gamma \in [0, 1)$ satisfying (39), is not unique. Given H_1 and H_2, we want to find the smallest such constant, the CBS constant. We now recall a standard result from [19] which leads to a numerical method for computing the CBS constant associated with (39).

Suppose first that $M \in \mathbb{R}^{N\times N}$ is symmetric and positive definite with $N := m + n$ for $m, n \in \mathbb{N}$. Then $(\mathbb{R}^N, (\cdot,\cdot)_M)$ is a Hilbert space with respect to the inner product $(\mathbf{u}, \mathbf{v})_M := \mathbf{u}^\top M\mathbf{v}$. Now consider $U, V \subset \mathbb{R}^N$ given by

$$U := \left\{ \begin{pmatrix} \mathbf{u}_1 \\ 0 \end{pmatrix},\ \mathbf{u}_1 \in \mathbb{R}^m \right\}, \qquad V := \left\{ \begin{pmatrix} 0 \\ \mathbf{v}_2 \end{pmatrix},\ \mathbf{v}_2 \in \mathbb{R}^n \right\}, \tag{46}$$

which satisfies $U \cap V = \{\mathbf{0}\}$. When M has a particular block structure, Theorem 1 along with U and V in (46), leads to the following result (see [19] for a proof).

Corollary 1 *Let $M \in \mathbb{R}^{N\times N}$ be symmetric and positive definite with*

$$M = \begin{bmatrix} B & C^\top \\ C & A \end{bmatrix}, \tag{47}$$

where $N = m + n$, $B \in \mathbb{R}^{m\times m}$ and $A \in \mathbb{R}^{n\times n}$. There exists a constant $\gamma \in [0, 1)$ such that

$$(\mathbf{u}_1^\top C^\top \mathbf{v}_2)^2 \le \gamma^2 \left(\mathbf{u}_1^\top B\mathbf{u}_1\right)\left(\mathbf{v}_2^\top A\mathbf{v}_2\right), \quad \forall \mathbf{u}_1 \in \mathbb{R}^m, \quad \forall \mathbf{v}_2 \in \mathbb{R}^n. \tag{48}$$

Furthermore, the smallest such constant, $\gamma_{\min} \in [0, 1)$ (the CBS constant), satisfying (48) is the square root of the largest eigenvalue $\theta_{\max}$ of the generalised eigenvalue problem

$$CB^{-1}C^\top \mathbf{v}_2 = \theta A\mathbf{v}_2. \tag{49}$$

We now demonstrate that the CBS constant associated with (39) for various choices of H_1 and H_2 can be computed by solving an eigenvalue problem of the form (49). Recall that $H_1, H_2 \subset H_0^1(D)$ with $H_1 \cap H_2 = \{0\}$. For now, we assume $a_0(\mathbf{x}) = 1$. Note that due to the symmetry of (38) we may compute the CBS constant associated with the equivalent result; there exists a $\gamma \in [0, 1)$ such that

$$|\langle u, v\rangle| \le \gamma \langle u, u\rangle^{1/2}\langle v, v\rangle^{1/2} \quad \text{for all } u \in H_2, \quad \text{for all } v \in H_1. \tag{50}$$

Given $H_1 := \text{span}\{\phi_i(\mathbf{x})\}_{i=1}^n$ and $H_2 := \text{span}\{\psi_i(\mathbf{x})\}_{i=1}^m$, we can define the augmented subspace

$$H := H_2 \oplus H_1 \subset H_0^1(D), \tag{51}$$

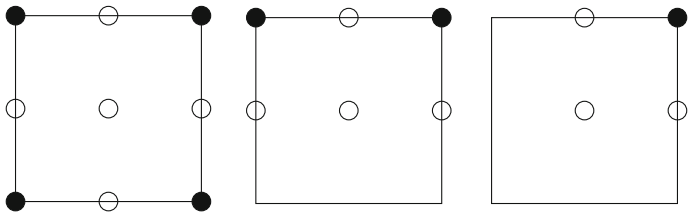

Fig. 1 Internal (**a**), edge (**b**), and corner (**c**) elements for Example 1. The black and clear markers are the nodes at which the basis functions of H_1 and H_2 are defined, respectively

of dimension $N = m + n$. We have $H = \text{span}\{\Phi_i(\mathbf{x})\}_{i=1}^N$, where $\Phi_i = \psi_i$, for $i = 1, 2, \ldots, m$, and $\Phi_{m+i} = \phi_i$, for $i = 1, 2, \ldots, n$. Then, for all $u \in H_2$ and $v \in H_1$, $\langle u, v\rangle = \mathbf{u}^\top M\mathbf{v}$ for some $\mathbf{u} \in U$ and $\mathbf{v} \in V$ in (46).

The matrix $M \in \mathbb{R}^{N\times N}$ is symmetric and positive definite and has the structure (47) with $[A]_{ij} = \langle\phi_i, \phi_j\rangle$, for $i, j = 1, \ldots, n$, $[B]_{ij} = \langle\psi_i, \psi_j\rangle$ for $i, j = 1, \ldots, m$ and $[C]_{ij} = \langle\phi_i, \psi_j\rangle$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. By Corollary 1, there exists a constant $\gamma \in [0, 1)$ such that (48) holds, which is equivalent to (50). Therefore, the CBS constant $\gamma_{\min}$ satisfying (50) can be computed numerically by solving the eigenvalue problem (49).

Given a fixed space H_1 associated with a uniform mesh $\mathcal{T}_h$ on the spatial domain D, we construct H_2 element-wise. That is, we insist that H_2 admits the decomposition

$$H_2 = \bigoplus_{\square_k\in\mathcal{T}_h} H_{k,2}, \qquad H_{k,2} = \text{span}\left\{\psi_i^k(\mathbf{x})\right\}_{i=1}^{m_k}, \tag{52}$$

where $\square_k$ denotes an element in $\mathcal{T}_h$. We choose the functions ψ_i^k to be bubble functions so that $H_{k,2}$ contains functions which only have non-zero support on $\square_k$. The linear system associated with the estimator e_{Y_1} satisfying (35) then decouples. Since $e_{Y_1} \in Y_1 := H_2 \otimes P$, on each $\square_k$, we have to solve a problem of size $m_k \times \dim(P)$. Since $\dim(P)$ may be large, to keep costs reasonable, we must restrict the dimension of $H_{k,2}$. Note that the necessity to restrict the dimension of H_2 is not as prevalent in the deterministic PDE setting, where each local problem for the analogous error estimate is of dimension m_k, not $m_k \times \dim(P)$. The goal is to find the best space (the one leading to the tightest error bound), of a fixed small dimension. Below, we fix H_1 and consider various choices of H_2 of the same dimension. We vary the mesh size h, and estimate the CBS constant by solving the eigenvalue problem (49).

Example 1 Let $D = [-1, 1]^2$ and let $\mathcal{T}_h$ denote a uniform mesh of square elements, with edge length h. Now let H_1 be the space of continuous functions that are piecewise bilinear on $\mathcal{T}_h$ (denoted $H_1 = \mathbb{Q}_1(h)$). On each $\square_k$ we construct a local space $H_{k,2}$ of dimension $m_k \leq 5$, by defining bubble functions at the edge midpoints and the element centroid (the $\mathbb{Q}_1$ nodes that would be introduced by a uniform mesh refinement). We consider the following options. The name given to the resulting space H_2 is shown in brackets.

1. *Biquadratic bubble functions* ($\mathbb{Q}_2(h)$) Consider the standard set of nine biquadratic ($\mathbb{Q}_2$) element basis functions and keep only those associated with the five selected nodes.
2. *Biquartic bubble functions* ($\mathbb{Q}_4(h)$) Consider the standard set of twenty-five biquartic ($\mathbb{Q}_4$) element basis functions and keep only the desired five.
3. *Piecewise bilinear bubble functions* ($\mathbb{Q}_1(h/2)$) Subdivide each element into four smaller ones of size $h/2$, and concatenate the standard $\mathbb{Q}_1$ basis functions associated with each new element at the five chosen nodes.

