ORNL REPORT

ORNL/TM-2015/328
Unlimited Release
Printed July 2015

Explicit cost bounds of stochastic Galerkin approximations for parameterized PDEs with random coefficients

N. C. Dexter, C. G. Webster, and G. Zhang

Prepared by
Department of Computational and Applied Mathematics
Computer Science and Mathematics Division
Oak Ridge National Laboratory
One Bethel Valley Road, Oak Ridge, Tennessee 37831

The Oak Ridge National Laboratory is operated by UT-Battelle, LLC,
for the United States Department of Energy under Contract DE-AC05-00OR22725.
Approved for public release; further dissemination unlimited.

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via the U.S. Department of Energy (DOE) Information Bridge.

Web site http://www.osti.gov/bridge

Reports produced before January 1, 1996, may be purchased by members of the public from the following source.

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Oak Ridge, TN 37831
Telephone 703-605-6000 (1-800-553-6847)
TDD 703-487-4639
Fax 703-605-6900
E-mail info@ntis.gov
Web site http://www.ntis.gov/support/ordernowabout.htm

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and International Nuclear Information System (INIS) representatives from the following source.

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Telephone 865-576-8401
Fax 865-576-5728
E-mail reports@osti.gov
Web site http://www.osti.gov/contact.html

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

ORNL/TM-2015/328
Date Published: July 2015

EXPLICIT COST BOUNDS OF STOCHASTIC GALERKIN APPROXIMATIONS FOR PARAMETERIZED PDES WITH RANDOM COEFFICIENTS

Nick Dexter * Clayton G. Webster † Guannan Zhang ‡

Abstract. This work analyzes the overall computational complexity of the stochastic Galerkin finite element method (SGFEM) for approximating the solution of parameterized elliptic partial differential equations with both affine and non-affine random coefficients. To compute the fully discrete solution, such approaches employ a Galerkin projection in both the deterministic and stochastic domains, produced here by a combination of finite elements and a global orthogonal basis, defined on an isotopic total degree index set, respectively. To account for the sparsity of the resulting system, we present a rigorous cost analysis that considers the total number of coupled finite element systems that must be simultaneously solved in the SGFEM. However, to maintain sparsity as the coefficient becomes increasingly nonlinear in the parameterization, it is necessary to also approximate the coefficient by an additional orthogonal expansion. In this case we prove a rigorous complexity estimate for the number of floating point operations (FLOPs) required per matrix-vector multiplication of the coupled system. Based on such complexity estimates we also develop explicit cost bounds in terms of FLOPs to solve the stochastic Galerkin (SG) systems to a prescribed tolerance, which are used to compare with the minimal complexity estimates of a stochastic collocation finite element method (SCFEM), shown in our previous work [16]. Finally, computational evidence complements the theoretical estimates and supports our conclusion that, in the case that the coefficient is affine, the coupled SG system can be solved more efficiently than the decoupled SC systems. However, as the coefficient becomes more nonlinear, it becomes prohibitively expensive to obtain an approximation with the SGFEM.

Key words. stochastic Galerkin, stochastic collocation, sparse polynomial approximation, complexity analysis, explicit cost bounds, finite elements

1. Introduction.

Nowadays, stochastic polynomial methods are widely used alternatives to Monte Carlo methods (see, e.g., [15]) for predicting the solution to physical and engineering problems described by parameterized partial differential equations (PDEs) with a finite number of random variables. In the last decade, two classes of such methods have been proposed that often feature much faster convergence rates: *intrusive* stochastic Galerkin (SG) methods and *non-intrusive* stochastic collocation (SC) methods. Both approaches typically employ a Galerkin projection in the physical domain, produced here by finite elements, and the resulting fully discrete approximations only differ in their choice of multivariate polynomials for the discretization in the stochastic domain. For details about the relations between these methods see [17–19, 21, 24], and for computational comparisons between the SG and SC methods see, e.g., [3, 13, 19].

