Truncation preconditioners for Stochastic Galerkin finite element discretizations

Bespalov, Alex; Loghin, Daniel; Youngnoi, Rawin

DOI:
10.1137/20M1345645

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Bespalov, A, Loghin, D & Youngnoi, R 2021, 'Truncation preconditioners for Stochastic Galerkin finite element discretizations', *SIAM Journal on Scientific Computing*, vol. 2021, pp. S92-S116.
https://doi.org/10.1137/20M1345645

SIAM J. SCI. COMPUT.
Vol. 0, No. 0, pp. S92–S116

TRUNCATION PRECONDITIONERS FOR STOCHASTIC GALERKIN FINITE ELEMENT DISCRETIZATIONS*

ALEX BESPALOV†, DANIEL LOGHIN†, AND RAWIN YOUNGNOI†

Abstract. The stochastic Galerkin finite element method (SGFEM) provides an efficient alternative to traditional sampling methods for the numerical solution of linear elliptic partial differential equations with parametric or random inputs. However, computing stochastic Galerkin approximations for a given problem requires the solution of large coupled systems of linear equations. Therefore, an effective and bespoke iterative solver is a key ingredient of any SGFEM implementation. In this paper, we analyze a class of truncation preconditioners for SGFEM. Extending the idea of the mean-based preconditioner, these preconditioners capture additional significant components of the stochastic Galerkin matrix. Focusing on the parametric diffusion equation as a model problem and assuming affine-parametric representation of the diffusion coefficient, we perform spectral analysis of the preconditioned matrices and establish optimality of truncation preconditioners with respect to SGFEM discretization parameters. Furthermore, we report the results of numerical experiments for model diffusion problems with affine and nonaffine parametric representations of the coefficient. In particular, we look at the efficiency of the solver (in terms of iteration counts for solving the underlying linear systems) and compare truncation preconditioners with other existing preconditioners for stochastic Galerkin matrices, such as the mean-based and the Kronecker product ones.

Key words. stochastic Galerkin methods, parametric PDEs, iterative solvers, Krylov methods, preconditioning, Gauss–Seidel approximation

AMS subject classifications. 35R60, 65C20, 65F10, 65F08, 65N22, 65N30

DOI. 10.1137/20M1345645

1. Introduction.

Over the last two decades, many new challenging problems in the field of computational partial differential equations (PDEs) have been motivated by the rapidly developing area of uncertainty quantification. Efficient numerical solution of PDE problems with parametric or uncertain inputs is one of these challenges. Several numerical methods have been developed and analyzed in this context. In particular, the stochastic Galerkin finite element method (SGFEM) [14, 1, 15] has emerged as an efficient and rapidly convergent alternative to traditional Monte Carlo sampling. However, the implementation of the SGFEM requires the solution of huge (although highly structured) linear systems. For realistic applications, such linear systems can only be solved using iterative methods equipped with effective, bespoke preconditioners. To that end, a range of linear algebra techniques have been employed including the multigrid and multilevel methods [18, 8, 25, 4, 19, 24, 9], domain decomposition methods [27, 32, 33], hierarchical methods [22, 30, 31], and Krylov methods [13, 22, 23, 16, 26, 34].

In this work, we focus on Krylov methods; in particular, for a parametric elliptic PDE problem with solution approximated by the SGFEM, we employ iterative methods of Krylov subspace type for which we design and analyze a suitable class of preconditioners.

*Received by the editors June 15, 2020; accepted for publication (in revised form) January 15, 2021; published electronically March 15, 2021.
https://doi.org/10.1137/20M1345645
Funding: The work of the first author 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, Birmingham, B15 2TT, UK (a.bespalov@bham.ac.uk, d.loghin@bham.ac.uk, rxy618@student.bham.ac.uk).

As a model problem we consider the parametric steady-state diffusion equation subject to homogeneous Dirichlet boundary conditions:

$$\begin{aligned} -\nabla \cdot (a(\boldsymbol{x},\boldsymbol{y})\nabla u(\boldsymbol{x},\boldsymbol{y})) &= f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega,\ \boldsymbol{y} \in \boldsymbol{\Gamma}, \\ u(\boldsymbol{x},\boldsymbol{y}) &= 0, \qquad \boldsymbol{x} \in \partial\Omega,\ \boldsymbol{y} \in \boldsymbol{\Gamma}, \end{aligned} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ is a bounded (spatial) domain with Lipschitz polygonal boundary $\partial\Omega$, $f \in H^{-1}(\Omega)$, and $\boldsymbol{\Gamma} := \prod_{m=1}^{\infty} \Gamma_m$ is the parameter domain with bounded intervals $\Gamma_m \subset \mathbb{R}$, $m \in \mathbb{N}$. We also note that ∇ denotes the spatial gradient operator $\nabla_{\boldsymbol{x}}$.

The SGFEM applied to problem (1.1) generates approximations in tensor product spaces $X \otimes S$, where X is a finite element space associated with the physical domain Ω, and S is a space of multivariate polynomials over a finite-dimensional manifold $\boldsymbol{\Gamma}_M \subset \boldsymbol{\Gamma}$ (here, $M \in \mathbb{N}$ refers to the number active parameters in the SGFEM approximation). A typical SGFEM discretization of problem (1.1) yields a structured linear system $A\mathbf{u} = \mathbf{f}$ with the coefficient matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N_{\boldsymbol{y}}} \\ A_{21} & A_{22} & \cdots & A_{2N_{\boldsymbol{y}}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N_{\boldsymbol{y}}1} & A_{N_{\boldsymbol{y}}2} & \cdots & A_{N_{\boldsymbol{y}}N_{\boldsymbol{y}}} \end{bmatrix},$$

where the blocks A_{ij} are $N_{\boldsymbol{x}} \times N_{\boldsymbol{x}}$ matrices with $N_{\boldsymbol{x}} = \dim(X)$ and $N_{\boldsymbol{y}} = \dim(S)$. If the parametric coefficient $a(\boldsymbol{x},\boldsymbol{y})$ in (1.1) is represented via a (truncated or infinite) series expansion that is *affine* in parameters, e.g.,

$$a(\boldsymbol{x},\boldsymbol{y}) = a_0(\boldsymbol{x}) + \sum_{m=1}^{\infty} a_m(\boldsymbol{x}) y_m, \quad \boldsymbol{x} \in \Omega,\ \boldsymbol{y} \in \boldsymbol{\Gamma},$$

then it is well known (see, e.g., [21, Chapter 9]) that the system matrix A can be written as a sum of Kronecker products:

$$A = G_0 \otimes K_0 + \sum_{m=1}^{M} G_m \otimes K_m. \tag{1.2}$$

Here, $G_m \in \mathbb{R}^{N_{\boldsymbol{y}} \times N_{\boldsymbol{y}}}$ are the (parametric) matrices built from polynomial basis functions in S, and $K_m \in \mathbb{R}^{N_{\boldsymbol{x}} \times N_{\boldsymbol{x}}}$ are the (spatial) stiffness matrices associated with coefficients $a_m(\boldsymbol{x})$ in the series expansion of $a(\boldsymbol{x},\boldsymbol{y})$.

The numerical solution of stochastic Galerkin linear systems presents significant challenges. On the one hand, it is evident from the structure of A indicated above that such matrices can reach huge sizes very quickly. For example, if S is the space of complete polynomials of degree $\leq k$ in M parameters, then $N_{\boldsymbol{y}} = \binom{M+k}{k}$ grows very fast with M and k. On the other hand, in the case of affine-parametric expansion of the coefficient $a(\boldsymbol{x},\boldsymbol{y})$ as given above, the matrix A is block-sparse due to the sparsity of matrices G_m in (1.2). This feature, however, is not guaranteed for other parametric representations of $a(\boldsymbol{x},\boldsymbol{y})$ (see subsection 5.2 for one example of such a representation). Thus, the availability of effective preconditioning techniques is of paramount importance in order to enable the application of the SGFEM to a range of parametric PDE problems.

In an early effort to provide an efficient solver technique for stochastic Galerkin linear systems, the *mean-based preconditioner* was proposed by Ghanem and Kruger in [13] and subsequently analyzed by Powell and Elman in [23]. In the notation employed above, this preconditioner is defined as

$$P_0 := G_0 \otimes K_0. \tag{1.3}$$

It has been shown that under certain standard boundedness conditions on the diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$, the performance of the conjugate gradient (CG) method equipped with the preconditioner P_0 is independent of the dimensions $N_{\boldsymbol{x}}$ and $N_{\boldsymbol{y}}$ of the underlying spatial and parametric approximation spaces. This is essentially due to the mean component $a_0(\boldsymbol{x})$ strongly dominating other terms in the expansion of $a(\boldsymbol{x}, \boldsymbol{y})$. When this is not the case, the performance deteriorates and dependence on $N_{\boldsymbol{y}}$ may arise (e.g., via dependence on the number M of active parameters and/or the polynomial degree k of parametric approximations).

An alternative approach that takes into account contributions from all component matrices in (1.2) was suggested by Ullmann in [34]. This preconditioner, which we denote by $P_\otimes$, is also defined as a Kronecker product:

$$P_\otimes := G \otimes K_0, \tag{1.4}$$

where the matrix $G \in \mathbb{R}^{N_{\boldsymbol{y}} \times N_{\boldsymbol{y}}}$ is constructed in order to minimize the Frobenius norm of the difference between the system matrix and the preconditioner, i.e.,

$$G := \arg\min_Q \|A - Q \otimes K_0\|_F .$$

While the eigenvalue bounds for the preconditioned system derived in [34] are not sharp and one cannot generally expect the iteration counts of the $P_\otimes$-preconditioned CG to be independent of the dimension $N_{\boldsymbol{y}}$ of the parametric approximation space S, the *Kronecker product preconditioner* $P_\otimes$ outperforms the mean-based preconditioner P_0, particularly in the case of the approximation space S comprising polynomials of large degree k.

A preconditioning strategy that exploits the hierarchical structure of stochastic Galerkin matrices was proposed by Sousedík and Ghanem in [30]. In this strategy, the inverses of submatrices are *approximated* by inverses of their diagonal blocks in the action of a hierarchical symmetric block Gauss–Seidel preconditioner. This preconditioner is further enhanced in [30] by performing the matrix-vector multiplications in its action using only a subset of component matrices in (1.2) selected according to the size of the norm of stiffness matrices K_m. In particular, a monotonic decay of the norms of K_m effectively results in *truncating* the sum in (1.2). Extensive numerical experiments for a model problem with truncated lognormal diffusion coefficient have demonstrated the effectiveness and competitiveness of this combined preconditioning strategy (called the *truncated hierarchical preconditioning*) in terms of both convergence of iterations and computational cost. The results of these experiments have also shown that truncated versions of the nonhierarchical symmetric block Gauss–Seidel preconditioner and the approximate hierarchical Gauss–Seidel preconditioner are largely comparable in terms of convergence of iterations.

In this paper, we study a preconditioning technique based on truncating the sum of Kronecker products in (1.2) as follows:

$$P_r := G_0 \otimes K_0 + \sum_{m=1}^{r} G_m \otimes K_m.$$

We will refer to this class of solvers as *truncation preconditioners.* While it includes the mean-based preconditioner as a special case, by capturing additional significant components of the stochastic Galerkin matrix A one aims to improve the preconditioner's performance retaining its optimality with respect to the discretization parameters (i.e., $N_{\boldsymbol{x}}$, M and k). Truncation preconditioners of this type were considered in [17, section 4.2] and analyzed therein for two extreme cases, namely, $r = 0$ and $r = M - 1$. While the preconditioning matrix P_r has a block-diagonal structure in the case of the *tensor-product* polynomial space S (with appropriately ordered basis functions; see [17]), this property, in general, does not hold for the matrix P_r if S is the space of *complete* polynomials. Therefore, in the latter case, considered in the present paper, a practical implementation of the truncation preconditioner P_r requires an additional technique to ensure efficient application of the action of P_r^{-1} on a given vector. To this end, similarly to the strategy in [30], we propose to use a symmetric block Gauss–Seidel approximation $\widetilde{P}_r$ of the truncation preconditioner P_r.

Focusing on the case of affine-parametric representation of $a(\boldsymbol{x}, \boldsymbol{y})$, our main goal in this paper is to perform spectral analysis of the preconditioned matrices and establish optimality of the preconditioners P_r and $\widetilde{P}_r$ with respect to all discretization parameters. By doing this we fill a gap in the theoretical analysis of preconditioning techniques for the numerical solution of stochastic Galerkin linear systems.