Table 1 Computed values of $\gamma^2_{\min}$ for Example 1, for varying h. H_1 is the usual $\mathbb{Q}_1$ finite element space and four choices of H_2 are considered

Mesh	h	N	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
4×4	2^{-1}	73	0.4106	0.0109	0.3381	0.0401
8×8	2^{-2}	337	0.4454	0.0119	0.3673	0.0437
16×16	2^{-3}	1441	0.4527	0.0121	0.3735	0.0445
32×32	2^{-4}	5953	0.4541	0.0121	0.3747	0.0446
64×64	2^{-5}	24193	0.4544	0.0121	0.3749	0.0446
Converged value			0.4545	0.0121	0.3750	0.0446

4. *Piecewise biquadratic bubble functions* ($\mathbb{Q}_2(h/2)$) This is the same as option 3 but with $\mathbb{Q}_2$ basis functions.

Figure 1 displays an arbitrary internal, edge, and corner element. In Table 1 we record $\gamma^2_{\min}$ for each choice of H_2 for varying h. In [8], the authors choose $H_2 = \mathbb{Q}_1(h/2)$ to define the error estimator $\eta = \|e_Y\|_{B_0}$ described in Sect. 3 when u_X is computed with $H_1 = \mathbb{Q}_1(h)$. Hence, when $a_0 = 1$, the associated CBS constant is $\gamma_{\min} \leq \sqrt{0.375} = \sqrt{3/8}$. However, of the four choices considered, $H_2 = \mathbb{Q}_4(h)$ yields the smallest CBS constant.

Example 2 Let $D = [-1, 1]^2$ and let $\mathcal{T}_h$ denote a uniform mesh of square elements. Now let H_1 be the set of continuous functions that are piecewise biquadratic on $\mathcal{T}_h$ (denoted $H_1 = \mathbb{Q}_2(h)$). On each $\square_k$ we construct a local space $H_{k,2}$ of dimension $m_k \leq 16$, by defining a set of bubble functions associated with the additional $\mathbb{Q}_2$ nodes that would be introduced by performing a uniform mesh refinement (see Fig. 2). We consider the following options.

1. *Biquartic bubble functions* ($\mathbb{Q}_4(h)$) Consider the set of twenty-five $\mathbb{Q}_4$ element basis functions associated with $\square_k$ but retain only those associated with the nodes indicated by the clear and grey markers in Fig. 2.
2. *Piecewise biquadratic bubble functions* ($\mathbb{Q}_2(h/2)$) Subdivide each element into four smaller ones of size $h/2$, and concatenate the standard $\mathbb{Q}_2$ basis functions associated with the new elements at the nodes indicated by the clear and grey markers in Fig. 2.

For our third and fourth choices of H_2 we modify the first two spaces by removing the basis functions associated with the nodes indicated by the grey markers in Fig. 2. This configuration is motivated by error estimation results presented in [23, p. 39]. We denote the resulting 'reduced' spaces by $\mathbb{Q}^r_4(h)$ and $\mathbb{Q}^r_2(h/2)$ and denote the number of degrees of freedom by N_r.

In Table 2 we record $\gamma^2_{\min}$ for each choice of H_2 for varying h. In [10], the authors choose $H_2 = \mathbb{Q}^r_4(h)$ to define the error estimator $\eta = \|e_Y\|_{B_0}$ described in Sect. 3, when u_X is computed with $H_1 = \mathbb{Q}_2(h)$. When $a_0 = 1$, the CBS constant is $\gamma_{\min} \leq \sqrt{0.36}$. Of the four spaces considered, $H_2 = \mathbb{Q}^r_4(h)$ yields the smallest CBS constant.

4.1 Local CBS Constants

Solving the eigenvalue problem (49) to compute the CBS constant when N is large is not practical. Alternatively, we may derive a small eigenvalue problem associated with a single

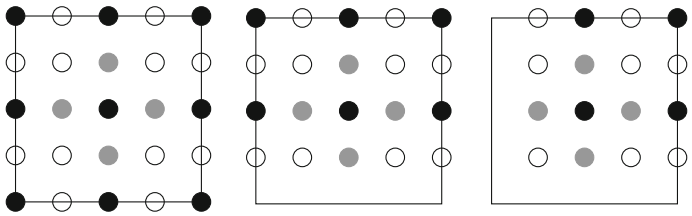

Fig. 2 Internal (**a**), edge (**b**), and corner (**c**) elements for Example 2. The black and clear/grey markers are the nodes at which the basis functions of H_1 and H_2 are defined, respectively

Table 2 Computed values of $\gamma_{\min}^2$ for Example 2, for varying h. H_1 is the usual $\mathbb{Q}_2$ finite element space and four choices of H_2 are considered

Mesh	h	N	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	N_r	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2×2	2^{-0}	57	0.3834	0.6764	41	0.3208	0.4904
4×4	2^{-1}	273	0.4341	0.6911	209	0.3565	0.5579
8×8	2^{-2}	1185	0.4391	0.6911	929	0.3595	0.5723
16×16	2^{-3}	4929	0.4399	0.6911	3905	0.3599	0.5758
32×32	2^{-4}	20097	0.4401	0.6911	16001	0.3600	0.5766
Converged value			0.4401	0.6911		0.3600	0.5769

element. For all $u \in H_2$ and $v \in H_1$ we have

$$\langle a_0 u, v\rangle = \sum_{\square_k \in \mathcal{T}_h} \int_{\square_k} a_0(\mathbf{x}) \nabla u|_k \cdot \nabla v|_k \, d\mathbf{x} =: \sum_{\square_k \in \mathcal{T}_h} \langle a_0 u_k, v_k\rangle_k, \tag{53}$$

where $u_k := u|_k \in H_{k,2} := H_2|_k$, and $v_k := v|_k \in H_{k,1} := H_1|_k$ with $H_{k,1}, H_{k,2} \subset H^1(\square_k)$ having dimensions $n_k = \dim(H_{k,1})$ and $m_k = \dim(H_{k,2})$ (recall that we now list functions in H_2 before functions in H_1). Here, $|_k$ denotes the restriction to element $\square_k$. For all $u_k \in H_{k,1}$ and $v_k \in H_{k,2}$, $\langle a_0 u_k, v_k\rangle_k = \mathbf{u}_k^\top M_k \mathbf{v}_k$, where the matrix $M_k \in \mathbb{R}^{N_k \times N_k}$ for $N_k := n_k + m_k$ has the same 2×2 block structure as before with $A_k \in \mathbb{R}^{n_k \times n_k}$, $B_k \in \mathbb{R}^{m_k \times m_k}$ and $C_k \in \mathbb{R}^{n_k \times m_k}$. Since $\langle a_0 \cdot, \cdot\rangle_k$ only induces a seminorm on $H^1(\square_k)$, M_k is positive semidefinite and Corollary 1 is not applicable. For our block matrices of interest, we require the following result from [19].