*Department of Mathematics, 227 Ayres Hall. 1403 Circle Drive. University of Tennessee, Knoxville, TN 37996-1320 (ndexter@utk.edu)

†Department of Computational and Applied Mathematics, Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, MS-6164, Oak Ridge, TN 37831-6164 (webstercg@ornl.gov)

‡Department of Computational and Applied Mathematics, Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008, MS-6164, Oak Ridge, TN 37831-6164 (zhangg@ornl.gov)

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

The focus of this paper is to provide explicit cost bounds for applying the stochastic Galerkin finite element method (SGFEM) to the solution of an elliptic PDE, with stochastic diffusion coefficient parameterized by finitely many random variables. In particular, we focus on the cost of constructing isotropic total degree SG approximations when the coefficient has both affine and non-affine dependence on the parameters. Under very basic assumptions on the coefficient, the solution to this problem has been shown to have analytic regularity in the random variables (see [33]). As a result, SG approximations that employ a global orthogonal basis have been shown to be optimal projections in the L^2 sense, converging sub-exponentially with respect to the cardinality of the polynomial subspace [32]. However, the computational cost of solving the coupled SG system does not grow linearly in the cardinality of the given subspace. Therefore, the convergence estimates do not indicate the total complexity of obtaining the approximation for a prescribed tolerance.

When the diffusion coefficient can be written as a sum of separable functions of the physical and random parameters, the SG system can be written as a sum of Kronecker products of mass and stiffness matrices. For every mass matrix, each nonzero element leads to a nonzero block of the coupled SG system, where the size of the block equals the size of the finite element stiffness matrix. To solve the SG system, one must simultaneously solve all the coupled finite element problems. In the case that the coefficient is affine in the parameters, the number of nonzeros in each mass matrix is of order $\mathcal{O}(M_p)$ [14], where M_p is the cardinality of the isotropic total degree polynomial subspace of order $p \in \mathbb{N}$. Thus, a matrix-vector product involving the coupled SG system requires $\mathcal{O}(J_h M_p)$ floating point operations (FLOPs), where J_h is the number of physical degrees of freedom. Therefore, the work of solving the coupled SG system when employing an iterative method, e.g., conjugate gradient (CG), is of the order $\mathcal{O}(J_h M_p N_{\text{iter}}^{\text{SG}})$ where $N_{\text{iter}}^{\text{SG}}$ is the number of iterations required to achieve a prescribed accuracy of the fully discrete approximation [3, 14, 34].

On the other hand, when the diffusion coefficient is a general non-affine function of the random parameters, the cost of obtaining an approximation with the SGFEM is not as obvious as before. In this setting we consider two cases, namely, the coefficient is: (1) a polynomial with respect to the random variables, and; (2) a transcendental function with respect to the random variables. In the first case, as we increase the order of the polynomial, the block-sparsity of the SG system decreases, resulting in a SG system that *incrementally* becomes block-dense [12, 14, 21, 23, 34, 35]. In the second case, a separable representation can be guaranteed with the use of an orthogonal expansion [36, 37], such that, substituting the expansion into the discretized PDE recovers the Kronecker product structure. However, when the expansion is not truncated, the SG system is known to be *entirely* block-dense [14, 23]. Without a priori knowledge on the exact sparsity of the mass matrices in this case, it was estimated that the complexity of matrix-vector multiplications of the SG Kronecker product system is between $\mathcal{O}(J_h M_p^2)$ and $\mathcal{O}(J_h M_p^3)$ [34]. As such, it is impossible to make a conclusive statement about the computational cost, and, more importantly, does not account

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

for the two cases above, i.e., when the coefficient is possibly a truncated polynomial of fixed total degree $r \in \mathbb{N}$ such that $1 \leq r < \infty$. In these cases, the work of solving the coupled SG system with an iterative method is given by $\mathcal{O}(J_h \mathcal{M}(p,r) N_{\text{iter}}^{\text{SG}})$, where $\mathcal{M}(p,r)$ is the total number of $\mathcal{O}(J_h)$ finite element problems that must be simultaneously solved.

The key challenge of estimating the cost of solving the SG system when the coefficient is a (truncated) polynomial of finite order is to provide bounds on the block-sparsity of the matrix, i.e., nonzeros of the SG system. To achieve this, we provide a rigorous counting argument, which can be seen as a generalization of [14], for the exact sparsity of the mass matrices for an arbitrary order orthogonal expansion of a non-affine coefficient. As a result, we are able to provide bounds for $\mathcal{M}(p,r)$ of the order $\mathcal{O}(M_p M_r \min\{2^r, M_{\lceil r/2 \rceil}\})$, where M_r is the cardinality of the total degree polynomial subspace used in an orthogonal expansion of order r of the coefficient. This result provides sharper estimates than the bounds in the case of the full orthogonal expansion from [35] since it depends on the truncation order r, and allows us to estimate the total complexity of solving the coupled system for general non-affine coefficients. Since the counting argument for the sparsity of the SG system relies only on the SG discretization of an elliptic operator in terms of orthogonal polynomials, we note that this argument can be reused to estimate the complexity of solving similarly defined PDEs with this method.