The paper is structured as follows. In the next section we present a detailed problem formulation, including specific assumptions on the parametric diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$, the variational formulation of (1.1), and the definitions and properties of the matrices A_{ij}, G_m, K_m. In section 3, we introduce and analyze a class of preconditioners P_r based on truncation of the series representation of the parametric diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$. A symmetric block Gauss–Seidel approximation to P_r is introduced and analyzed in section 4. Section 5 includes a range of numerical results, with conclusions and potential extensions summarized in section 6.

2. Problem formulation. In this section we outline some standard background results and assumptions and introduce the variational formulation required for the numerical solution of (1.1) via the SGFEM.

2.1. Functional analytic framework. Let $y_m \in \Gamma_m$ be the images of independent bounded random variables with cumulative density function $\pi_m(y_m)$ and probability density function $p_m(y_m) = \mathrm{d}\pi_m(y_m)/\mathrm{d}y_m$. The joint cumulative density function and the joint probability density function of the associated multivariable random variable $\boldsymbol{y} \in \boldsymbol{\Gamma}$ are defined, respectively, as

$$\pi(\boldsymbol{y}) := \prod_{m=1}^{\infty} \pi_m(y_m) \quad \text{and} \quad p(\boldsymbol{y}) := \prod_{m=1}^{\infty} p_m(y_m).$$

Without loss of generality, we assume for all $m \in \mathbb{N}$ that $\Gamma_m := [-1, 1]$; additionally, we assume that p_m is even. 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 $(\boldsymbol{\Gamma}, \mathcal{B}(\boldsymbol{\Gamma}))$, where $\mathcal{B}(\boldsymbol{\Gamma})$ is the Borel σ-algebra on $\boldsymbol{\Gamma}$. Then $L^2_{\pi_m}(\Gamma_m)$, $L^2_{\pi}(\boldsymbol{\Gamma})$ represent the weighted Lebesgue spaces with associated inner products

$$\langle f, g\rangle_{\pi_m} := \int_{\Gamma_m} p_m(y_m) f(y_m) g(y_m) \mathrm{d}y_m, \qquad f, g \in L^2_{\pi_m}(\Gamma_m),$$

$$\langle f, g\rangle_{\pi} := \int_{\boldsymbol{\Gamma}} p(\boldsymbol{y}) f(\boldsymbol{y}) g(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}, \qquad f, g \in L^2_{\pi}(\boldsymbol{\Gamma}).$$

Finally, a space relevant to the weak formulation of problem (1.1) is $L^2_\pi(\mathbf{\Gamma}; H^1_0(\Omega))$, which is the space of strongly measurable functions $v : \Omega \times \mathbf{\Gamma} \to \mathbb{R}$ such that

$$\|v\|_{L^2_\pi(\mathbf{\Gamma};H^1_0(\Omega))} := \Big\| \|v(\cdot,\boldsymbol{y})\|_{H^1_0(\Omega)} \Big\|_{L^2_\pi(\mathbf{\Gamma})} := \left[\int_{\mathbf{\Gamma}} p(\boldsymbol{y}) \, \|v(\cdot,\boldsymbol{y})\|^2_{H^1_0(\Omega)} \, \mathrm{d}\boldsymbol{y} \right]^{1/2} < +\infty,$$

where $H^1_0(\Omega)$ is the usual Sobolev space of functions in $H^1(\Omega)$ that vanish at the boundary $\partial\Omega$ in the sense of traces. It is known that $L^2_\pi(\mathbf{\Gamma}; H^1_0(\Omega))$ is isometrically isomorphic to the following tensor product Hilbert space (see [28, Remark C.24]):

$$V := L^2_\pi(\mathbf{\Gamma}) \otimes H^1_0(\Omega). \tag{2.1}$$

We will use the space V to derive the variational formulation of problem (1.1) in subsection 2.3 below. Before doing this, let us make some specific assumptions on the parametric diffusion coefficient $a(\boldsymbol{x},\boldsymbol{y})$.

2.2. The parametric diffusion coefficient. We will assume that the diffusion coefficient a is *affine-parametric*, i.e.,

$$a(\boldsymbol{x},\boldsymbol{y}) = a_0(\boldsymbol{x}) + \sum_{m=1}^{\infty} a_m(\boldsymbol{x}) y_m, \quad \boldsymbol{x} \in \Omega, \ \boldsymbol{y} \in \mathbf{\Gamma}, \tag{2.2}$$

with $a_m \in L^\infty(\Omega)$, $m \in \mathbb{N}_0$, and with the series converging uniformly in $L^\infty(\Omega)$. The representation (2.2) is motivated by the Karhunen–Loève expansion of a second-order random field a with given mean $\mathbb{E}[a]$ and covariance function $\mathrm{Cov}[a](\boldsymbol{x},\boldsymbol{x}')$. In this case, $a(\boldsymbol{x},\boldsymbol{y})$ is represented as in (2.2) with $a_0(\boldsymbol{x}) = \mathbb{E}[a]$ and $a_m(\boldsymbol{x}) = \sqrt{\lambda_m}\varphi_m(\boldsymbol{x})$ $(m = 1, 2\ldots)$, where $\{(\lambda_m, \varphi_m)\}_{m=1}^\infty$ are the eigenpairs of the integral operator $\int_\Omega \mathrm{Cov}[a](\boldsymbol{x},\boldsymbol{x}')\varphi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'$ such that $\lambda_1 \geq \lambda_2 \geq \cdots > 0$, and y_m $(m = 1, 2\ldots)$ are the images of pairwise uncorrelated random variables with zero mean and unit variance (see, e.g., [14, section 2.3]).

As is the case for deterministic diffusion problems, the standard conditions for well-posedness of the weak formulation of problem (1.1) are the positivity and boundedness of the diffusion coefficient a. In order to ensure that the coefficient $a(\boldsymbol{x},\boldsymbol{y})$ given by (2.2) satisfies these conditions, we assume that (cf. [28, Proposition 2.22])

$$a_0^{\min} \leq a_0(\boldsymbol{x}) \leq a_0^{\max} \quad \text{a.e. in } \Omega \tag{2.3}$$

for some constants $0 < a_0^{\min} \leq a_0^{\max} < \infty$ and that

$$\tau := \frac{1}{a_0^{\min}} \Big\| \sum_{m=1}^{\infty} |a_m| \Big\|_\infty < 1, \tag{2.4}$$

where $\|\cdot\|_\infty$ denotes the norm in $L^\infty(\Omega)$. In this case, an elementary calculation shows that

$$0 < a_{\min} \leq a(\boldsymbol{x},\boldsymbol{y}) \leq a_{\max} < \infty \quad \text{a.e. in } \Omega \times \mathbf{\Gamma} \tag{2.5}$$

with $a_{\min} := a_0^{\min}(1-\tau)$ and $a_{\max} := a_0^{\max} + a_0^{\min}\tau$.

Remark 2.1. As an alternative to (2.4), one can make a weaker assumption:

$$\widetilde{\tau} := \Big\| a_0^{-1} \sum_{m=1}^{\infty} |a_m| \Big\|_\infty < 1. \tag{2.6}$$

In this case, one has

$$0 < \widetilde{a}_{\min} \le a_0(\boldsymbol{x})(1-\widetilde{\tau}) \le a(\boldsymbol{x},\boldsymbol{y}) \le a_0(\boldsymbol{x})(1+\widetilde{\tau}) \le \widetilde{a}_{\max} < \infty \quad \text{a.e. in } \Omega \times \boldsymbol{\Gamma} \tag{2.7}$$

with $\widetilde{a}_{\min} := a_0^{\min}(1-\widetilde{\tau})$ and $\widetilde{a}_{\max} := a_0^{\max}(1+\widetilde{\tau})$. We refer to [17, section 2.3] for a detailed discussion of assumption (2.6) and its comparison to the one in (2.4).

2.3. Variational formulation and Galerkin approximations. Let V be defined as in (2.1). The weak formulation of problem (1.1) reads as follows: find $u \in V$ such that

$$\mathcal{A}(u,v) = \mathcal{F}(v) \quad \forall\, v \in V, \tag{2.8}$$

where the bilinear form $\mathcal{A} : V \times V \to \mathbb{R}$ and the linear functional $\mathcal{F} : V \to \mathbb{R}$ are defined by

$$\begin{aligned} \mathcal{A}(v,w) &:= \int_{\boldsymbol{\Gamma}} p(\boldsymbol{y}) \int_{\Omega} a(\boldsymbol{x},\boldsymbol{y}) \nabla v(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla w(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \\ \mathcal{F}(w) &:= \int_{\boldsymbol{\Gamma}} p(\boldsymbol{y}) \int_{\Omega} f(\boldsymbol{x}) w(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}. \end{aligned} \tag{2.9}$$

The boundedness of the parametric coefficient a (see (2.5) and (2.7)) implies a unique solvability of (2.8) due to the Lax–Milgram theorem.

We discretize problem (2.8) by using the Galerkin projection onto a finite-dimensional subspace of V constructed as a tensor product of finite-dimensional subspaces of $H_0^1(\Omega)$ and $L^2_\pi(\boldsymbol{\Gamma})$. Let us describe these subspaces in detail. Let $\mathcal{T}_h$ denote a conforming, quasi-uniform partition of Ω into simplices K of maximum diameter h. We define

$$X_h := \{\phi : \Omega \to \mathbb{R} : \phi|_K \in \mathbb{P}_q(K)\} \cap C^0(\Omega) \subset H_0^1(\Omega),$$

where $\mathbb{P}_q(K)$ denotes the space of polynomials of degree q defined on K. We will assume that

$$X_h = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_{N_{\boldsymbol{x}}}\},$$

where $\phi_j(\boldsymbol{x})$ have local support and $N_{\boldsymbol{x}} := \dim(X_h)$.

Let us now introduce a polynomial subspace of $L^2_\pi(\boldsymbol{\Gamma})$. To this end, for each $m \in \mathbb{N}$, we first consider the sequence $\{P_j^m(y_m) : j \in \mathbb{N}_0\}$ of univariate polynomials that are orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle_{\pi_m}$ such that P_j^m is a polynomial of degree $j \in \mathbb{N}_0$. These polynomials form an orthonormal basis of $L^2_{\pi_m}(\Gamma_m)$, i.e., $L^2_{\pi_m}(\Gamma_m) = \operatorname{span}\left\{P_j^m : j \in \mathbb{N}_0\right\}$. Moreover, they satisfy the following three-term recurrence [12, Theorem 1.29]:

$$c_{j+1}^m P_{j+1}^m(y_m) = (y_m - a_j^m) P_j^m(y_m) - c_j^m P_{j-1}^m(y_m), \quad j = 0,1,2,\ldots, \tag{2.10}$$

with $P_{-1}^m(y_m) = 0$, $P_0^m(y_m) = 1/c_0^m$, where a_j^m, c_j^m are defined in terms of inner products involving the monic orthogonal polynomial counterparts to P_j^m (for details, see [12, Chapter 1]). For our choice of intervals $\Gamma_m = [-1,1]$, the recurrence coefficients a_j^m, c_j^m are bounded; in particular $a_j^m = 0$ (since the measure π_m is symmetric) and there holds (see [12, Theorem 1.28])

$$0 < c_j^m \le 1, \quad m \in \mathbb{N}, \quad j = 0,1,2,\ldots. \tag{2.11}$$

Consider now the *index set* of multi-indices $\boldsymbol{\alpha}$ with finite support

$$\mathbb{I} := \left\{\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots) \in \mathbb{N}_0^{\mathbb{N}} : \max(\operatorname{supp}\boldsymbol{\alpha}) < \infty\right\},$$

where $\operatorname{supp}\boldsymbol{\alpha} = \{m \in \mathbb{N};\ \alpha_m \neq 0\}$. For each multi-index $\boldsymbol{\alpha} \in \mathbb{I}$, we define the multivariate polynomial

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{y}) := \prod_{m \in \operatorname{supp}\boldsymbol{\alpha}} P^m_{\alpha_m}(y_m).$$

The set $\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{I}\}$ is an orthonormal basis of $L^2_\pi(\boldsymbol{\Gamma})$; see [28, Theorem 2.12]. Thus, any finite index set $\mathbb{I}_n \subset \mathbb{I}$ induces a finite dimensional subspace $\operatorname{span}\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{I}_n\}$ of $L^2_\pi(\boldsymbol{\Gamma})$. In this paper, we employ finite index sets of the following type:

$$\mathbb{I}^M_k := \{\boldsymbol{\alpha} \in \mathbb{I} : |\boldsymbol{\alpha}| \leq k \text{ and } \alpha_m = 0 \ \ \forall\, m > M\}, \quad k \in \mathbb{N}_0, \quad M \in \mathbb{N},$$

where $|\boldsymbol{\alpha}| = \sum_{m \in \operatorname{supp}\boldsymbol{\alpha}} \alpha_m$. We denote the corresponding finite-dimensional subspaces of $L^2_\pi(\boldsymbol{\Gamma})$ as

$$S^M_k := \operatorname{span}\left\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{I}^M_k\right\}. \tag{2.12}$$

Thus, S^M_k is the space of complete polynomials of degree $\leq k$ in M variables; its dimension is given by

$$N_{\boldsymbol{y}} := \dim(S^M_k) = \operatorname{card}\mathbb{I}^M_k = \binom{M+k}{k}.$$

Furthermore, there exists a bijection $\boldsymbol{\kappa} : \{1, 2, \ldots, N_{\boldsymbol{y}}\} \to \mathbb{I}^M_k$, so that we can also describe S^M_k via the span $S^M_k = \operatorname{span}\left\{\psi_{\boldsymbol{\kappa}(j)} : 1 \leq j \leq N_{\boldsymbol{y}}\right\}$.