Corollary 2 *Let $M_k \in \mathbb{R}^{N_k \times N_k}$ be symmetric and positive semidefinite with*

$$M_k = \begin{bmatrix} B_k & C_k^\top \\ C_k & A_k \end{bmatrix}, \quad \mathcal{N}(M_k) = \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_2 \end{pmatrix}, \; A_k \mathbf{v}_2 = \mathbf{0}, \; C_k^T \mathbf{v}_2 = \mathbf{0}, \; \mathbf{v}_2 \in \mathbb{R}^{n_k} \right\}, \tag{54}$$

where $N_k = m_k + n_k$, $B_k \in \mathbb{R}^{m_k \times m_k}$ is invertible and $A_k \in \mathbb{R}^{n_k \times n_k}$. Let $U_k, V_k \subset \mathbb{R}^{N_k}$ have the same structure as U and V in (46), but for n_k and m_k in place of n and m. Then, there exists a constant $\gamma_k \in [0, 1)$ such that

$$(\mathbf{u}_1^\top C_k^\top \mathbf{v}_2)^2 \le \gamma_k^2 \left(\mathbf{u}_1^\top B_k \mathbf{u}_1\right)\left(\mathbf{v}_2^\top A_k \mathbf{v}_2\right), \quad \forall \mathbf{u}_1 \in \mathbb{R}^{m_k}, \quad \forall \mathbf{v}_2 \in \mathbb{R}^{n_k}. \tag{55}$$

If $H_{k,1}$ and $H_{k,2}$ are chosen so that M_k satisfies the conditions of Corollary 2, there exists a constant $\gamma_k \in [0, 1)$ such that (55) holds, or equivalently

$$\langle a_0 u_k, v_k\rangle_k^2 \leq \gamma_k^2 \langle a_0 u_k, u_k\rangle_k \langle a_0 v_k, v_k\rangle_k, \quad \forall u_k \in H_{k,2}, \quad \forall v_k \in H_{k,1}. \tag{56}$$

Furthermore, the local CBS constant $\gamma_{k,\min}$ associated with (56) is the square root of $\theta_{\max}$ satisfying

$$C_k B_k^{-1} C_k^\top \mathbf{v}_2 = \theta A_k \mathbf{v}_2, \qquad \mathbf{v}_2 \notin \mathcal{N}(A_k), \tag{57}$$

see [19]. It is straightforward to show that

$$\langle a_0 u, v\rangle \leq \sup_{\square_k} \gamma_{k,\min} \sum_{\square_k} \langle a_0 u_k, u_k\rangle_k^{1/2} \langle a_0 v_k, v_k\rangle_k^{1/2}$$

and comparing with (39) gives $\gamma_{\min} \leq \sup_{\square_k} \gamma_{k,\min}$. When the mesh is uniform and $a_0(\mathbf{x})$ is constant on each element, $\gamma_{k,\min}$ does not depend on h or $a_0|_k$. To estimate $\gamma_{\min}$, we only need to compute $\gamma_{k,\min}$ for a single internal element (as this is larger than the constant associated with corner/edge elements).

We now revisit Example 1 and compute local CBS constants associated with $\square_{\text{ref}} := [-1, 1]^2$. We choose $H_{k,1}$ to be the local $\mathbb{Q}_1$ finite element space whose basis functions are associated with the black markers shown in Fig. 3, and ordered as shown. Then, $\dim(H_{k,1}) = 4$ and, if $a_0 = 1$, we have

$$A_k = \begin{bmatrix} 2/3 & -1/6 & -1/3 & -1/6 \\ -1/6 & 2/3 & -1/6 & -1/3 \\ -1/3 & -1/6 & 2/3 & -1/6 \\ -1/6 & -1/3 & -1/6 & 2/3 \end{bmatrix}. \tag{58}$$

For the four choices of $H_{k,2}$ considered in Example 1 (which all have dimension five), we construct the matrix M_k in (54) and calculate $\gamma_{k,\min}^2$ by solving the eigenvalue problem (57). The ordering of the basis functions of $H_{k,2}$ is as illustrated by the clear markers in Fig. 3.

First, let $H_{k,2} = \mathbb{Q}_2(h)$. The matrices C_k and B_k are

$$C_k = \frac{1}{3} P, \quad B_k = \begin{bmatrix} 88/45 & -16/45 & 0 & -16/45 & -16/15 \\ -16/45 & 88/45 & -16/45 & 0 & -16/15 \\ 0 & -16/45 & 88/45 & -16/45 & -16/15 \\ -16/45 & 0 & -16/45 & 88/45 & -16/15 \\ -16/15 & -16/15 & -16/15 & -16/15 & 256/45 \end{bmatrix},$$

where we define

$$P := \begin{bmatrix} 1 & -1 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ -1 & 1 & 1 & -1 & 0 \\ -1 & -1 & 1 & 1 & 0 \end{bmatrix}. \tag{59}$$

Solving (57) gives $\gamma_{k,\min}^2 = \frac{5}{11} \approx 0.4545$, which agrees with the bound reported in [28]. If we choose $H_{k,2} = \mathbb{Q}_4(h)$ (a non-standard choice), we have

$$C_k = \frac{1}{15} P, \quad B_k = \begin{bmatrix} 373/127 & -39/197 & 1/2326 & -39/197 & 84/247 \\ -39/197 & 373/127 & -39/197 & 1/2326 & 84/247 \\ 1/2326 & -39/197 & 373/127 & -39/197 & 84/247 \\ -39/197 & 1/2326 & -39/197 & 373/127 & 84/247 \\ 84/247 & 84/247 & 84/247 & 84/247 & 3166/203 \end{bmatrix},$$

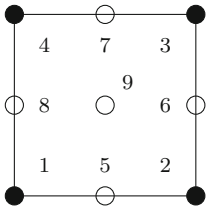

Fig. 3 An arbitrary internal $\mathbb{Q}_1$ element $\square_k \in \mathcal{T}_h$. The numbering of the solid black and clear makers illustrates the chosen ordering of the basis functions of $H_{k,1}$ and $H_{k,2}$, respectively

and $\gamma^2_{k,\min} \approx 0.0121$. Next, if $H_{k,2} = \mathbb{Q}_1(h/2)$, we have

$$C_k = \frac{1}{4}P, \quad B_k = \begin{bmatrix} 4/3 & -1/3 & 0 & -1/3 & -1/3 \\ -1/3 & 4/3 & -1/3 & 0 & -1/3 \\ 0 & -1/3 & 4/3 & -1/3 & -1/3 \\ -1/3 & 0 & -1/3 & 4/3 & -1/3 \\ -1/3 & -1/3 & -1/3 & -1/3 & 8/3 \end{bmatrix},$$

and $\gamma^2_{k,\min} = \frac{3}{8} = 0.3750$, which agrees with [27]. Finally, let $H_{k,2} = \mathbb{Q}_2(h/2)$ (another non-standard choice). Then,

$$C_k = \frac{1}{12}P, \quad B_k = \begin{bmatrix} 56/45 & -1/45 & 0 & -1/45 & -1/15 \\ -1/45 & 56/45 & -1/45 & 0 & -1/15 \\ 0 & -1/45 & 56/45 & -1/45 & -1/15 \\ -1/45 & 0 & -1/45 & 56/45 & -1/15 \\ -1/15 & -1/15 & -1/15 & -1/15 & 112/45 \end{bmatrix},$$

and $\gamma^2_{k,\min} \approx 0.0446$. Comparing with the results in Table 1, we confirm the relationship $\gamma^2_{\min} \leq \gamma^2_{\min,k}$.

Remark 1 If $a_0(\mathbf{x})$ varies in space, we may assume that $a_0(\mathbf{x})$ can be approximated by a function $a_0^h(\mathbf{x})$ that is constant in each element in $\mathcal{T}_h$. Then, on each $\square_k$, we have a symmetric and positive semidefinite matrix

$$M_k = \alpha_k \begin{bmatrix} B_k & C_k^\top \\ C_k & A_k \end{bmatrix}, \qquad \alpha_k := a_0^h|_k,$$

where B_k is invertible and B_k, C_k, A_k do not depend on α_k. The local eigenvalue problem, equivalent to (57), is $(\alpha_k C_k)(\alpha_k B_k)^{-1}(\alpha_k C_k^\top)\mathbf{v}_2 = \theta(\alpha_k A_k)\mathbf{v}_2$ for $\mathbf{v}_2 \notin \mathcal{N}(A_k)$, and thus the local CBS constant $\gamma_{k,\min}$ satisfying (56) is independent of α_k and $a_0^h(\mathbf{x})$.