In addition, we also develop explicit cost bounds in terms of FLOPs to solve the SG system. Our approach relies on ε-complexity analysis, wherein we balance the errors arising from the approximation with the SGFEM and the iterative solver, e.g., CG, so as to ensure the solution to the fully discrete approximation achieves a given tolerance of $\varepsilon > 0$. With this result, we are able to provide a direct comparison with ε-complexity estimates for the stochastic collocation finite element method (SCFEM) in our previous work [16]. Finally, we present numerical results in agreement with the theoretical work estimates for both the SGFEM and SCFEM all cases described above.

An outline of the paper is as follows. In §2, we provide a discussion on the model problem, and requirements on the diffusion coefficient. In §3, we define the parameterized finite element and SG approximations, derive the SG system, and provide examples of the resulting linear systems that arise from the SG discretization with various coefficients. We then define the cost of solving the SG system and discuss preconditioning strategies. In §4, we derive the exact number of coupled finite element problems in the SG system and bounds on the sparsity in the non-affine case, and present explicit cost bounds of the SGFEM. We also discuss the conditioning of the system in the non-affine case in order to provide a comparison with similar results from [29]. In §5, we briefly describe the SCFEM, and provide theoretical comparison with results from [16] in terms of minimum work to reach a given tolerance, both in the affine and non-affine cases. Finally, in §6, we present illustrative numerical examples corroborating our theoretical results.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

2. Problem setting. We consider the simultaneous solution of the parameterized linear elliptic PDE:

$$\begin{cases} -\nabla \cdot (a(x,\mathbf{y})\nabla u(x,\mathbf{y})) &= f(x) \quad \forall x \in D,\ \mathbf{y} \in \Gamma \\ u(x,\mathbf{y}) &= 0 \quad \forall x \in \partial D,\ \mathbf{y} \in \Gamma \end{cases} \tag{2.1}$$

where $f \in L^2(D)$ is a fixed function of x, $D \subset \mathbb{R}^d$, $d = 1, 2, 3$, is a bounded Lipschitz domain, and $\mathbf{y}(\omega) = (y_1(\omega), \ldots, y_N(\omega)) : \Omega \to \Gamma = \prod_{i=1}^N \Gamma_i \subseteq \mathbb{R}^N$ is a random vector with $\omega \in \Omega$ and Ω the set of outcomes. In this setting we assume the components of $\mathbf{y}$ have a joint probability density (PDF) $\varrho : \Gamma \to \mathbb{R}_+$, with $\varrho(\mathbf{y}) = \prod_{i=1}^N \varrho_i(y_i) \in L^\infty(\Gamma)$ known directly through, e.g., truncations of correlated random fields [22] in $(\Gamma, \mathcal{B}(\Gamma), \varrho(\mathbf{y})d\mathbf{y})$, where $\mathcal{B}(\mathbf{y})$ denotes the Borel σ-algebra on Γ and $\varrho(\mathbf{y})d\mathbf{y}$ is the probability measure of $\mathbf{y}$. Without loss of generality, we assume that $\Gamma_i = [-1, 1]$ and $\varrho_i(y_i) = \frac{1}{2}\ \forall y_i \in \Gamma_i$, $i = 1, \ldots, N$, so that y_i are uniformly distributed (i.e., $y_i \sim \mathcal{U}(-1, 1)\ \forall i = 1, \ldots, N$). We require the following assumptions related to the continuity, coercivity, and holomorphic dependence of the coefficient $a(x, \mathbf{y})$. Namely:

(A1) *There exist constants* $0 < a_{\min} \le a_{\max} < \infty$ *such that for all* $x \in \overline{D}$ *and* $\mathbf{y} \in \Gamma$, $a_{\min} \le a(x, \mathbf{y}) \le a_{\max}$.

(A2) *The complex continuation of* $a(x, \mathbf{y})$, *denoted* $a^* : \mathbb{C}^N \to L^\infty$, *is a* $L^\infty(D)$-*valued holomorphic function on* $\mathbb{C}^N$.