We can now define the following finite-dimensional subspace of V:

$$V^M_{hk} := X_h \otimes S^M_k = \operatorname{span}\big\{\varphi_{ij}(\boldsymbol{x},\boldsymbol{y}) := \phi_i(\boldsymbol{x})\psi_{\boldsymbol{\kappa}(j)}(\boldsymbol{y}) : 1 \leq i \leq N_{\boldsymbol{x}},\ 1 \leq j \leq N_{\boldsymbol{y}}\big\}. \tag{2.13}$$

The resulting discrete formulation of (2.8) reads as follows: find $u^M_{hk} \in V^M_{hk}$ such that

$$\mathcal{A}(u^M_{hk}, v) = \mathcal{F}(v) \quad \forall\, v \in V^M_{hk}. \tag{2.14}$$

Using the definition of V^M_{hk} in (2.13) we write the Galerkin approximation u^M_{hk} as

$$u^M_{hk}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N_{\boldsymbol{x}}} \sum_{j=1}^{N_{\boldsymbol{y}}} u_{ij}\varphi_{ij}(\boldsymbol{x},\boldsymbol{y}). \tag{2.15}$$

The Galerkin projection onto the finite-dimensional space V^M_{hk} defined via the choice of finite index set $\mathbb{I}^M_k$ can be shown to correspond to a discrete weak formulation involving a truncation at $m = M$ of the parametric diffusion coefficient a given in (2.2). More precisely, if we let

$$a_M(\boldsymbol{x},\boldsymbol{y}) := a_0(\boldsymbol{x}) + \sum_{m=1}^{M} a_m(\boldsymbol{x})y_m, \quad \boldsymbol{x} \in \Omega,\ \boldsymbol{y} \in \boldsymbol{\Gamma},$$

then the associated bilinear form

$$\mathcal{A}_M(v,w) := \int_{\boldsymbol{\Gamma}} p(\boldsymbol{y}) \int_\Omega a_M(\boldsymbol{x},\boldsymbol{y})\nabla v(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla w(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

satisfies (see, e.g., [2, page A349])

$$\mathcal{A}_M(v,w) = \mathcal{A}(v,w) \quad \forall v, w \in V^M_{hk}. \tag{2.16}$$

Using representation (2.15) of the Galerkin solution, and setting $v = \varphi_{st}$ in (2.14) for $s = 1, \ldots, N_{\boldsymbol{x}}$ and $t = 1, \ldots, N_{\boldsymbol{y}}$, we obtain the following linear system:

$$\sum_{i=1}^{N_{\boldsymbol{x}}} \sum_{j=1}^{N_{\boldsymbol{y}}} u_{ij} \mathcal{A}(\varphi_{ij}, \varphi_{st}) = \mathcal{F}(\varphi_{st}).$$

This system becomes, using (2.16) and the separable form (2.13) of φ_{ij}, φ_{st} and of each term in the series expansion (2.2) of the diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$,

$$\sum_{m=0}^{M} \sum_{i=1}^{N_{\boldsymbol{x}}} \sum_{j=1}^{N_{\boldsymbol{y}}} u_{ij} \int_{\Omega} a_m \nabla \phi_i \cdot \nabla \phi_s \mathrm{d}\boldsymbol{x} \int_{\boldsymbol{\Gamma}} y_m \psi_{\boldsymbol{\kappa}(j)} \psi_{\boldsymbol{\kappa}(t)} p \mathrm{d}\boldsymbol{y} = \int_{\Omega} f \phi_s \mathrm{d}\boldsymbol{x} \int_{\boldsymbol{\Gamma}} \psi_{\boldsymbol{\kappa}(t)} p \mathrm{d}\boldsymbol{y},$$

where we set $y_0 = 1$. Therefore, the discrete formulation (2.14) yields a linear system $A\mathbf{u} = \mathbf{f}$ with block structure. Specifically, the coefficient matrix A, the solution vector $\mathbf{u}$, and the right-hand side vector $\mathbf{f}$ are given by

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N_{\boldsymbol{y}}} \\ A_{21} & A_{22} & \cdots & A_{2N_{\boldsymbol{y}}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N_{\boldsymbol{y}}1} & A_{N_{\boldsymbol{y}}2} & \cdots & A_{N_{\boldsymbol{y}}N_{\boldsymbol{y}}} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_{N_{\boldsymbol{y}}} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_{N_{\boldsymbol{y}}} \end{bmatrix}, \tag{2.17}$$

respectively, where

$$A_{tj} = \langle \psi_{\boldsymbol{\kappa}(j)}, \psi_{\boldsymbol{\kappa}(t)} \rangle_{\pi} K_0 + \sum_{m=1}^{M} \langle y_m \psi_{\boldsymbol{\kappa}(j)}, \psi_{\boldsymbol{\kappa}(t)} \rangle_{\pi} K_m, \quad t, j = 1, \ldots, N_{\boldsymbol{y}}$$

with finite element (stiffness) matrices K_m, $m = 0, 1, \ldots, M$, defined by

$$[K_m]_{si} := \int_{\Omega} a_m \nabla \phi_i \cdot \nabla \phi_s \, \mathrm{d}\boldsymbol{x}, \quad s, i = 1, \ldots, N_{\boldsymbol{x}},$$

$$\mathbf{u}_j := [u_{1j} \; u_{2j} \; \ldots \; u_{N_{\boldsymbol{x}}j}]^T, \quad j = 1, \ldots, N_{\boldsymbol{y}},$$

and

$$[\mathbf{f}_t]_s := \langle 1, \psi_{\boldsymbol{\kappa}(t)} \rangle_{\pi} \int_{\Omega} f \phi_s \, \mathrm{d}\boldsymbol{x}, \quad s = 1, \ldots, N_{\boldsymbol{x}}, \; t = 1, \ldots, N_{\boldsymbol{y}}.$$

Using Kronecker products for matrices, it is convenient to write the coefficient matrix A in the following compact form:

$$A = G_0 \otimes K_0 + \sum_{m=1}^{M} G_m \otimes K_m, \tag{2.18}$$

where

$$[G_0]_{tj} := \langle \psi_{\boldsymbol{\kappa}(j)}, \psi_{\boldsymbol{\kappa}(t)} \rangle_{\pi}, \quad [G_m]_{tj} := \langle y_m \psi_{\boldsymbol{\kappa}(j)}, \psi_{\boldsymbol{\kappa}(t)} \rangle_{\pi} \tag{2.19}$$

for $m = 1, \ldots, M$ and $t, j = 1, \ldots, N_{\boldsymbol{y}}$.

The stochastic Galerkin matrix A is symmetric and positive definite. Furthermore, as it follows from the theorem below, A is block sparse with no more than $2M + 1$ nonzero block matrices per row.

THEOREM 2.1 ([21, Theorems 9.58 and 9.59]). *Consider the matrices G_m defined in* (2.19) *for $m = 0, 1, \dots, M$. The matrix G_0 is the $N_{\boldsymbol{y}} \times N_{\boldsymbol{y}}$ identity matrix, and each matrix G_m for $m = 1, 2, \dots, M$ has at most two nonzero entries per row. More precisely,*

$$[G_m]_{tj} = \begin{cases} c^m_{\gamma_m+1} & \text{if } \gamma_m = \beta_m - 1 \text{ and } \gamma_\ell = \beta_\ell \ \ \forall\, \ell \in \mathbb{N} \setminus \{m\}, \\ c^m_{\gamma_m} & \text{if } \gamma_m = \beta_m + 1 \text{ and } \gamma_\ell = \beta_\ell \ \ \forall\, \ell \in \mathbb{N} \setminus \{m\}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\gamma} = \boldsymbol{\kappa}(t)$, $\boldsymbol{\beta} = \boldsymbol{\kappa}(j)$, and $c^m_{\gamma_m}$ are the coefficients arising in the three-term recurrence (2.10) *which defines the orthonormal polynomials P^m_j.*

3. Truncation preconditioners. In this section, we consider a class of preconditioners that are induced by bilinear forms associated with truncations of the series representation of the parametric diffusion coefficient a (cf. [30, 17]). We will show that these *truncated* bilinear forms are equivalent to the bilinear form arising in the variational formulation of our PDE. The immediate consequence of this fact is that the resulting *truncation preconditioners* will be optimal in some sense to be described below (see Definition 3.4).

3.1. Equivalent bilinear forms and preconditioning. A generic approach to preconditioner design for discretizations of variational problems is based on approximating the bilinear forms arising in the formulation of the problem. For symmetric and coercive problems, the well-known concept of equivalence of bilinear forms translates into spectral equivalence between the coefficient matrix and the preconditioner induced by the approximating bilinear form; in turn, spectral equivalence enables both the design and analysis of effective preconditioning techniques. We summarize this approach in Proposition 3.3 below, which requires the following two definitions.

DEFINITION 3.1. *We say that positive definite symmetric bilinear forms $\mathcal{A}, \mathcal{B} : V \times V \to \mathbb{R}$ are equivalent if there exist positive constants θ, Θ such that for all $v \in V$ there holds*

$$\theta \mathcal{B}(v, v) \le \mathcal{A}(v, v) \le \Theta \mathcal{B}(v, v).$$

DEFINITION 3.2. *We say that symmetric positive definite matrices $A, B \in \mathbb{R}^{n \times n}$ are spectrally equivalent if there exist positive constants θ, Θ independent of n such that for all $\mathbf{v} \in \mathbb{R}^n$ there holds*

$$\theta \mathbf{v}^T B \mathbf{v} \le \mathbf{v}^T A \mathbf{v} \le \Theta \mathbf{v}^T B \mathbf{v}.$$

In this case, we write $A \sim B$.

Remark 3.1. The relation $\sim$ is an equivalence relation. In particular, transitivity will be relevant in our subsequent discussion.

Bilinear form equivalence is connected to the well-known concepts of operator and spectral equivalence (see [6, 11]) as well as norm-equivalent preconditioners (see [20]). In this context, the following result is key to our subsequent analysis.

PROPOSITION 3.3. *Let $\mathcal{A}, \mathcal{B}$ denote positive definite symmetric bilinear forms on $V \times V$ which are equivalent. Let $V_n = \operatorname{span}\{\varphi_1, \dots, \varphi_n\} \subset V$, and let $A, B \in \mathbb{R}^{n \times n}$ be defined as follows:*

$$A_{ij} = \mathcal{A}(\varphi_j, \varphi_i), \qquad B_{ij} = \mathcal{B}(\varphi_j, \varphi_i).$$

Then $A \sim B$ and the spectrum of $B^{-1}A$ satisfies

$$\Lambda(B^{-1}A) \subset [\theta, \Theta], \tag{3.1}$$

where θ, Θ are the constants of equivalence for $\mathcal{A}, \mathcal{B}$.

The above result motivates the following definition.

DEFINITION 3.4. *Let $A, B \in \mathbb{R}^{n\times n}$ satisfy (3.1) with constants θ, Θ independent of n. Then B is said to be an optimal preconditioner for A with respect to the problem size n.*

Preconditioner optimality translates into performance optimality. In particular, it is well-known that the preconditioned CG algorithm applied to the linear system $A\mathbf{u} = \mathbf{f}$ with optimal preconditioner B converges in a number of steps independent of n. Our aim is to construct optimal preconditioners with respect to the problem size for the coefficient matrix in (2.17). We do this by first adapting the result of Proposition 3.3 to the parametric elliptic problem (1.1). We will need the following auxiliary result.