5 Theoretical Estimates of the CBS Constant

In this section we fix $H_{k,1}$ to be the local $\mathbb{Q}_1$ finite element space so that A_k is given by (58) and assume that the degrees of freedom are numbered as shown in Fig. 3. We also assume that $H_{k,2}$ is chosen so that $\dim(H_{k,2}) = 5$ and the resulting matrices B_k and C_k have a particular structure. Exploiting this structure, and using only linear algebra arguments, we show that the local CBS constant $\gamma_{k,\min}$ can be calculated analytically without assembling and solving (57). To simplify notation, we drop the subscript k.

Theorem 8 *Let $M \in \mathbb{R}^{9\times 9}$ be a symmetric and positive semidefinite matrix with the 2×2 block structure (47), where $B \in \mathbb{R}^{5\times 5}$ is symmetric and positive definite and A is the $\mathbb{Q}_1$ element stiffness matrix defined in (58). If the matrix $CB^{-1}C^\top \in \mathbb{R}^{4\times 4}$ is of the form*

$$CB^{-1}C^\top = \beta \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} =: \beta Q, \tag{60}$$

for some constant $\beta \in \mathbb{R}^+$, then there exists a constant $\gamma \in [0, 1)$ such that

$$(\mathbf{u}_1^\top C^\top \mathbf{v}_2)^2 \le \gamma^2 \mathbf{u}_1^\top B\mathbf{u}_1 \mathbf{v}_2^\top A\mathbf{v}_2, \quad \forall \mathbf{u}_1 \in \mathbb{R}^5, \quad \forall \mathbf{v}_2 \in \mathbb{R}^4. \tag{61}$$

Proof It is sufficient to show that $\mathcal{N}(M)$ is given by the definition in (54). The result then follows from Corollary 2. Let $\mathbf{x}^T = (\mathbf{u}_1^T, \mathbf{v}_2^T) \in \mathbb{R}^9$ for $\mathbf{u}_1 \in \mathbb{R}^5$ and $\mathbf{v}_2 \in \mathbb{R}^4$ be such that $M\mathbf{x} = \mathbf{0}$. Then

$$B\mathbf{u}_1 + C^T\mathbf{v}_2 = \mathbf{0}, \tag{62}$$

$$C\mathbf{u}_1 + A\mathbf{v}_2 = \mathbf{0}, \tag{63}$$

and $S\mathbf{v}_2 = \mathbf{0}$ for the Schur complement $S = A - CB^{-1}C^T = A - \beta Q$. Since

$$S = \begin{bmatrix} \frac{2}{3} - \beta & -\frac{1}{6} & \beta - \frac{1}{3} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{2}{3} - \beta & -\frac{1}{6} & \beta - \frac{1}{3} \\ \beta - \frac{1}{3} & -\frac{1}{6} & \frac{2}{3} - \beta & -\frac{1}{6} \\ -\frac{1}{6} & \beta - \frac{1}{3} & -\frac{1}{6} & \frac{2}{3} - \beta \end{bmatrix}$$

and A are circulant matrices with zero row sums, we have

$$\mathcal{N}(S) = \mathcal{N}(A) = \text{span}\{(1, 1, 1, 1)^T\} \tag{64}$$

and thus $\mathbf{v}_2 \in \mathcal{N}(A)$. We now show that $\mathbf{u}_1 = \mathbf{0}$ and $C^T\mathbf{v}_2 = \mathbf{0}$ for all $\mathbf{v}_2 \in \mathcal{N}(A)$. If $\mathbf{v}_2 \in \mathcal{N}(A)$, from (63) it follows that $C\mathbf{u}_1 = \mathbf{0}$. Since B is invertible, (62) gives $0 = \mathbf{v}_2^T C\mathbf{u}_1 = -(C^T\mathbf{v}_2)^T B^{-1}(C^T\mathbf{v}_2)$ and $(B\mathbf{u}_1)^T B^{-1}(B\mathbf{u}_1) = 0$. Since B^{-1} is also invertible, we conclude that $B\mathbf{u}_1 = \mathbf{0}$ and $\mathbf{u}_1 = \mathbf{0}$. Finally, $\mathbf{u}_1 = \mathbf{0}$ and (62) gives $C^T\mathbf{v}_2 = \mathbf{0}$. □

Theorem 9 *Let $M \in \mathbb{R}^{9\times 9}$ be as in Theorem 8, then the smallest constant $\gamma \in [0, 1)$ satisfying (61), denoted $\gamma_{\min}$ (the CBS constant), is given by*

$$\gamma_{\min}^2 = 2\beta, \tag{65}$$

where $\beta \in \mathbb{R}^+$ is the constant in (60).

Proof Recall from (57) that $\gamma_{\min}^2$ is the largest eigenvalue $\theta_{\max}$ satisfying

$$CB^{-1}C^\top \mathbf{v}_2 = \theta A\mathbf{v}_2, \quad \mathbf{v}_2 \notin \mathcal{N}(A). \tag{66}$$

By considering the expression $Q\mathbf{u} = \mathbf{0}$ it is easy to show that

$$\mathcal{N}(Q) = \text{span}\left\{(1, 0, 1, 0)^\top, (0, 1, 0, 1)^\top\right\}, \tag{67}$$

and so $\mathcal{N}(A) \subset \mathcal{N}(Q)$. Under the stated assumptions, we have

$$CB^{-1}C^\top = \beta Q = \beta \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} =: \beta Q_1 \otimes Q_2,$$

and the set of eigenvalues is $\{2\beta, 2\beta, 0, 0\}$. The basis vectors of $\mathcal{N}(Q)$ in (67) are eigenvectors corresponding to the zero eigenvalues. In addition,

$$\mathbf{P}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \end{bmatrix} =: \mathbf{p}_1 \otimes \mathbf{p}_2,$$

is an eigenvector corresponding to $\theta = 2\beta$. To see this, note that

$$CB^{-1}C^\top \mathbf{P}_1 = \beta\,(Q_1 \otimes Q_2)\,(\mathbf{p}_1 \otimes \mathbf{p}_2) = \beta \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \otimes \mathbf{p}_2$$
$$= \beta \begin{bmatrix} 2 \\ -2 \end{bmatrix} \otimes \mathbf{p}_2 = 2\beta \mathbf{p}_1 \otimes \mathbf{p}_2 = 2\beta \mathbf{P}_1.$$

The same is true for $\mathbf{P}_2 = [-1, 1, 1, -1]^\top$. Furthermore, the vectors $\mathbf{P}_1$ and $\mathbf{P}_2$ also satisfy $A\mathbf{P}_1 = \mathbf{P}_1$ and $A\mathbf{P}_2 = \mathbf{P}_2$ (and clearly do not belong to $\mathcal{N}(A)$, see (64)) and hence are eigenvectors of A with eigenvalue $\theta = 1$. Thus

$$CB^{-1}C^\top \mathbf{P}_i = \beta Q \mathbf{P}_i = 2\beta \mathbf{P}_i = 2\beta(1)\mathbf{P}_i = 2\beta A \mathbf{P}_i, \quad i = 1, 2.$$

That is, $\mathbf{P}_1$ and $\mathbf{P}_2$ are eigenvectors in (66) with $\theta = 2\beta$. If we take $\mathbf{u}$ to be a member of $\mathcal{N}(Q)$ but not $\mathcal{N}(A)$, then (66) is trivially satisfied with $\theta = 0$. Hence, $\gamma_{\min}^2 = \max\{0, 2\beta\} = 2\beta$. □

The next results show that if the matrices B and C have certain structures, then $CB^{-1}C^\top$ always has the structure (60) and an explicit expression is available for the constant β in (65), and hence the CBS constant.