LEMMA 3.5. *Let $p : \mathbf{\Gamma} \to \mathbb{R}_+$, and assume that $b_i : \Omega \times \mathbf{\Gamma} \to \mathbb{R}_+$ $(i = 1, 2)$ satisfy*

$$0 < \beta_i^{\min} \le b_i(\boldsymbol{x}, \boldsymbol{y}) \le \beta_i^{\max} \quad \text{a.e. in } \Omega \times \mathbf{\Gamma} \ni (\boldsymbol{x}, \boldsymbol{y}).$$

Define the bilinear forms $\mathcal{B}_i : V \times V \to \mathbb{R}$ via

$$\mathcal{B}_i(v, w) = \int_{\mathbf{\Gamma}} p(\boldsymbol{y}) \int_\Omega b_i(\boldsymbol{x}, \boldsymbol{y}) \nabla v(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla w(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \quad i = 1, 2.$$

Then the bilinear forms $\mathcal{B}_i$ are equivalent:

$$\theta \mathcal{B}_2(v, v) \le \mathcal{B}_1(v, v) \le \Theta \mathcal{B}_2(v, v) \quad \forall\, v \in V,$$

where

$$\theta = \frac{\beta_1^{\min}}{\beta_2^{\max}}, \qquad \Theta = \frac{\beta_1^{\max}}{\beta_2^{\min}}.$$

Proof. For any $v \in V$, we have

$$\begin{aligned}
\mathcal{B}_1(v, v) &= \int_{\mathbf{\Gamma}} p(\boldsymbol{y}) \int_\Omega \left(\frac{b_1(\boldsymbol{x}, \boldsymbol{y})}{b_2(\boldsymbol{x}, \boldsymbol{y})} \right) b_2(\boldsymbol{x}, \boldsymbol{y}) \nabla v(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\
&\le \left(\operatorname*{ess\,sup}_{(\boldsymbol{x},\boldsymbol{y}) \in \Omega \times \mathbf{\Gamma}} \frac{b_1(\boldsymbol{x}, \boldsymbol{y})}{b_2(\boldsymbol{x}, \boldsymbol{y})} \right) \int_{\mathbf{\Gamma}} p(\boldsymbol{y}) \int_\Omega b_2(\boldsymbol{x}, \boldsymbol{y}) \nabla v(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla v(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\
&\le \frac{\beta_1^{\max}}{\beta_2^{\min}} \mathcal{B}_2(v, v).
\end{aligned}$$

The lower bound follows analogously. ∎

The boundedness required in the above lemma for $b_i(\boldsymbol{x}, \boldsymbol{y})$ holds for the parametric diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$ (see (2.5)). In the next subsection we show that assumptions (2.3) and (2.4), which guarantee (2.5), also yield boundedness of *truncated* expansions of the coefficient $a(\boldsymbol{x}, \boldsymbol{y})$. That result will enable our main goal—the analysis of truncation preconditioners for stochastic Galerkin matrices.

3.2. Spectral analysis. In analogy with (2.4), we define

$$\tau_0 := 0, \quad \tau_r := \frac{1}{a_0^{\min}} \left\| \sum_{m=1}^{r} |a_m| \right\|_\infty, \quad r \in \mathbb{N}. \tag{3.2}$$

Note that $(\tau_r)_{r\in\mathbb{N}_0}$ is a monotonic increasing sequence, which is bounded from above (cf. (2.4)), i.e.,

$$0 \le \tau_r \le \tau_{r+1} \le \tau < 1, \quad r \in \mathbb{N}_0.$$

First, we establish the boundedness of *truncated* expansions of the parametric diffusion coefficient a.

LEMMA 3.6. *Assume that* (2.3) *and* (2.4) *hold. Let* $r \in \mathbb{N}_0$*, and define* $a_r : \Omega \times \mathbf{\Gamma} \to \mathbb{R}$ *to be the finite sum*

$$a_r(\boldsymbol{x}, \boldsymbol{y}) := a_0(\boldsymbol{x}) + \sum_{m=1}^{r} a_m(\boldsymbol{x}) y_m. \tag{3.3}$$

Then $a_r(\boldsymbol{x}, \boldsymbol{y})$ *is positive and bounded almost everywhere in* $\Omega \times \mathbf{\Gamma}$.

Proof. The case $r = 0$ follows from the boundedness and positivity assumptions (2.3) on $a_0(\boldsymbol{x})$. Consider now $r \in \mathbb{N}$. Since $|y_m| \le 1$, using the definition (3.2) of τ_r we obtain

$$|a_r(\boldsymbol{x}, \boldsymbol{y}) - a_0(\boldsymbol{x})| = \left| \sum_{m=1}^{r} a_m(\boldsymbol{x}) y_m \right| \le \sum_{m=1}^{r} |a_m(\boldsymbol{x})| \le a_0^{\min} \tau_r \qquad \text{a.e. in } \Omega \times \mathbf{\Gamma}.$$

Hence,

$$a_0(\boldsymbol{x}) - a_0^{\min} \tau_r \le a_r(\boldsymbol{x}, \boldsymbol{y}) \le a_0(\boldsymbol{x}) + a_0^{\min} \tau_r \tag{3.4}$$

and using the boundedness of $a_0(\boldsymbol{x})$ (see (2.3)), we get

$$\eta_r^{\min} := a_0^{\min} - a_0^{\min} \tau_r \le a_r(\boldsymbol{x}, \boldsymbol{y}) \le a_0^{\max} + a_0^{\min} \tau_r =: \eta_r^{\max} \quad \text{a.e. in } \Omega \times \mathbf{\Gamma}. \tag{3.5}$$

The proof concludes by noting that $\eta_r^{\min} > 0$ since $\tau_r \le \tau < 1$ (cf. (3.2) and (2.4)). ∎

Combining Lemmas 3.5 and 3.6, we obtain the following result.

THEOREM 3.7. *Let* $a : \Omega \times \mathbf{\Gamma} \to \mathbb{R}_+$ *be a parametric diffusion coefficient given by* (2.2)*, and let* $\mathcal{A} : V \times V \to \mathbb{R}$ *be the associated bilinear form defined in* (2.9)*. Assume* (2.3) *and* (2.4) *hold. Let* a_r *be given by* (3.3)*, and define the associated bilinear form as*

$$\mathcal{A}_r(v, w) := \int_{\mathbf{\Gamma}} p(\boldsymbol{y}) \int_\Omega a_r(\boldsymbol{x}, \boldsymbol{y}) \nabla v(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla w(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}.$$

Then $\mathcal{A}$ *and* $\mathcal{A}_r$ *are equivalent for any* $r \in \mathbb{N}_0$.

Proof. By (2.5), the diffusion coefficients a is bounded. Furthermore, by Lemma 3.6, the coefficient a_r, $r \in \mathbb{N}_0$, is bounded as well. Consequently, by Lemma 3.5, the bilinear forms are equivalent. In particular,

$$\theta_r \mathcal{A}_r(v, v) \le \mathcal{A}(v, v) \le \Theta_r \mathcal{A}_r(v, v),$$

where

$$\theta_r := \frac{a_{\min}}{\eta_r^{\max}} = \frac{(1-\tau) a_0^{\min}}{a_0^{\max} + a_0^{\min} \tau_r}, \qquad \Theta_r := \frac{a_{\max}}{\eta_r^{\min}} = \frac{a_0^{\max} + a_0^{\min} \tau}{(1-\tau_r) a_0^{\min}} \tag{3.6}$$

with $\eta_r^{\min}$ and $\eta_r^{\max}$ defined in (3.5). ∎

Remark 3.2. The equivalence of the bilinear form $\mathcal{A}_0$ associated with the parameter-free term $a_0(\boldsymbol{x})$ in (2.2) and the bilinear form $\mathcal{A}$ is well known (see, e.g., [3, equation (2.5)]). Theorem 3.7 extends this result to the case of arbitrary finite truncation of the affine-parametric coefficient $a(\boldsymbol{x}, \boldsymbol{y})$.

Remark 3.3. The constants θ_r, Θ_r in (3.6) depend on τ, $a_0^{\min}$, $a_0^{\max}$, and indirectly on r via τ_r.

Theorem 3.7, combined with Proposition 3.3, indicates that the bilinear form $\mathcal{A}_r$ induces a family of preconditioners for the SGFEM matrix A in (2.17). This is made precise in the next theorem which is the main result of this section.

THEOREM 3.8. *Let $\mathcal{A}$, $\mathcal{A}_r$ be defined as in Theorem* 3.7*, and assume* (2.3) *and* (2.4) *hold. Let $\{\varphi_{ij}\}$ be the tensor-product basis for the finite dimensional space V_{hk}^M in* (2.13) *and $A = [\mathcal{A}(\varphi_{ij}, \varphi_{st})]$ be the associated SGFEM matrix. For a fixed $r \in \mathbb{N}_0$ define the preconditioner P_r via*

$$P_r := [\mathcal{A}_r(\varphi_{ij}, \varphi_{st})]\,.$$

Then $P_r \sim A$ and the spectrum of $P_r^{-1}A$ satisfies

$$\Lambda(P_r^{-1}A) \subset [\theta_r, \Theta_r] \tag{3.7}$$

with θ_r, Θ_r defined in (3.6).

Remark 3.4. To obtain alternative spectral bounds under the assumption in (2.6), we can define (cf. (3.2))

$$\widetilde{\tau}_0 := 0, \quad \widetilde{\tau}_r := \left\| a_0^{-1} \sum_{m=1}^{r} |a_m| \right\|_\infty, \quad r \in \mathbb{N}.$$

Note that $(\widetilde{\tau}_r)_{r\in\mathbb{N}_0}$ is a monotonic increasing sequence bounded from above by $\widetilde{\tau}$. Then, instead of (3.5) we obtain by using (2.7)

$$\frac{1-\widetilde{\tau}_r}{1+\widetilde{\tau}}\, a(\boldsymbol{x}, \boldsymbol{y}) \le (1-\widetilde{\tau}_r) a_0(\boldsymbol{x}) \le a_r(\boldsymbol{x}, \boldsymbol{y}) \le (1+\widetilde{\tau}_r) a_0(\boldsymbol{x}) \le \frac{1+\widetilde{\tau}_r}{1-\widetilde{\tau}}\, a(\boldsymbol{x}, \boldsymbol{y}) \text{ a.e. in } \Omega\times\boldsymbol{\Gamma}.$$

This implies the equivalence of bilinear forms $\mathcal{A}$, $\mathcal{A}_0$, $\mathcal{A}_r$ as follows:

$$(1-\widetilde{\tau}_r)\mathcal{A}_0(v,v) \le \mathcal{A}_r(v,v) \le (1+\widetilde{\tau}_r)\mathcal{A}_0(v,v) \quad \forall\, v \in V$$

and

$$\frac{1-\widetilde{\tau}}{1+\widetilde{\tau}_r}\, \mathcal{A}_r(v,v) \le \mathcal{A}(v,v) \le \frac{1+\widetilde{\tau}}{1-\widetilde{\tau}_r}\, \mathcal{A}_r(v,v) \quad \forall\, v \in V.$$

Therefore, by Proposition 3.3, the following spectral bounds hold:

$$\Lambda(P_0^{-1}P_r) \subset [1-\widetilde{\tau}_r, 1+\widetilde{\tau}_r] \text{ and } \Lambda(P_r^{-1}A) \subset \left[\frac{1-\widetilde{\tau}}{1+\widetilde{\tau}_r}, \frac{1+\widetilde{\tau}}{1-\widetilde{\tau}_r}\right].$$

The preconditioners P_r, which we will refer to as *truncation preconditioners*, are induced by the bilinear form $\mathcal{A}_r$, $r \in \mathbb{N}_0$. Therefore, by (2.16), there holds $P_r = A$ for all $r \ge M$. In practice, the values of r are expected to be small in order to allow for sparse approximations of A which can be efficiently implemented. Note also that P_r

can be written as a sum of Kronecker products, just as was the case for the SGFEM matrix A (cf. (2.18)):

$$P_r = G_0 \otimes K_0 + \sum_{m=1}^{r} G_m \otimes K_m. \tag{3.8}$$

The result of Theorem 3.8 indicates that the performance of the preconditioned CG method will be independent of discretization parameters but may depend on the choice of truncation parameter r. Since

$$\mathcal{A}(v,v) = \mathcal{A}_r(v,v) + \mathcal{R}(v,v)$$

with

$$\mathcal{R}(v,v) := \int_{\boldsymbol{\Gamma}} p(\boldsymbol{y}) \int_{\Omega} (a(\boldsymbol{x},\boldsymbol{y}) - a_r(\boldsymbol{x},\boldsymbol{y})) \nabla v(\boldsymbol{x},\boldsymbol{y}) \cdot \nabla v(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

a smallness assumption of the form

$$|\mathcal{R}(v,v)| \leq \varepsilon_r \mathcal{A}(v,v), \quad 0 < \varepsilon_r < 1, \tag{3.9}$$

would allow for the following equivalence of $\mathcal{A}$ and $\mathcal{A}_r$:

$$(1-\varepsilon_r)\mathcal{A}(v,v) \leq \mathcal{A}_r(v,v) \leq (1+\varepsilon_r)\mathcal{A}(v,v). \tag{3.10}$$

Since

$$a(\boldsymbol{x},\boldsymbol{y}) - a_r(\boldsymbol{x},\boldsymbol{y}) = \sum_{m=r+1}^{\infty} a_m(\boldsymbol{x}) y_m,$$

by the definition of $\mathcal{R}$, assumption (3.9) holds for sufficiently large r. As a result, (3.10) implies the following eigenvalue bounds:

$$\Lambda(P_r^{-1} A) \subset \left[\frac{1}{1+\varepsilon_r}, \frac{1}{1-\varepsilon_r} \right].$$

This suggests that the closer a_r approximates a, the tighter the preconditioned spectrum will be clustered around unity. We will investigate this conclusion numerically in section 5.