Lemma 1 *If the matrix $C \in \mathbb{R}^{4\times 5}$ has the form*

$$C = \alpha \begin{bmatrix} 1 & -1 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ -1 & 1 & 1 & -1 & 0 \\ -1 & -1 & 1 & 1 & 0 \end{bmatrix} =: \alpha P, \tag{68}$$

for some $\alpha \in \mathbb{R}^+$ and if $B \in \mathbb{R}^{5\times 5}$ is an invertible bordered matrix of the form

$$B = \begin{bmatrix} \bar{B} & \mathbf{b} \\ \mathbf{b}^\top & \mu \end{bmatrix}, \tag{69}$$

where $\bar{B} \in \mathbb{R}^{4\times 4}$ is a symmetric circulant matrix, $\mathbf{b} \in \mathbb{R}^4$ is a constant vector, and $\mu \in \mathbb{R}$ is a constant, then the matrix $CB^{-1}C^\top$ has the form (60).

Proof First we show that if B has the form (69) then so does B^{-1}. We have

$$B^{-1} = \begin{bmatrix} \bar{B}^{-1} + \nu^{-1}\bar{B}^{-1}\mathbf{b}\mathbf{b}^\top\bar{B}^{-1} & -\nu^{-1}\bar{B}^{-1}\mathbf{b} \\ -\nu^{-1}\mathbf{b}^\top\bar{B}^{-1} & \nu^{-1} \end{bmatrix}$$

where $\nu := \mu - \mathbf{b}^\top\bar{B}^{-1}\mathbf{b} \in \mathbb{R}$ is the Schur complement. Since $\bar{B}$ is symmetric and circulant, so is its inverse (see [13]). Consequently, $\mathbf{q} := \bar{B}^{-1}\mathbf{b} \in \mathbb{R}^{4\times 1}$ is a constant vector and $\mathbf{q}\mathbf{q}^\top \in \mathbb{R}^{4\times 4}$ is a constant matrix. This is because $\mathbf{b}$ is a constant vector and the row sums of a circulant matrix are equal. Therefore

$$B^{-1} = \begin{bmatrix} \hat{B} & \hat{\mathbf{b}} \\ \hat{\mathbf{b}}^\top & \nu^{-1} \end{bmatrix} \tag{70}$$

where $\hat{B} := \bar{B}^{-1} + \nu^{-1}\mathbf{q}\mathbf{q}^\top \in \mathbb{R}^{4\times 4}$ is a symmetric circulant matrix bordered by $\hat{\mathbf{b}} := -\nu^{-1}\mathbf{q} \in \mathbb{R}^{4\times 1}$ and $\nu \in \mathbb{R}$ is a constant. Hence, B^{-1} has the form

$$B^{-1} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_2 & \times \\ \alpha_2 & \alpha_1 & \alpha_2 & \alpha_3 & \times \\ \alpha_3 & \alpha_2 & \alpha_1 & \alpha_2 & \times \\ \alpha_2 & \alpha_3 & \alpha_2 & \alpha_1 & \times \\ \times & \times & \times & \times & \times \end{bmatrix}, \tag{71}$$

for some $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$ and, for the rest of the proof, the elements marked with $\times$ are not important. Now, elementary matrix multiplication with C gives

$$CB^{-1} = (\alpha_1 - \alpha_3)\alpha \begin{bmatrix} 1 & -1 & -1 & 1 & \bar{\times} \\ 1 & 1 & -1 & -1 & \bar{\times} \\ -1 & 1 & 1 & -1 & \bar{\times} \\ -1 & -1 & 1 & 1 & \bar{\times} \end{bmatrix}$$

(again the elements marked with $\bar{\times}$ are not important) and

$$CB^{-1}C^\top = 4(\alpha_1 - \alpha_3)\alpha^2 \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} = \beta Q$$

with $\beta := 4(\alpha_1 - \alpha_3)\alpha^2$. □

Combining the last two results, we see that to compute the CBS constant, we only need to know α (one entry of C) and α_1 and α_3 (two entries of the first column of B^{-1}). We can determine the latter analytically by exploiting the spectral decomposition for circulant matrices. Lemma 2 is standard (for example, see [13]) and we apply it to $\bar{B}$ in (69) in Lemma 3.

Lemma 2 *Let $D \in \mathbb{R}^{n\times n}$ be a circulant matrix with first column given by $\boldsymbol{d} = [d_0, d_1, \ldots, d_{n-1}]^\top \in \mathbb{R}^n$. The eigenvalues λ_j and eigenvectors $\boldsymbol{v}_j$ of D are*

$$\lambda_j = \sum_{k=0}^{n-1} d_k \omega_j^k, \qquad \boldsymbol{v}_j = n^{(-1/2)}[1, \omega_j, \omega_j^2, \ldots, \omega_j^{n-1}]^\top \tag{72}$$

where $\omega_j = \exp(2\pi i j/n)$ and $i = \sqrt{-1}$.

Lemma 3 *Let the principle minor $\bar{B}$ in (69) of the matrix B be given by*

$$\bar{B} = \begin{bmatrix} b_1 & b_2 & b_3 & b_2 \\ b_2 & b_1 & b_2 & b_3 \\ b_3 & b_2 & b_1 & b_2 \\ b_2 & b_3 & b_2 & b_1 \end{bmatrix}. \tag{73}$$

Then the eigenvalues of $\bar{B}$ are

$$\lambda_1 = b_1 - b_3, \quad \lambda_2 = b_1 - 2b_2 + b_3, \quad \lambda_3 = \lambda_1, \quad \lambda_4 = b_1 + 2b_2 + b_3, \tag{74}$$

and the eigenvectors of $\bar{B}$ are given by the columns of the unitary matrix

$$F^* = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ i & -1 & -i & 1 \\ -1 & 1 & -1 & 1 \\ -i & -1 & i & 1 \end{bmatrix}.$$

Moreover, $\bar{B} = F^* diag(\boldsymbol{\lambda}) F$, where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]^\top$.

Proof Since $\bar{B}$ is circulant, its eigenpairs are given by (72). Here, $n = 4$ and we have $\omega_1 = \mathrm{i}, \omega_2 = -1, \omega_3 = -\mathrm{i}$, and $\omega_4 = 1$. The decomposition is standard (see [34, Corollary 5.16]). □

Combining the above results, gives the following final result.

Theorem 10 *Let the assumptions of Theorem 8 and Lemma 1 hold, with the entries of $\bar{B}$ labelled as in (73). Then, the square of the CBS constant associated with (61) is given by*

$$\gamma_{\min}^2 = 8\alpha^2 (b_1 - b_3)^{-1}. \tag{75}$$

Proof From Lemma 1, we have $C B^{-1} C^\top = \beta Q$ with $\beta = 4\alpha^2(\alpha_1 - \alpha_3)$ where α_1 and α_3 are elements of the matrix $\hat{B}$ in (70), which depends on the inverse of $\bar{B}$ in (69). By Lemma 3,

$$\bar{B}^{-1} = F^* \begin{bmatrix} \lambda_1^{-1} & 0 & 0 & 0 \\ 0 & \lambda_2^{-1} & 0 & 0 \\ 0 & 0 & \lambda_1^{-1} & 0 \\ 0 & 0 & 0 & \lambda_4^{-1} \end{bmatrix} F.$$

Since $\bar{B}^{-1}$ is circulant, its entries are known once we specify its first column $\bar{\mathbf{c}}$. Furthermore, since

$$F = (F^*)^* = \frac{1}{2}\begin{bmatrix} 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

we have $\bar{\mathbf{c}} := \bar{B}^{-1}\mathbf{e}_1 = F^* \mathrm{diag}(1./\boldsymbol{\lambda}) F \mathbf{e}_1 = \frac{1}{2} F^* (1./\boldsymbol{\lambda})$. It follows that

$$\bar{\mathbf{c}} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ i & -1 & -i & 1 \\ -1 & 1 & -1 & 1 \\ -i & -1 & i & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^{-1} \\ \lambda_2^{-1} \\ \lambda_1^{-1} \\ \lambda_4^{-1} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2\lambda_1^{-1} + \lambda_2^{-1} + \lambda_4^{-1} \\ \lambda_4^{-1} - \lambda_2^{-1} \\ \lambda_2^{-1} + \lambda_4^{-1} - 2\lambda_1^{-1} \\ \lambda_4^{-1} - \lambda_2^{-1} \end{bmatrix}. \tag{76}$$

Now, since $\hat{B} := \bar{B}^{-1} + \nu^{-1}\mathbf{q}\mathbf{q}^\top$ in Lemma 1, we know that $\alpha_1 = [\bar{\mathbf{c}}]_1 + \tau$ and $\alpha_3 = [\bar{\mathbf{c}}]_3 + \tau$ for some $\tau \in \mathbb{R}$, and consequently, by considering (76) and the eigenvalues (74), we have

$$\alpha_1 - \alpha_3 = [\bar{\mathbf{c}}]_1 - [\bar{\mathbf{c}}]_3 = \frac{1}{4}\left(2\lambda_1^{-1} + 2\lambda_1^{-1}\right) = \lambda_1^{-1} = (b_1 - b_3)^{-1}.$$

Since B is symmetric and positive definite, so is $\bar{B}$. Consequently, $\lambda_1 > 0$ and $\beta = 4\alpha^2 (b_1 - b_3)^{-1} > 0$. The result follows by Theorem 9. □

Table 3 The constants $\alpha, b_1, b_3 \in \mathbb{R}$ required to compute $\gamma^2_{k,\min} = 8\alpha^2(b_1 - b_3)$, when $H_{k,1}$ is the local $\mathbb{Q}_1$ space and $H_{k,2}$ is chosen as in Example 1

$H_{k,2}$	α	b_1	b_3	$\gamma^2_{k,\min}$
$\mathbb{Q}_2(h)$	1/3	88/45	0	0.4545
$\mathbb{Q}_4(h)$	1/15	373/127	1/2326	0.0121
$\mathbb{Q}_1(h/2)$	1/4	4/3	0	0.3750
$\mathbb{Q}_2(h/2)$	1/12	56/45	0	0.0446

Table 4 Effectivity indices $\hat{\theta}_{\text{eff}}$ for Test Problem 1 with $H_1 = \mathbb{Q}_1(h)$ and four choices of H_2, for varying h (with p fixed) and varying p (with h fixed)

h	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2^{-2}	1.9254×10^{-2}	0.97	0.16	0.87	1.04
2^{-3}	1.0244×10^{-2}	0.93	0.22	0.84	0.99
2^{-4}	6.2277×10^{-3}	0.80	0.31	0.73	0.85
2^{-5}	4.7187×10^{-3}	0.62	0.39	0.59	0.65
p	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
$2, \ldots, 6$	$\approx 1.02 \times 10^{-2}$	0.93	0.22	0.84	0.99

The matrices B and C associated with the examples in Sect. 4.1 (corresponding to the four choices of $H_{k,2}$ from Example 1) *all* have the desired structure. The associated values of α, b_1 and b_3, and the squares of the CBS constants are recorded in Table 3 (to stress that these are local quantities we reintroduce the subscript k). The results match the numerical estimates obtained in Sect. 4.1. With the new approach, the matrices A, B and C do not need to be assembled, and no eigenvalue problem needs to be solved.

6 Numerical Results

We now return to (1a)–(1b) and assess the quality of the energy error estimator η in (37), extending the discussion in [8] and [10]. First, we select $X = H_1 \otimes P$ and compute $u_X \in X$ by solving (9). We choose either $H_1 = \mathbb{Q}_1(h)$ or $H_1 = \mathbb{Q}_2(h)$ on a uniform square partition of D and fix P to be the space of global polynomials with total degree less than or equal to p in $y_1, y_2, \ldots, y_M$. Each parameter y_m is assumed to be the image of a mean zero uniform random variable. Hence, for a given multi-index $\mu \in J_P$ we construct the basis functions in (30) by tensorizing univariate Legendre polynomials. Next, we compute $\eta = \eta(u_X)$ in (37) by solving (35) and (36), choosing H_2 and Q so that the conditions in (31) are satisfied. For H_2, we consider the spaces from Examples 1 and 2. We choose Q to be the space of polynomials associated with $J_Q := \hat{J}_Q \backslash J_P$ where $\hat{J}_Q$ is the set of multi-indices associated with the space of polynomials with total degree less than or equal to $p+1$ in $y_1, y_2, \ldots, y_M, y_{M+1}$.

By Theorem 6, the *effectivity index* $\theta_{\text{eff}} := \eta(u_X)/||u - u_X||_B$ satisfies

$$\frac{\sqrt{1-\gamma^2}\sqrt{1-\beta^2}}{\sqrt{\Lambda}} \leq \theta_{\text{eff}} \leq \frac{1}{\sqrt{\lambda}}. \tag{77}$$

Table 5 Effectivity indices $\hat{\theta}_{\text{eff}}$ for Test Problem 1 with $H_1 = \mathbb{Q}_2(h)$ and four choices of H_2, for varying h (with p fixed) and varying p (with h fixed)

h	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2^{-1}	4.1729×10^{-3}	0.48	0.49	0.46	0.57
$2^{-2,-3,-4}$	$\approx 4.09 \times 10^{-3}$	0.44	0.44	0.44	0.44
p	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2	4.1134×10^{-3}	0.45	0.45	0.45	0.45
$3, \ldots, 6$	$\approx 4.10 \times 10^{-3}$	0.44	0.44	0.44	0.44

Table 6 Effectivity indices $\hat{\theta}_{\text{eff}}$ for Test Problem 1 with $H_1 = \mathbb{Q}_2(h)$ and four choices of H_2, for varying h (with p fixed) and varying p (with h fixed). Modified choice of Q

h	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2^{-1}	4.1729×10^{-3}	0.83	0.83	0.81	0.82
$2^{-2,-3,-4}$	$\approx 4.09 \times 10^{-3}$	0.81	0.82	0.81	0.81
p	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2	4.1191×10^{-3}	0.82	0.82	0.82	0.82
$3, \ldots, 6$	$\approx 4.10 \times 10^{-3}$	0.81	0.81	0.81	0.81

Since $||u - u_X||_B$ cannot be computed, we examine $\hat{\theta}_{\text{eff}} := \eta(u_X)/||u_{\text{ref}} - u_X||_B$, where $u_{\text{ref}} \in X_{\text{ref}}$ is a surrogate solution obtained by solving (9) over a sufficiently rich subspace $X_{\text{ref}} \subset V$ where $X_{\text{ref}} \supset X$. We define X_{ref} in the same way as X, with $M_{\text{ref}} = 10$, $h_{\text{ref}} = 2^{-7}$ and $p_{\text{ref}} = 8$. Fixing H_1, P and Q, we investigate which choice of H_2 consistently leads to $\hat{\theta}_{\text{eff}} \approx 1$.