4. Modified truncation preconditioners. Any practical implementation of a preconditioner requires an efficient technique for applying the action of its inverse on a given vector. Standard approaches include constructing sparse factorizations, or employing multigrid or multilevel techniques; domain decomposition methods represent yet another approach. The potential for parallelism could also be a deciding factor in the choice of solution method.

The preconditioner P_r introduced in the previous section is block-sparse, with sparsity deteriorating with increasing r (cf. Theorem 2.1). In the case when $r = 1$, the structure can be shown to be block-tridiagonal under a certain permutation—this is not an ideal situation, as it requires additional techniques to ensure an efficient application of the preconditioner. For this reason, we replace P_r with its corresponding symmetric block Gauss–Seidel (SBGS) approximation:

$$\widetilde{P}_r := \left(G_0 \otimes K_0 + \sum_{m=1}^{r} L_m \otimes K_m \right) (G_0 \otimes K_0)^{-1} \left(G_0 \otimes K_0 + \sum_{m=1}^{r} L_m^T \otimes K_m \right), \tag{4.1}$$

where $L_m + L_m^T = G_m$. The matrix $\widetilde{P}_r$ thus represents a sparse approximation to P_r involving block-triangular and block-diagonal matrices. In the remainder of this section we prove that $P_r \sim \widetilde{P}_r$ and provide complexity considerations, including a discussion of implementation.

4.1. Analysis of SBGS approximation of P_r. In this subsection we assume that the ordering of multi-indices in the index set $\mathbb{I}_k^M$ is such that the matrices L_m in (4.1) have at most one nonzero entry per row and per column (this property holds, e.g., for lexicographic or antilexicographic ordering as well as for ascending or descending ordering by the total degree of the associated complete polynomials in S_k^M). Let us define

$$S_r := \sum_{m=1}^{r} L_m \otimes K_m, \quad D_0 := G_0 \otimes K_0 \tag{4.2}$$

so that

$$P_r \overset{(3.8)}{=} D_0 + S_r + S_r^T \tag{4.3}$$

and

$$\widetilde{P}_r \overset{(4.1)}{=} (D_0 + S_r) D_0^{-1} (D_0 + S_r^T) = P_r + S_r D_0^{-1} S_r^T.$$

Our spectral analysis focuses on deriving bounds for the generalized Rayleigh quotient

$$\frac{\mathbf{v}^T \widetilde{P}_r \mathbf{v}}{\mathbf{v}^T P_r \mathbf{v}} = 1 + \frac{\mathbf{v}^T S_r D_0^{-1} S_r^T \mathbf{v}}{\mathbf{v}^T (D_0 + S_r + S_r^T)\mathbf{v}}, \qquad \mathbf{v} \in \mathbb{R}^{N_{\boldsymbol{x}} N_{\boldsymbol{y}}} \setminus \{\mathbf{0}\}. \tag{4.4}$$

Since the lower bound is 1, we restrict our attention to deriving an upper bound for the second term on the right-hand side of (4.4), which we write using the change of variable $\mathbf{w} = D_0^{1/2}\mathbf{v}$ as

$$\rho(\mathbf{w}) := \frac{\mathbf{w}^T \widetilde{S}_r \widetilde{S}_r^T \mathbf{w}}{\mathbf{w}^T (I + \widetilde{S}_r + \widetilde{S}_r^T)\mathbf{w}}.$$

Here,

$$\widetilde{S}_r := D_0^{-1/2} S_r D_0^{-1/2} = \sum_{m=1}^{r} L_m \otimes \widetilde{K}_m \tag{4.5}$$

with $\widetilde{K}_m := K_0^{-1/2} K_m K_0^{-1/2}$, using the fact that $G_0 = I_{N_{\boldsymbol{y}}}$. Hence,

$$\rho(\mathbf{w}) \le \max_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T \widetilde{S}_r \widetilde{S}_r^T \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \cdot \max_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T \mathbf{w}}{\mathbf{w}^T (I + \widetilde{S}_r + \widetilde{S}_r^T)\mathbf{w}} = \frac{\sigma_{\max}^2(\widetilde{S}_r)}{\lambda_{\min}(I + \widetilde{S}_r + \widetilde{S}_r^T)}, \tag{4.6}$$

where $\sigma_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote, respectively, the largest singular value and the smallest eigenvalue of a matrix. In the next lemma, we provide bounds for $\sigma_{\max}(\widetilde{S}_r)$ and $\lambda_{\min}(I + \widetilde{S}_r + \widetilde{S}_r^T)$ in order to conclude the derivation of the upper bound on ρ.

LEMMA 4.1. *Suppose that* (2.3) *and* (2.4) *hold and* τ_r *is defined in* (3.2). *Let* $\widetilde{S}_r$ *be defined by* (4.5). *Then*

$$\lambda_{\min}(I + \widetilde{S}_r + \widetilde{S}_r^T) \ge 1 - \tau_r \tag{4.7}$$

and

$$\sigma_{\max}(\widetilde{S}_r) \le \frac{1}{a_0^{\min}} \sum_{m=1}^{r} \|a_m\|_\infty. \tag{4.8}$$

Proof. Since

$$I+\widetilde{S}_r+\widetilde{S}_r^T = I+\sum_{m=1}^{r} G_m\otimes\widetilde{K}_m = D_0^{-1/2}\Big(D_0+\sum_{m=1}^{r} G_m\otimes K_m\Big)D_0^{-1/2} = D_0^{-1/2}P_r D_0^{-1/2},$$

the eigenvalues of $I + \widetilde{S}_r + \widetilde{S}_r^T$ are the eigenvalues of $D_0^{-1}P_r = P_0^{-1}P_r$. To find the bounds on the spectrum of $P_0^{-1}P_r$, recall the inequalities in (3.4) and the lower bound for $a_0(\boldsymbol{x})$ in (2.3), which together imply that

$$(1-\tau_r)a_0(\boldsymbol{x}) \le a_r(\boldsymbol{x},\boldsymbol{y}) \le (1+\tau_r)a_0(\boldsymbol{x}) \quad \text{a.e. in } \Omega\times\mathbf{\Gamma}.$$

Hence, the bilinear forms $\mathcal{A}_0$ and $\mathcal{A}_r$ are equivalent, and by Proposition 3.3 there holds

$$\Lambda(P_0^{-1}P_r) \subset [1-\tau_r, 1+\tau_r]\,. \tag{4.9}$$

This proves (4.7).

On the other hand, since $\sigma_{\max}$ defines a norm, we use the triangle inequality to estimate

$$\sigma_{\max}(\widetilde{S}_r) \le \sum_{m=1}^{r} \sigma_{\max}(L_m\otimes\widetilde{K}_m) \le \sum_{m=1}^{r} \sigma_{\max}(L_m)\sigma_{\max}(\widetilde{K}_m). \tag{4.10}$$

Now, $\sigma_{\max}^2(L_m) = \lambda_{\max}(L_m L_m^T)$; since $L_m L_m^T$ is diagonal for every m (due to L_m having at most one nonzero entry per row and per column), it follows that for all m

$$\sigma_{\max}(L_m) = \max_{i,j}[G_m]_{ij} \le \max_k c_k^m \le 1 \tag{4.11}$$

with c_k^m being bounded by 1 (cf. Theorem 2.1 and inequalities (2.11)). Finally, since the eigenvalues of $\widetilde{K}_m$ are the eigenvalues of $K_0^{-1}K_m$, we find that

$$\sigma_{\max}(\widetilde{K}_m) = \max_i \big|\lambda_i(K_0^{-1}K_m)\big| \le \frac{\|a_m\|_\infty}{a_0^{\min}},$$

and then inequality (4.8) follows from (4.10) and (4.11). This finishes the proof. ∎

We summarize our discussion in the following result.

PROPOSITION 4.2. *Suppose that* (2.3) *and* (2.4) *hold and* τ_r *is defined in* (3.2). *Let* P_r *be defined in* (3.8) *and let* $\widetilde{P}_r$ *be its SBGS approximation* (4.1). *Then* $\widetilde{P}_r \sim P_r$ *and the spectrum of* $P_r^{-1}\widetilde{P}_r$ *satisfies*

$$\Lambda(P_r^{-1}\widetilde{P}_r) \subset [1,\, 1+\delta_r], \tag{4.12}$$

where

$$\delta_r := \frac{1}{1-\tau_r}\left(\frac{1}{a_0^{\min}}\sum_{m=1}^{r}\|a_m\|_\infty\right)^2. \tag{4.13}$$

The proof of the upper bound in (4.12) is completed by substituting the estimates (4.7), (4.8) into (4.6) and then using the resulting bound for $\rho(\mathbf{w})$ in (4.4).

The following result is a straightforward consequence of Theorem 3.8, Proposition 4.2, and the transitivity of spectral equivalence.

THEOREM 4.3. *Let $\widetilde{P}_r$ be the SBGS approximation* (4.1) *to P_r. Let θ_r, Θ_r be defined in* (3.6)*, and let δ_r be defined in* (4.13)*. Then $\widetilde{P}_r \sim A$ and the spectrum of $\widetilde{P}_r^{-1}A$ satisfies*

$$\Lambda(\widetilde{P}_r^{-1}A) \subset \left[\frac{\theta_r}{1+\delta_r}, \Theta_r\right]. \tag{4.14}$$

4.2. Implementation. Complexity considerations. Our proposed solution method for solving the linear system (2.14) is the preconditioned CG (PCG) method, for which the main computational effort at each step comprises a matrix-vector product with the matrix A and the solution of a linear system with the preconditioning matrix. Since the main computational cost is associated with the latter operation, we discuss this in detail. We will denote by $\mathcal{H}(operation)$ the complexity, i.e., number of flops required to perform an *operation*. The number of nonzeros of a matrix will be denoted by $\mathrm{nnz}(\cdot)$.

As indicated previously, the action of the inverse of P_r needs to be approximated due to its sparse (but nondiagonal) block structure. We achieve this by replacing P_r with its SBGS approximation $\widetilde{P}_r$. Let us consider the implementation of the action of $\widetilde{P}_r^{-1}$ onto a given vector $\mathbf{v}$; for general $r \in \mathbb{N}_0$, this can be achieved as follows (see (4.2)–(4.3) for the definitions of the respective matrices):

1. solve $(D_0 + S_r)\mathbf{w} = \mathbf{v}$;
2. solve $(D_0 + S_r^T)\mathbf{z} = D_0\mathbf{w}$.

Both of the above steps involve the solution of a block-triangular system, with the main computational cost arising from solving linear systems with the diagonal blocks K_0. Specifically, since P_r has at most $2r+1$ nonzero block matrices per row, we find

$$\mathcal{H}(\widetilde{P}_r^{-1}\mathbf{v}) \approx (2rN_{\boldsymbol{y}})\mathrm{nnz}(K_0) + 2N_{\boldsymbol{y}}\mathcal{H}(K_0^{-1}\mathbf{b})$$

for some vector $\mathbf{b}$ of size $N_{\boldsymbol{x}}$. Below we consider two special cases.