6.1 Test Problem 1

To start, we consider a test problem from [8]. We choose $f(\mathbf{x}) = \frac{1}{8}(2 - x_1^2 - x_2^2)$ for $\mathbf{x} = (x_1, x_2)^T \in D := [-1, 1]^2$ and assume that $a(\mathbf{x}, \mathbf{y})$ is the parametric form of a second order random field with mean $\mathbb{E}[a](\mathbf{x})$ and covariance function

$$C[a](\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{|x_1 - x_1'|}{l_1} - \frac{|x_2 - x_2'|}{l_2}\right), \quad \mathbf{x}, \mathbf{x}' \in D. \tag{78}$$

We may then expand $a(\mathbf{x}, \mathbf{y})$ using the Karhunen-Loève expansion, namely;

$$a(\mathbf{x}, \mathbf{y}) = \mathbb{E}[a](\mathbf{x}) + \sigma\sqrt{3}\sum_{m=1}^{\infty}\sqrt{\lambda_m}\phi_m(\mathbf{x})y_m, \quad y_m \in \Gamma_m = [-1, 1], \tag{79}$$

where (λ_m, ϕ_m) are the eigenpairs of the covariance operator. We choose $\mathbb{E}[a](\mathbf{x}) = 1$ (the mean), $\sigma = 0.15$ (the standard deviation) and $l_1 = l_2 = 2$ (the correlation lengths). In [8], (79) is truncated a priori after M terms, so that the problem is posed on $D \times \bar{\Gamma}$, where $\bar{\Gamma} = \prod_{m=1}^{M} \Gamma_m$. In that case, u_X and u_{ref} are both functions of M parameters. Here, u depends on infinitely many parameters and u_{ref} is a function of $M_{\text{ref}} > M$ parameters.

In our first experiment we choose $H_1 = \mathbb{Q}_1(h)$ and fix $M = 5$ in the definition of P. In Table 4 we record $\hat{\theta}_{\text{eff}}$ for varying h with fixed $p = 4$, and varying p with fixed $h = 2^{-3}$. We see that $H_2 = \mathbb{Q}_2(h/2)$ yields the best error estimator. Interestingly, $H_2 = \mathbb{Q}_4(h)$ defines the worst estimator, despite the fact that its associated CBS constant is the smallest ($\gamma^2_{\min} \leq 0.0121$). Recall from Theorem 7 that $||e_{Y_1}||_{B_0}$ and $||e_{Y_2}||_{B_0}$ provide estimates of the energy error reductions associated with augmenting H_1 with H_2, and P with Q, respectively. When $H_2 = \mathbb{Q}_4(h)$, since the CBS constant is small, we know $||e_{Y_1}||_{B_0}$ is a good estimate. When *both* $||e_{Y_1}||_{B_0}$ and $||e_{Y_2}||_{B_0}$ are much smaller than $||e||_B$ (which is true when we use the stated Q and $H_2 = \mathbb{Q}_4(h)$), the saturation constant $\beta \approx 1$. This causes η to be much smaller than $||e||_B$, resulting in a poor effectivity index.

We now repeat the experiment with $H_1 = \mathbb{Q}_2(h)$. Note that for a fixed h, the spatial error associated with u_X is smaller than for $H_1 = \mathbb{Q}_1(h)$. Results are presented in Table 5. Now, as we vary both h (for p=4 fixed) and p (for $h = 2^{-3}$ fixed), the error $||u_{\text{ref}} - u_X||_B$ stagnates. The estimated errors behave the same way, but $\hat{\theta}_{\text{eff}}$ is not close to one. There is little benefit in computing a new Galerkin solution by augmenting either H_1 with H_2 (for any of the choices of H_2) or P with Q. The saturation constant is close to one in all cases. However, if introducing more parameters into the approximation space leads to a smaller saturation constant, a better estimate of the error should be obtained by modifying Q to include more parameters.

We fix P as before with $M = 5$ but now choose Q to be the space of polynomials associated with $J_Q := \hat{J}_Q \backslash J_P$ where $\hat{J}_Q$ is the set of multi-indices associated with polynomials with total degree less than or equal to $p + 1$ in the first $M + 3$ parameters. Results are presented in Table 6. The effectivity indices are much improved. It is well known that the eigenvalues λ_m associated with (78) decay very slowly ($\sqrt{\lambda_m} = O(m^{-1})$, see [24]). To achieve a small saturation constant, and hence an accurate error estimator, a large number of parameters need to be incorporated into Q. We now study a problem with faster decaying coefficients.

6.2 Test Problem 2

We consider a problem as in [10], first introduced in [16]. We choose $f(\mathbf{x}) = 1$ for $\mathbf{x} = (x_1, x_2)^T \in D := [0, 1]^2$ and

$$a(\mathbf{x}, \mathbf{y}) = 1 + \sum_{m=1}^{\infty} \alpha_m \cos(2\pi\beta_m^1 x_1) \cos(2\pi\beta_m^2 x_2) y_m, \quad y_m \in \Gamma_m, \tag{80}$$

where $\beta_m^1 = m - k_m(k_m + 1)/2$ and $\beta_m^2 = k_m - \beta_m^1$ for $m \in \mathbb{N}$ with $k_m = \lfloor -1/2 + (1/4 + 2m)^{1/2} \rfloor$. As in [16] we select $\alpha_m = 0.547 m^{-2}$. We conduct the same experiments as in Sect. 6.1 using the original definition of $\hat{J}_Q$ (with $M + 1$ parameters). Effectivity indices are shown in Tables 7 and 8. When $H_1 = \mathbb{Q}_1(h)$, $H_2 = \mathbb{Q}_2(h)$ yields the best estimator, very closely followed by $\mathbb{Q}_2(h/2)$. When $H_1 = \mathbb{Q}_2(h)$, $H_2 = \mathbb{Q}_4(h)$ yields the best error estimator, closely followed by $\mathbb{Q}_2^r(h/2)$ (recall that $\mathbb{Q}_4^r(h)$ has the smallest CBS constant).

7 Summary and Conclusions

Using classical theory from [1,6] for Galerkin approximation, we provided an alternative derivation of an error estimator from [10] and the associated bound. Our approach highlights the straightforward extension of an error estimation strategy for standard Galerkin FEMs for

Table 7 Effectivity indices $\hat{\theta}_{\text{eff}}$ for Test Problem 2 with $H_1 = \mathbb{Q}_1(h)$ and four choices of H_2, for varying h (with p fixed) and varying p (with h fixed)

h	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2^{-3}	3.0684×10^{-2}	0.95	0.13	1.31	0.93
2^{-4}	1.5396×10^{-2}	0.95	0.14	1.32	0.94
2^{-5}	7.7745×10^{-3}	0.95	0.18	1.32	0.93
p	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_2(h)$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_1(h/2)$	$\mathbb{Q}_2(h/2)$
2	3.0723×10^{-2}	0.95	0.13	1.31	0.93
$3, \ldots, 6$	$\approx 3.09 \times 10^{-2}$	0.95	0.12	1.31	0.93

Table 8 Effectivity indices $\hat{\theta}_{\text{eff}}$ for Test Problem 2 with $H_1 = \mathbb{Q}_2(h)$ and four choices of H_2, for varying h (with p fixed) and varying p (with h fixed). $M = 5$ in the definition of P

h	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2^{-3}	2.4871×10^{-3}	1.10	1.17	0.82	0.95
2^{-4}	1.4017×10^{-3}	0.82	0.85	0.73	0.77
2^{-5}	1.2813×10^{-3}	0.71	0.71	0.70	0.70
p	$\|\|u_{\text{ref}} - u_X\|\|_B$	$\mathbb{Q}_4(h)$	$\mathbb{Q}_2(h/2)$	$\mathbb{Q}_4^r(h)$	$\mathbb{Q}_2^r(h/2)$
2	3.1896×10^{-3}	1.03	1.08	0.86	0.93
3	2.5511×10^{-3}	1.10	1.16	0.83	0.95
4, 5, 6	$\approx 2.49 \times 10^{-3}$	1.10	1.18	0.82	0.95

deterministic PDEs to SGFEMs for parameter-dependent PDEs. The quality of the estimator depends on a CBS constant associated with two finite element spaces H_1 and H_2. For $H_1 = \mathbb{Q}_1(h)$ and $H_1 = \mathbb{Q}_2(h)$ we investigated non-standard choices of H_2 which lead to small CBS constants. When $H_1 = \mathbb{Q}_1(h)$ and H_2 satisfies certain conditions, we derived new theoretical estimates for the CBS constant. In Sect. 6 we demonstrated that the best choice of H_2 for constructing an effective error estimator is not necessarily the space that leads to the smallest CBS constant. Through numerical experiments, we demonstrated that Q must also be carefully selected and tailored to properties of the diffusion coefficient. When both H_2 and Q are chosen appropriately, the estimator exhibits effectivity indices close to one.

Choosing H_2 and Q so that the effectivity index is close to one is not the end of the story. If the estimated error associated with $u_X \in X = H_1 \otimes P$ is too high, we need to decide how to enrich X and compute a new approximation. The error estimate needs to be accurate, but to derive adaptive algorithms using (44)–(45), we should only work with spaces H_2 and Q such that it is straight-forward to compute new SGFEM approximations in $(H_1 \oplus H_2) \otimes P$ and/or $H_1 \otimes (P \oplus Q)$. For example, when $H_1 = \mathbb{Q}_1(h)$, choosing $H_2 = \mathbb{Q}_2(h)$ yields an accurate error estimate for the current approximation, but does not give a feasible spatial adaptive enrichment strategy. Choosing $H_2 = \mathbb{Q}_1(h/2)$ is more natural. Fortunately, this space also yields a good error estimator. When $H_1 = \mathbb{Q}_2(h)$, $H_2 = \mathbb{Q}_2^r(h/2)$ yields an excellent estimator. Although using $H_2 = \mathbb{Q}_2(h/2)$ is more natural for adaptivity, we recommend using $H_2 = \mathbb{Q}_2^r(h/2)$ to estimate the error. Not only is this cheapest option of all those

considered, since $\mathbb{Q}_2(h/2)$ is richer, the estimated spatial error reduction $||e_{Y_1}||_{B_0}$ obtained using $H_2 = \mathbb{Q}_2^r(h/2)$ is still informative, if we wish to assess the benefit of computing a new approximation in $(\mathbb{Q}_2(h) \oplus \mathbb{Q}_2(h/2)) \otimes P$.

Acknowledgements We thank David Silvester and Alexei Bespalov for valuable discussions and contributions to the MATLAB toolbox S-IFISS [7], which we used to produce our numerical results.

References

1. Ainsworth, M., Oden, J.T.: A Posteriori Error Estimation in Finite Element Analysis. Wiley, New York (2000)
2. Axelsson, O., Blaheta, R., Neytcheva, M., Pultarová, I.: Preconditioning of iterative methods–theory and applications. Numer. Linear Algebra Appl. **22**, 901–902 (2015)
3. Axelsson, O., Vassilevski, P.S.: Algebraic multilevel preconditioning methods. II. SIAM J. Numer. Anal. **27**, 1569–1590 (1990)
4. Babuška, I.M., Tempone, R., Zouraris, G.E.: Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal. **42**, 800–825 (2004)
5. Babuška, I.M., Tempone, R., Zouraris, G.E.: Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. Comput. Methods Appl. Mech. Eng. **194**, 1251–1294 (2005)
6. Bank, R.E., Weiser, A.: Some a posteriori error estimators for elliptic partial differential equations. Math. Comput. **44**, 283–301 (1985)
7. Bespalov, A., Powell, C.E., Silvester, D.: Stochastic IFISS (S-IFISS) version 1.1. http://www.manchester.ac.uk/ifiss/s-ifiss1.0.tar.gz (2016). Accessed 6 Mar 2018
8. Bespalov, A., Powell, C.E., Silvester, D.: Energy norm a posteriori error estimation for parametric operator equations. SIAM J. Sci. Comput. **36**, A339–A363 (2014)
9. Bespalov, A., Rocchi, L.: Efficient adaptive algorithms for elliptic PDEs with random data. SIAM/ASA J. Uncertain. Quantif. **6**, 243–272 (2018)
10. Bespalov, A., Silvester, D.: Efficient adaptive stochastic Galerkin methods for parametric operator equations. SIAM J. Sci. Comput. **38**, A2118–A2140 (2016)
11. Bieri, M., Andreev, R., Schwab, C.: Sparse tensor discretization of elliptic SPDEs. SIAM J. Sci. Comput. **31**, 4281–4304 (2009)
12. Bieri, M., Schwab, C.: Sparse high order FEM for elliptic sPDEs. Comput. Methods Appl. Mech. Eng. **198**, 1149–1170 (2009)
13. Davis, P.J.: Circulant Matrices. Wiley, New York (1979)
14. Deb, M.K., Babuška, I.M., Oden, J.T.: Solution of stochastic partial differential equations using Galerkin finite element techniques. Comput. Methods Appl. Mech. Eng. **190**, 6359–6372 (2001)
15. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
16. Eigel, M., Gittelson, C.J., Schwab, C., Zander, E.: Adaptive stochastic Galerkin FEM. Comput. Methods Appl. Mech. Eng. **270**, 247–269 (2014)
17. Eigel, M., Gittelson, C.J., Schwab, C., Zander, E.: A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes. ESAIM Math. Model. Numer. Anal. **49**, 1367–1398 (2015)
18. Eigel, M., Merdon, C.: Local equilibration error estimators for guaranteed error control in adaptive stochastic higher-order Galerkin finite element methods. SIAM/ASA J. Uncertain. Quantif. **4**, 1372–1397 (2016)
19. Eijkhout, V., Vassilevski, P.: The role of the strengthened Cauchy–Buniakowskiĭ-Schwarz inequality in multilevel methods. SIAM Rev. **33**, 405–419 (1991)
20. Ernst, O.G., Powell, C.E., Silvester, D.J., Ullmann, E.: Efficient solvers for a linear stochastic Galerkin mixed formulation of diffusion problems with random data. SIAM J. Sci. Comput. **31**, 1424–1447 (2008)
21. Ernst, O.G., Ullmann, E.: Stochastic Galerkin matrices. SIAM J. Matrix Anal. Appl. **31**, 1848–1872 (2009)

22. Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements: A Spectral Approach. Springer, New York (1991)
23. Liao, Q.: Error estimation and stabilization for low order finite elements. Ph.D. Thesis. The University of Manchester, Manchester (2010)
24. Lord, G.J., Powell, C.E., Shardlow, T.: An Introduction to Computational Stochastic PDEs. Cambridge University Press, New York (2014)
25. Powell, C.E., Elman, H.C.: Block-diagonal preconditioning for spectral stochastic finite-element systems. IMA J. Numer. Anal. **29**, 350–375 (2009)
26. Powell, C.E., Ullmann, E.: Preconditioning stochastic Galerkin saddle point systems. SIAM J. Matrix Anal. Appl. **31**, 2813–2840 (2010)
27. Pultarová, I.: The strengthened C.B.S. inequality constant for second order elliptic partial differential operator and for hierarchical bilinear finite element functions. Appl. Math. **50**, 323–329 (2005)
28. Pultarová, I.: Preconditioning and a posteriori error estimates using h- and p-hierarchical finite elements with rectangular supports. Numer. Linear Algebra Appl. **16**, 415–430 (2009)
29. Pultarová, I.: Adaptive algorithm for stochastic Galerkin method. Appl. Math. **60**, 551–571 (2015)
30. Pultarová, I.: Hierarchical preconditioning for the stochastic Galerkin method: upper bounds to the strengthened CBS constants. Comput. Math. Appl. **71**, 949–964 (2016)
31. Schwab, C., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numer. **20**, 291–467 (2011)
32. Silvester, D.: Pranjal: an optimal solver for linear systems arising from stochastic FEM approximation of diffusion equations with random coefficients. SIAM/ASA J. Uncertain. Quantif. **4**, 298–311 (2016)
33. Ullmann, E.: A Kronecker product preconditioner for stochastic Galerkin finite element discretizations. SIAM J. Sci. Comput. **32**, 923–946 (2010)
34. Vogel, C.R.: Computational Methods for Inverse Problems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002)