4.2.1. Special case: $r = 0$. The preconditioner P_0 is the mean-based preconditioner introduced in [13]. The complexity associated with the action of the inverse on a given vector is

$$\mathcal{H}(P_0^{-1}\mathbf{v}) = N_{\boldsymbol{y}}\mathcal{H}(K_0^{-1}\mathbf{b}).$$

4.2.2. Special case: $r = 1$. The structure of P_r simplifies greatly when $r = 1$. In particular, G_1 has a block-diagonal structure under a certain permutation [21, section 9.5]:

$$G_1 = \mathrm{diag}(T_{k+1}, T_k, \ldots, T_k, \ldots, T_1, \ldots, T_1),$$

where $T_j \in \mathbb{R}^{j\times j}$ $(j = 1, \ldots, k+1)$ are tridiagonal, with zero main diagonal. Note in particular that $T_1 = 0$. As a result, P_1 will have a block-diagonal structure, where each diagonal block is a block-tridiagonal matrix of size $jN_{\boldsymbol{x}}$. Specifically,

$$P_1 = \bigoplus_{j=1}^{k+1}\bigoplus_{i=1}^{n_j}\left(I \otimes K_0 + T_j \otimes K_1\right),$$

where, assuming $M > 1$,

$$n_j = \binom{k+M-j-1}{M-2}.$$

Given this structure, the complexity for the implementation of the action of $\widetilde{P}_1^{-1}$ on a given vector $\mathbf{v}$ is

$$\mathcal{H}(\widetilde{P}_1^{-1}\mathbf{v}) \approx 2\left(\sum_{j=2}^{k+1} jn_j\right)\left(\mathrm{nnz}(K_0) + \mathcal{H}(K_0^{-1}\mathbf{b})\right) + n_1\mathcal{H}(K_0^{-1}\mathbf{b}).$$

Under the assumption that $\mathcal{H}(K_0^{-1}\mathbf{b})$ dominates the computation, we deduce that the implementation of $\widetilde{P}_1$ is at most twice as expensive as the implementation of P_0.

4.2.3. Kronecker preconditioner. We end this section with a discussion of the complexity required for an implementation of the Kronecker preconditioner [34]. Since

$$P_\otimes = G \otimes K_0 = (G \otimes I_{N_{\boldsymbol{x}}})(I_{N_{\boldsymbol{y}}} \otimes K_0) = (G \otimes I_{N_{\boldsymbol{x}}})P_0,$$

the complexity is given by

$$\mathcal{H}(P_\otimes^{-1}\mathbf{v}) = N_{\boldsymbol{y}}\mathcal{H}(K_0^{-1}\mathbf{b}) + N_{\boldsymbol{x}}\mathcal{H}(G^{-1}\mathbf{d})$$

for some vector $\mathbf{d}$ of size $N_{\boldsymbol{y}}$. Thus, the complexity exceeds that of P_0 by a computational cost dependent on the sparsity of G. In particular, it was estimated in [34] that this additional cost would amount to $\mathcal{H}(G^{-1}\mathbf{d}) \sim O((2M+1)^2)$ operations, excluding the cost of performing a Cholesky factorization of G.

5. Numerical experiments. In this section, we investigate the effectiveness of the preconditioning strategies considered in sections 3–4. In particular, we verify the theoretical optimality of truncation preconditioners P_r and $\widetilde{P}_r$ ($r \geq 1$) with respect to discretization parameters and compare their performance with that of the mean-based preconditioner P_0 and the Kronecker product preconditioner $P_\otimes$ defined in (1.4) (see [34] for details and analysis).

We chose to use test problems satisfying the descriptions and assumptions in this paper, as well as problems outside the theoretical framework. Thus, we solved model problem (1.1) using the following choices of parametric diffusion coefficient:

- $a(\boldsymbol{x}, \boldsymbol{y})$ has the affine representation (2.2), with the coefficients $a_m(\boldsymbol{x})$ satisfying (2.3) and (2.4);
- $a(\boldsymbol{x}, \boldsymbol{y})$ is a lognormal diffusion coefficient, i.e., $a(\boldsymbol{x}, \boldsymbol{y}) = \exp(b(\boldsymbol{x}, \boldsymbol{y}))$, where $b(\boldsymbol{x}, \boldsymbol{y})$ is assumed to have the affine representation (2.2) but with unbounded parameters y_m (we note that this choice of coefficient a is not covered by our theoretical analysis).

In all our tests, we chose $\Omega = (0,1)^2$ and $f(\boldsymbol{x}) = 1$. We used the MATLAB toolbox S-IFISS [29] to generate SGFEM discretizations of our model problem for a range of discretization parameters. We used uniform subdivisions of Ω into square elements of edge length h, with h ranging between 2^{-3} to 2^{-7}. The discretization parameters k, M had ranges $1 \leq k \leq 6$, $1 \leq M \leq 8$. We solved the resulting linear systems using the PCG method with tolerance $tol = 10^{-6}$ and zero initial guess.

5.1. Test problem 1: Affine-parametric diffusion coefficient. For this test problem, the diffusion coefficient $a(\boldsymbol{x}, \boldsymbol{y})$ had the affine-parametric form (2.2), with the coefficients $a_m(\boldsymbol{x})$ in the expansion chosen such that they exhibit either slow or fast decay. As indicated at the end of section 3, this is expected to affect the performance of the truncation preconditioners P_r. In particular, our numerical experiments will highlight the dependence on the truncation parameter r.

Following [7, section 11.1], we select the expansion coefficients $a_m(\boldsymbol{x})$ ($m \in \mathbb{N}_0$) in (2.2) to represent planar Fourier modes of increasing total order, i.e.,

$$a_0(\boldsymbol{x}) = 1, \ a_m(\boldsymbol{x}) = \bar{\alpha} m^{-\widetilde{\sigma}} \cos\left(2\pi\beta_1(m)x_1\right) \cos\left(2\pi\beta_2(m)x_2\right), \ \boldsymbol{x} = (x_1, x_2) \in \Omega. \tag{5.1}$$

Here, $\widetilde{\sigma} > 1$, $0 < \bar{\alpha} < 1/\zeta(\widetilde{\sigma})$, where ζ denotes the Riemann zeta function, and β_1, β_2 are given by

$$\beta_1(m) = m - k(m)\left(k(m)+1\right)/2, \qquad \beta_2(m) = k(m) - \beta_1(m)$$

with $k(m) = \lfloor -1/2 + \sqrt{1/4 + 2m} \rfloor$. Furthermore, we assume the parameters y_m ($m \in \mathbb{N}$) in (2.2) to be the images of independent uniformly distributed mean-zero random variables on $\Gamma_m = [-1, 1]$. In this case, $p_m(y_m) = 1/2$, $y_m \in \Gamma_m$, and the orthonormal polynomial basis of $L^2_{\pi_m}(\Gamma_m)$ consists of scaled Legendre polynomials. Note that with these settings, conditions (2.3) and (2.4) are satisfied with constants $a_0^{\min} = a_0^{\max} = 1$ and $\tau \leq \bar{\alpha}\zeta(\widetilde{\sigma})$, respectively.

The choice $\widetilde{\sigma} = 2$ in (5.1) yields coefficients a_m with slow decay of the amplitudes $\bar{\alpha} m^{-\widetilde{\sigma}}$, while the fast decay corresponds to the choice $\widetilde{\sigma} = 4$. In each case, we select the factor $\bar{\alpha}$ such that $\bar{\alpha}\zeta(\widetilde{\sigma}) = 0.9999$, which gives $\bar{\alpha} \approx 0.6079$ for $\widetilde{\sigma} = 2$ and $\bar{\alpha} \approx 0.9239$ for $\widetilde{\sigma} = 4$. The magnitudes of the expansion coefficients a_m for increasing m for slow and fast decay are displayed in Table 5.1.

While the magnitudes of a_m ($m = 1, 2, \ldots$) (and hence, the importance of the corresponding parameters y_m, $m = 1, 2, \ldots$) decay significantly faster for $\widetilde{\sigma} = 4$ (e.g., $\|a_1\|_\infty \approx 16\|a_2\|_\infty$ for $\widetilde{\sigma} = 4$, whereas $\|a_1\|_\infty \approx 4\|a_2\|_\infty$ for $\widetilde{\sigma} = 2$), we observe that the magnitude of a_1 is much closer to the magnitude of the mean field a_0 in the case of fast decay than in the case of slow decay. This suggests that for fast decay there holds $a(\boldsymbol{x}, \boldsymbol{y}) \approx a_1(\boldsymbol{x}, \boldsymbol{y})$, which in turn implies that equivalence (3.10) may hold for $r = 1$ and with a small ε_1. We therefore expect the performance of P_1 to be superior in the fast decay case. This is indeed confirmed by the iteration counts in Table 5.2. The results also confirm the optimal performance of P_r with respect to k.

A similar behavior can be observed also for the case where the preconditioners P_r are replaced by their SBGS approximations $\widetilde{P}_r$. Table 5.3 diplays the corresponding iteration counts for a range of r, as well as for the mean-based preconditioner P_0 and Kronecker preconditioner $P_\otimes$, for both fast and slow decay cases.

The results in Tables 5.2 and 5.3 indicate that the iteration counts corresponding to the approximations $\widetilde{P}_r$ of the preconditioners P_r are higher. This is expected given the theoretical deterioration of the spectral bounds in (4.14) as compared to

TABLE 5.1
Magnitudes $\|a_m\|_\infty$ *of expansion coefficients* (5.1) *for test problem* 1.

m	0	1	2	3	4	5	6
Slow decay ($\widetilde{\sigma} = 2$)	1.0000	0.6079	0.1520	0.0675	0.0380	0.0243	0.0169
Fast decay ($\widetilde{\sigma} = 4$)	1.0000	0.9239	0.0577	0.0114	0.0036	0.0015	0.0007

TABLE 5.2
PCG iterations counts for test problem 1; $h = 2^{-4}, M = 8$.

k	Fast decay							Slow decay						
	P_0	P_1	P_2	P_3	P_4	P_5	P_6	P_0	P_1	P_2	P_3	P_4	P_5	P_6
1	13	4	3	3	2	2	2	10	6	4	4	4	3	3
2	16	5	4	3	3	2	2	12	7	5	5	4	4	3
3	21	6	4	3	3	2	2	14	7	6	5	4	4	4
4	24	6	4	3	3	3	2	15	8	6	5	4	4	4

the bounds in (3.7). However, all truncation preconditioners require fewer iterations than their mean-based and Kronecker product counterparts. This improvement in the iteration counts is more pronounced in the case of fast decay than in the case of slow decay, which is consistent, in particular, with how the magnitudes of expansion coefficients a_m change with m (see Table 5.1 and the associated discussion above). For example, the numbers of iterations for the truncation preconditioner P_1 (resp., $\widetilde{P}_1$) are less than those for the mean-based preconditioner P_0 by factors of 3 to 4 (resp., by factors of about 2 to 2.5) in the case of fast decay. This is because in this case, the expansion coefficient a_1 has approximately the same magnitude as the mean field. In the case of slow decay, however, both P_1 and $\widetilde{P}_1$ outperform P_0 in terms of the number of iterations only by a factor between 1.5 and 2. It is worth recalling here that the computational cost for the truncation preconditioner $\widetilde{P}_1$ is about twice the cost of the mean-based preconditioner P_0 (see subsection 4.2); thus, in terms of the overall computational complexity, $\widetilde{P}_1$ performs at least the same as P_0; in the fast decay case, the overall computational cost for modified truncation preconditioners is significantly lower than that for P_0.

If more expansion coefficients are retained in P_r ($r \geq 2$), then the iteration counts naturally (and consistently) decrease; in particular, they decrease quicker in the case of fast decay of coefficient amplitudes than in the case of slow decay of the amplitudes; see Table 5.2. This is again in agreement with what one might expect and reflects different decay patterns of $\|a_m\|_\infty$ in each of these cases, as shown in Table 5.1. However, when applying the corresponding modified truncation preconditioners $\widetilde{P}_r$ ($r \geq 2$) and increasing the number r of retained expansion coefficients, the iteration counts decrease very slowly (in the case of fast decay they even stagnate for $r \geq 2$ in most cases); see Table 5.3. This indicates that no significant improvement is obtained by including additional terms a_m in the definition of P_r.

The above set of experiments demonstrates that the modified truncation preconditioners $\widetilde{P}_r$ provide sufficiently accurate approximations of stochastic Galerkin matrices and thus can be used as effective practical preconditioners for linear systems arising from SGFEM approximations of the model problem (1.1) with *affine-parametric*

TABLE 5.3
PCG iterations counts for test problem 1; $h = 2^{-4}, M = 8$.

k	Fast decay								Slow decay							
	$P_\otimes$	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$	$\widetilde{P}_3$	$\widetilde{P}_4$	$\widetilde{P}_5$	$\widetilde{P}_6$	$P_\otimes$	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$	$\widetilde{P}_3$	$\widetilde{P}_4$	$\widetilde{P}_5$	$\widetilde{P}_6$
1	12	13	7	6	6	6	6	6	9	10	6	5	5	5	5	5
2	16	16	8	7	7	7	7	7	12	12	7	6	6	6	5	5
3	20	21	9	9	8	8	8	8	14	14	8	7	6	6	6	6
4	24	24	10	9	9	9	9	9	15	15	9	7	7	6	6	6
5	26	27	11	10	10	10	10	10	16	16	9	7	7	7	6	6
6	29	29	12	11	11	11	11	11	17	17	10	8	7	7	7	7

TABLE 5.4
PCG iterations counts for test problem 1 for various h and M, for $k = 3$.

h	Fast decay						Slow decay					
	$M = 4$			$M = 8$			$M = 4$			$M = 8$		
	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$
2^{-3}	18	8	8	18	8	8	13	7	6	13	7	6
2^{-4}	21	9	9	21	9	9	14	8	7	14	8	7
2^{-5}	23	10	9	23	10	9	14	8	7	15	8	7
2^{-6}	24	10	10	24	10	10	15	8	7	15	8	7
2^{-7}	24	10	10	24	10	10	15	8	7	15	8	7

representation of the diffusion coefficient. Depending on the decay of magnitudes of the expansion coefficients, one may choose larger values of r to improve the efficiency of the solver (e.g., in the case of slow decay). However, in most cases, we recommend to choose $r = 1$ or $r = 2$.

We end the discussion of our first test problem with a numerical confirmation of optimality of the modified truncation preconditioners with respect to discretization parameters h and M. We consider again the cases of fast ($\widetilde{\sigma} = 4$) and slow ($\widetilde{\sigma} = 2$) decay of coefficient amplitudes $\bar{\alpha} m^{-\widetilde{\sigma}}$ in (5.1) and employ three preconditioners: the mean-based P_0 and the modified truncation preconditioners $\widetilde{P}_1$ and $\widetilde{P}_2$. Note that other modified truncation preconditioners (for $r > 2$) yield similar performance, and the corresponding results are not included here. We chose to work with two values of $M \in \{4, 8\}$ and several uniform subdivisions into squares of side lengths ranging from $h = 2^{-3}$ to $h = 2^{-7}$, while keeping fixed the polynomial degree $k = 3$. The results of these computations are presented in Table 5.4. These results show that indeed, the iteration counts do not change as M increases from 4 to 8 (for the same value of h) and as the spatial mesh gets sufficiently refined (for the same value of M).

5.2. Test problem 2: Nonaffine parametric diffusion coefficient. Consider again model problem (1.1), now with the following truncated *lognormal* diffusion coefficient

$$a(\boldsymbol{x}, \boldsymbol{y}) := \exp\big(b(\boldsymbol{x}, \boldsymbol{y})\big) := \exp\bigg(b_0(\boldsymbol{x}) + \sum_{m=1}^{N} b_m(\boldsymbol{x}) y_m\bigg), \ \boldsymbol{x} \in \Omega, \ \boldsymbol{y} \in \boldsymbol{\Gamma} := \prod_{m=1}^{N} \Gamma_m, \tag{5.2}$$

where $b_0, b_m \in L^\infty(\Omega)$ for all $m = 1, \ldots, N$ and the parameters $y_m \in \Gamma_m := \mathbb{R}$ are the images of independent normally distributed random variables with zero mean and unit variance. Accordingly, p_m now denotes the standard Gaussian probability density function, and the joint probability density function is $p(\boldsymbol{y}) = \prod_{m=1}^{N} p_m(y_m)$. The well-posedness of weak formulation (2.8) in this case has been studied in [5].

As a polynomial basis of $L^2_\pi(\boldsymbol{\Gamma})$ we choose the set of scaled Hermite polynomials $\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\Gamma} \to \mathbb{R} : \boldsymbol{\alpha} \in \mathbb{N}_0^N\}$ orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$. In this basis, the diffusion coefficient (5.2) has the representation

$$a(\boldsymbol{x}, \boldsymbol{y}) := \sum_{\boldsymbol{\alpha} \in \mathbb{N}_0^N} a_{\boldsymbol{\alpha}}(\boldsymbol{x}) \psi_{\boldsymbol{\alpha}}(\boldsymbol{y})$$

with (cf. [34, page 926])

$$a_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \langle a(\boldsymbol{x}, \cdot), \psi_{\boldsymbol{\alpha}} \rangle_\pi = \mathbb{E}[a(\boldsymbol{x}, \cdot)] \prod_{m \in \operatorname{supp} \boldsymbol{\alpha}} \frac{b_m^{\alpha_m}(\boldsymbol{x})}{\sqrt{\alpha_m}} \quad \forall\, \boldsymbol{\alpha} \in \mathbb{N}_0^N, \tag{5.3}$$

where

$$\mathbb{E}[a(\boldsymbol{x},\cdot)] = \int_{\boldsymbol{\Gamma}} \exp\left(b(\boldsymbol{x},\boldsymbol{y})\right) p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = \exp\left(b_0(\boldsymbol{x}) + \frac{1}{2}\sum_{m=1}^{N} b_m^2(\boldsymbol{x})\right) > 0 \quad \text{a.e. in } \Omega.$$

Let $M < N$. A Galerkin projection onto $X_h \otimes S_k^M$ (cf. (2.12)–(2.13)) yields a linear system with coefficient matrix

$$A = \sum_{\boldsymbol{\alpha} \in \mathbb{I}_{2k}^M} G_{\boldsymbol{\alpha}} \otimes K_{\boldsymbol{\alpha}}, \tag{5.4}$$

where for all $\boldsymbol{\alpha} \in \mathbb{I}_{2k}^M$ we have defined

$$[G_{\boldsymbol{\alpha}}]_{tj} := \left\langle \psi_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\kappa}(j)}, \psi_{\boldsymbol{\kappa}(t)} \right\rangle_{\pi} \qquad \text{for } t,j = 1,\ldots,N_{\boldsymbol{y}},$$

$$[K_{\boldsymbol{\alpha}}]_{si} := \int_{\Omega} a_{\boldsymbol{\alpha}} \nabla\phi_i \cdot \nabla\phi_s \,\mathrm{d}\boldsymbol{x} \qquad \text{for } s,i = 1,\ldots,N_{\boldsymbol{x}}$$

with $a_{\boldsymbol{\alpha}}$ given by (5.3). Note that all matrices $G_{\boldsymbol{\alpha}}$ have nonnegative entries; cf. [10, Appendix A]. In particular, $G_{\mathbf{0}} = I$, where $\mathbf{0} = (0,0,\ldots,0) \in \mathbb{I}_{2k}^M$. Furthermore, the well-posedness of weak formulation implies that the matrix A is positive definite.

In order to define a family of truncation preconditioners for A, we introduce an ordering of the terms in (5.4) based on the magnitude of $a_{\boldsymbol{\alpha}}$. To that end, we assume, without loss of generality, that all magnitudes $\|a_{\boldsymbol{\alpha}}\|_{\infty}$ for $\boldsymbol{\alpha} \in \mathbb{I}_{2k}^M$ are distinct. Then, the ordering is defined by the sequence $\left\{\boldsymbol{\alpha}_\ell : \ell = 0,1,\ldots,\operatorname{card}\mathbb{I}_{2k}^M - 1\right\}$ of all multi-indices in $\mathbb{I}_{2k}^M$ such that

$$\|a_{\boldsymbol{\alpha}_i}\|_{\infty} > \left\|a_{\boldsymbol{\alpha}_j}\right\|_{\infty}, \quad i < j.$$

Unlike in the affine case, this ordering is not sufficient to ensure positivity of the truncated diffusion coefficient

$$a_r(\boldsymbol{x},\boldsymbol{y}) := \sum_{\ell=0}^{r} a_{\boldsymbol{\alpha}_\ell}(\boldsymbol{x}) \psi_{\boldsymbol{\alpha}_\ell}(\boldsymbol{y}), \quad 0 \le r \le \operatorname{card}\mathbb{I}_{2k}^M - 1. \tag{5.5}$$

Consequently, the resulting truncation preconditioner

$$P_r := \sum_{\ell=0}^{r} G_{\boldsymbol{\alpha}_\ell} \otimes K_{\boldsymbol{\alpha}_\ell} \tag{5.6}$$

is not guaranteed to be positive definite. However, as in the affine case, we replace P_r by its SBGS approximation $\widetilde{P}_r$. As demonstrated in Proposition 5.1 below, the modified truncation preconditioner $\widetilde{P}_r$ is positive definite, provided that the mean field $a_{\mathbf{0}}(\boldsymbol{x}) = \mathbb{E}[a(\boldsymbol{x},\cdot)]$ is included in the truncation (5.5). It is important to note that the complexity associated with the action of $\widetilde{P}_r^{-1}$ remains unchanged from the affine case; cf. subsection 4.2.

PROPOSITION 5.1. *Let a be the lognormal diffusion coefficient given by* (5.2). *Let $0 \le r \le \operatorname{card}\mathbb{I}_{2k}^M - 1$, and assume that the truncated diffusion coefficient a_r in* (5.5) *satisfies $\mathbb{E}[a_r] = \mathbb{E}[a] = a_{\mathbf{0}} > 0$ a.e. in Ω. Then the truncation preconditioner P_r given by* (5.6) *can be represented as*

$$P_r = D + L + L^T, \tag{5.7}$$

where D is a block-diagonal symmetric positive definite matrix and L is a strictly lower block-triangular matrix. As a consequence, the SBGS approximation of P_r defined by

$$\widetilde{P}_r := (D + L)\, D^{-1}(D + L^T)$$

is a symmetric positive definite matrix.

Proof. Denote by $\mathbb{I}_{\text{even}}$ the set of multi-indices in $\mathbb{I}_{2k}^M$ with even entries, i.e.,

$$\mathbb{I}_{\text{even}} := \left\{ \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M) \in \mathbb{I}_{2k}^M;\ \alpha_m \text{ is even for all } m = 1, \ldots, M \right\}.$$

Let $\boldsymbol{\alpha} \in \mathbb{I}_{\text{even}}$. It follows from (5.3) that $a_{\boldsymbol{\alpha}}(\boldsymbol{x}) \geq 0$ a.e. in Ω. Therefore, the stiffness matrices $K_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \mathbb{I}_{\text{even}}$ are symmetric positive semidefinite; in particular, the matrix $K_{\mathbf{0}}$ is symmetric positive definite. Since all diagonal elements of the matrix $G_{\boldsymbol{\alpha}}$ are nonnegative, each nonzero diagonal block of $G_{\boldsymbol{\alpha}} \otimes K_{\boldsymbol{\alpha}}$ is a symmetric positive semidefinite matrix. In particular, the block-diagonal matrix $G_{\mathbf{0}} \otimes K_{\mathbf{0}} = I \otimes K_{\mathbf{0}}$ is symmetric positive definite.

Now let $\boldsymbol{\alpha} \in \mathbb{I}_{2k}^M \setminus \mathbb{I}_{\text{even}}$. In this case, all diagonal elements of the matrix $G_{\boldsymbol{\alpha}}$ are zeros. Indeed, for any $j = 1, \ldots, N_{\boldsymbol{y}}$, one has

$$[G_{\boldsymbol{\alpha}}]_{jj} = \left\langle \psi_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\kappa}(j)}, \psi_{\boldsymbol{\kappa}(j)} \right\rangle_\pi = \left\langle \psi_{\boldsymbol{\alpha}}, \psi_{\boldsymbol{\kappa}(j)}^2 \right\rangle_\pi = 0,$$

because there exists $m^* \in \{1, \ldots, M\}$ such that α_{m^*} is odd and the associated univariate Hermite polynomial is an odd function. Therefore, in this case, all diagonal blocks of $G_{\boldsymbol{\alpha}} \otimes K_{\boldsymbol{\alpha}}$ are zero matrices.

Overall, by combining the above observations and using the assumption that $\mathbb{E}[a_r] = \mathbb{E}[a]$, we conclude that the diagonal blocks of the truncation preconditioner P_r in (5.6) are symmetric positive definite matrices. This proves (5.7).

It is now easy to see that the SBGS approximation of P_r is a positive definite matrix. Indeed, for any nonzero vector $\mathbf{v}$ there holds

$$\mathbf{v}^T \widetilde{P}_r \mathbf{v} = \mathbf{v}^T (D + L)\, D^{-1}(D + L^T)\mathbf{v} = \mathbf{w}^T D^{-1} \mathbf{w} > 0$$

with nonzero $\mathbf{w} := (D + L^T)\mathbf{v}$. ∎

In numerical experiments, we set $N = 20$ and chose $b_m(\boldsymbol{x})$ in (5.2) to be the coefficients $a_m(\boldsymbol{x})$ in test problem 1 as defined in (5.1) with $\widetilde{\sigma} = 2$ and $\bar{\alpha} = 0.547$. In Table 5.5, for $M = k = 6$, we show first eight multi-indices in the sequence $\{\boldsymbol{\alpha}_\ell\}$ and the corresponding coefficient magnitudes $\|a_{\boldsymbol{\alpha}_\ell}\|_\infty$. We see that in this example,

TABLE 5.5
Multi-indices of first 8 largest magnitudes $\|a_{\boldsymbol{\alpha}}\|_\infty$ for test problem 2; $M = k = 6$.

ℓ	$\boldsymbol{\alpha}_\ell$	$\|a_{\boldsymbol{\alpha}_\ell}\|_\infty$
0	(0,0,0,0,0,0)	3.20
1	(1,0,0,0,0,0)	1.75
2	(2,0,0,0,0,0)	0.68
3	(0,1,0,0,0,0)	0.44
4	(1,1,0,0,0,0)	0.24
5	(3,0,0,0,0,0)	0.21
6	(0,0,1,0,0,0)	0.19
7	(0,0,0,1,0,0)	0.11

TABLE 5.6
PCG iterations counts for test problem 2; $h = 2^{-4}$, $M = 6$.

k	$P_\otimes$	P_0	$\widetilde{P}_1$	$\widetilde{P}_2$	$\widetilde{P}_3$	$\widetilde{P}_4$	$\widetilde{P}_5$	$\widetilde{P}_6$
1	12	12	6	7	6	6	6	6
2	18	19	8	10	9	9	8	8
3	25	26	**10**	12	11	11	10	10
4	32	34	**13**	15	13	13	12	11
5	40	43	**17**	19	16	17	**13**	**12**
6	49	52	**24**	22	19	20	**14**	**14**

the coefficient with the largest magnitude is the mean field, i.e., $a_{\boldsymbol{\alpha}_0} = a_{\mathbf{0}}$. While the distribution of indices inducing the ordering does not display any obvious pattern, we note a fast decay with ℓ in the magnitudes recorded, which is similar to the affine case.

Table 5.6 displays the PCG iteration counts corresponding to solving linear systems arising from SGFEM discretizations of the described test problem. We used the following discretization parameters: $h = 2^{-4}$, $M = 6$, and $k \in \{1, \dots, 6\}$. In our experiments, we employed the modified truncation preconditioners $\widetilde{P}_r$ with $r \in \{1, \dots, 6\}$, alongside the mean-based (P_0) and Kronecker ($P_\otimes$) preconditioners. These experiments included cases where preconditioners P_r defined by (5.6) were not positive definite (in Table 5.6, the iteration counts for such cases are shown in boldface).

The results in Table 5.6 indicate that the numbers of iterations by the modified truncation preconditioners are significantly lower than those corresponding to the mean-based and Kronecker preconditioners (it is worth noting here that while the computational cost for P_0 and $\widetilde{P}_r$ remains unchanged from the affine case, the cost for $P_\otimes$ in this test problem will be significantly higher than in the affine case, due to the density of the matrix G in (1.4) for the lognormal diffusion coefficient). For all preconditioners, the experiments show that the iteration counts grow with k, although this growth is much less pronounced for truncation preconditioners. Furthermore, while we see only a negligible improvement with increasing r for $k = 1, \dots, 4$, this becomes more pronounced for higher polynomial degrees ($k = 5, 6$).

6. Summary and future work. Efficient solution of large coupled linear systems is a key ingredient in successful implementation of the SGFEM. Truncation preconditioners represent a competitive alternative to existing solvers relying on the mean-based and Kronecker preconditioners. Our theoretical analysis shows that for elliptic problems with *affine-parametric* coefficients, truncation preconditioners are optimal with respect to discretization parameters. Our numerical experiments confirm this, while also demonstrating the improvement in the iteration count when compared with the mean-based and Kronecker preconditioners.

On a practical note, the superior efficiency of considered solvers requires, crucially, suitable fast (possibly parallel) implementation of the corresponding SBGS approximations, which were also analyzed and shown to be optimal. For simplicity, we considered a model diffusion problem, however, the analysis included in this work can be extended in a straightforward manner to the general case of elliptic PDE with parametric or uncertain inputs, under standard assumptions.

We have also applied truncation preconditioners in the case of *nonaffine* (specifically, lognormal) diffusion coefficient. The numerical experiments suggest this is a promising approach. Theoretical analysis of truncation preconditioners for this class of parametric problems will be the focus of future research on the topic.

REFERENCES

[1] I. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825, https://doi.org/10.1137/S0036142902418680.

[2] A. Bespalov, C. E. Powell, and D. Silvester, *Energy norm a posteriori error estimation for parametric operator equations*, SIAM J. Sci. Comput., 36 (2014), pp. A339–A363, https://doi.org/10.1137/130916849.

[3] A. Bespalov, D. Praetorius, L. Rocchi, and M. Ruggeri, *Convergence of adaptive stochastic Galerkin FEM*, SIAM J. Numer. Anal., 57 (2019), pp. 2359–2382, https://doi.org/10.1137/18M1229560.

[4] M. Brezina, A. Doostan, T. Manteuffel, S. McCormick, and J. Ruge, *Smoothed aggregation algebraic multigrid for stochastic PDE problems with layered materials*, Numer. Linear Algebra Appl., 21 (2014), pp. 239–255, https://doi.org/10.1002/nla.1924.

[5] J. Charrier, *Strong and weak error estimates for elliptic partial differential equations with random coefficients*, SIAM J. Numer. Anal., 50 (2012), pp. 216–246.

[6] Y. G. D'yakonov, *The construction of iterative methods based on the use of spectrally equivalent operators*, Comput. Math. Math. Phys., 6 (1966), pp. 14–46, https://doi.org/10.1016/0041-5553(66)90030-9.

[7] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, *Adaptive stochastic Galerkin FEM*, Comput. Methods Appl. Mech. Engrg., 270 (2014), pp. 247–269, https://doi.org/10.1016/j.cma.2013.11.015.

[8] H. Elman and D. Furnival, *Solving the stochastic steady-state diffusion problem using multigrid*, IMA J. Numer. Anal., 27 (2007), pp. 675–688, https://doi.org/10.1093/imanum/drm006.

[9] H. C. Elman and T. Su, *A low-rank multigrid method for the stochastic steady-state diffusion problem*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 492–509, https://doi.org/10.1137/17M1125170.

[10] O. G. Ernst and E. Ullmann, *Stochastic Galerkin matrices*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1848–1872.

[11] V. Faber, T. A. Manteuffel, and S. V. Parter, *On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations*, Adv. Appl. Math., 11 (1990), pp. 109–163.

[12] W. Gautschi, *Orthogonal Polynomials: Computation and Approximation*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2004.

[13] R. G. Ghanem and R. M. Kruger, *Numerical solution of spectral stochastic finite element systems*, Comput. Methods Appl. Mech. Engrg., 129 (1996), pp. 289–303, https://doi.org/10.1016/0045-7825(95)00909-4.

[14] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991, https://doi.org/10.1007/978-1-4612-3094-6.

[15] M. D. Gunzburger, C. G. Webster, and G. Zhang, *Stochastic finite element methods for partial differential equations with random input data*, Acta Numer., 23 (2014), pp. 521–650, https://doi.org/10.1017/S0962492914000075.

[16] C. Jin and X.-C. Cai, *A preconditioned recycling GMRES solver for stochastic Helmholtz problems*, Commun. Comput. Phys., 6 (2009), pp. 342–353, https://doi.org/10.4208/cicp.2009.v6.p342.

[17] M. Kubínová and I. Pultarová, *Block preconditioning of stochastic Galerkin problems: New two-sided guaranteed spectral bounds*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 88–113, https://doi.org/10.1137/19M125902X.

[18] O. P. Le Maître, O. M. Knio, B. J. Debusschere, H. N. Najm, and R. G. Ghanem, *A multigrid solver for two-dimensional stochastic diffusion equations*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 4723–4744, https://doi.org/10.1016/S0045-7825(03)00457-2.

[19] B. Lee, *Parallel preconditioners and multigrid solvers for stochastic polynomial chaos discretizations of the diffusion equation at the large scale*, Numer. Linear Algebra Appl., 23 (2016), pp. 5–36, https://doi.org/10.1002/nla.2000.

[20] D. Loghin and A. J. Wathen, *Analysis of preconditioners for saddle-point problems*, SIAM J. Sci. Comp., 25 (2004), pp. 2029–2049, https://doi.org/10.1137/S1064827502418203.

[21] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge Texts Appl. Math., Cambridge University Press, New York, 2014, https://doi.org/10.1017/CBO9781139017329.

[22] M. F. Pellissetti and R. G. Ghanem, *Iterative solution of systems of linear equations arising in the context of stochastic finite elements*, Adv. Eng. Softw., 31 (2000), pp. 607–616, https://doi.org/10.1016/s0965-9978(00)00034-x.

[23] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA J. Numer. Anal., 29 (2009), pp. 350–375, https://doi.org/10.1093/imanum/drn014.

[24] I. Pultarová, *Block and multilevel preconditioning for stochastic Galerkin problems with lognormally distributed parameters and tensor product polynomials*, Int. J. Uncertain. Quantif., 7 (2017), pp. 441–462, https://doi.org/10.1615/Int.J.UncertaintyQuantification.2017020377.

[25] E. Rosseel, T. Boonen, and S. Vandewalle, *Algebraic multigrid for stationary and time-dependent partial differential equations with stochastic coefficients*, Numer. Linear Algebra Appl., 15 (2008), pp. 141–163, https://doi.org/10.1002/nla.568.

[26] E. Rosseel and S. Vandewalle, *Iterative solvers for the stochastic finite element method*, SIAM J. Sci. Comput., 32 (2010), pp. 372–397, https://doi.org/10.1137/080727026.

[27] A. Sarkar, N. Benabbou, and R. Ghanem, *Domain decomposition of stochastic PDEs: Theoretical formulations*, Internat. J. Numer. Methods Engrg., 77 (2009), pp. 689–701, https://doi.org/10.1002/nme.2431.

[28] C. Schwab and C. J. Gittelson, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, Acta Numer., 20 (2011), pp. 291–467, https://doi.org/10.1017/S0962492911000055.

[29] D. J. Silvester, A. Bespalov, and C. E. Powell, *Stochastic IFISS (S-IFISS), version* 1.04, October 2017, http://www.manchester.ac.uk/ifiss/sifiss.html.

[30] B. Sousedík and R. G. Ghanem, *Truncated hierarchical preconditioning for the stochastic Galerkin FEM*, Int. J. Uncertain. Quantif., 4 (2014), pp. 333–348, https://doi.org/10.1615/Int.J.UncertaintyQuantification.2014007353.

[31] B. Sousedík, R. G. Ghanem, and E. T. Phipps, *Hierarchical Schur complement preconditioner for the stochastic Galerkin finite element methods*, Numer. Linear Algebra Appl., 21 (2014), pp. 136–151, https://doi.org/10.1002/nla.1869.

[32] W. Subber and S. Loisel, *Schwarz preconditioners for stochastic elliptic PDEs*, Comput. Methods Appl. Mech. Engrg., 272 (2014), pp. 34–57, https://doi.org/10.1016/j.cma.2013.12.016.

[33] W. Subber and A. Sarkar, *A domain decomposition method of stochastic PDEs: An iterative solution techniques using a two-level scalable preconditioner*, J. Comput. Phys., 257 (2014), pp. 298–317, https://doi.org/10.1016/j.jcp.2013.08.058.

[34] E. Ullmann, *A Kronecker product preconditioner for stochastic Galerkin finite element discretizations*, SIAM J. Sci. Comput., 32 (2010), pp. 923–946, https://doi.org/10.1137/080742853.