The holomorphic dependence on $\mathbf{y}$ of the coefficient $a(x, \mathbf{y})$ holds in many examples, including polynomial, exponential, and trigonometric functions of the variables $y_1, \ldots, y_N$ shown below.

Example 2.1 (The affine case)**.** *We consider an affine function of the random parameters, e.g.,*

$$a(x, \mathbf{y}) = a_0(x) + \sum_{k=1}^N y_k b_k(x), \quad x \in \overline{D},\ \mathbf{y} \in \Gamma, \tag{2.2}$$

where a_0, $\{b_k\}_{k=1}^N \subset L^\infty(D)$ *are such that* $a(x, \mathbf{y})$ *satisfies* (A1). *Such examples include general Karhunen-Loève expansions [22] or piecewise constant random fields.*

Example 2.2 (The non-affine, polynomial case)**.** *We consider a non-affine, polynomial function of the random parameters, e.g.,*

$$a(x, \mathbf{y}) = a_0(x) + \sum_{1 \le |\boldsymbol{\alpha}| \le \overline{r}} \mathbf{y}^{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}}(x), \quad x \in \overline{D},\ \mathbf{y} \in \Gamma, \tag{2.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ *is a multi-index,* $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_N$, $\mathbf{y}^{\boldsymbol{\alpha}} = y_1^{\alpha_1} \cdots y_N^{\alpha_N}$, $\overline{r} < \infty$ *is the polynomial order of* $a(x, \mathbf{y})$, *and* $\{c_{\boldsymbol{\alpha}}\}_{|\boldsymbol{\alpha}| \le \overline{r}} \subset L^\infty(D)$ *are such that* $a(x, \mathbf{y})$ *satisfies* (A1). *Examples include fixed-order Taylor or orthogonal expansions of a general random field.*

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Example 2.3 (The non-affine, transcendental case)**.** *We consider a non-affine, transcendental function of the random parameters, e.g.,*

$$a(x,\mathbf{y}) = a_0(x) + g(x,\mathbf{y}), \quad x \in \overline{D},\ \mathbf{y} \in \Gamma, \tag{2.4}$$

where $a_0, g \subset L^\infty(D)$*, and* $g(x,\mathbf{y})$ *is a general transcendental function of* x *and* $\mathbf{y}$*, such that* $a(x,\mathbf{y})$ *satisfies* (A1)*. Examples of* $g(x,\mathbf{y})$ *include the sine, logarithm, or exponential functions of* (2.2) *or* (2.3)*.*

Let $L^2_\varrho(\Gamma)$ be the space of square integrable functions with respect to the measure $\varrho(\mathbf{y})d\mathbf{y}$ and $L^\infty(\Gamma)$ be the space of essentially bounded functions, where the essential suppremum is taken with respect to ϱ. By $H^{-1}(D)$ we denote the dual of $H^1_0(D)$, the space of square integrable functions in D having zero trace on the boundary and square integrable distributional derivatives. We will often use the abbreviation $\mathcal{H}^2_\varrho$ to denote $H^1_0(D) \otimes L^2_\varrho(\Gamma)$ and $\mathcal{H}^\infty$ to denote $H^1_0(D) \otimes L^\infty(\Gamma)$. For the space $H^1_0(D)$ we have the energy norm $\|v\|_{H^1_0(D)} = \|\nabla v\|_{L^2(D)}$, hence $H^1_0(D) \otimes L^2_\varrho(\Gamma)$ is a Hilbert space with norm $\|v\|^2_{\mathcal{H}^2_\varrho} = \int_\Gamma \|v\|^2_{H^1_0(D)} \varrho d\mathbf{y}$. The *stochastic weak form* of problem (2.1) is given by: *find* $u \in H^1_0(D) \otimes L^2_\varrho(\Gamma)$ *such that* $\forall v \in H^1_0(D) \otimes L^2_\varrho(\Gamma)$

$$\int_\Gamma \mathcal{B}[u,v](\mathbf{y})\varrho(\mathbf{y})\, d\mathbf{y} = \int_\Gamma F(v)\varrho(\mathbf{y})\, d\mathbf{y}, \tag{2.5}$$

where

$$\mathcal{B}[u,v](\mathbf{y}) = \int_D a(x,\mathbf{y})\nabla u(x,\mathbf{y}) \cdot \nabla v(x,\mathbf{y}) dx, \quad F(v) = \int_D f(x)v(x,\mathbf{y})dx. \tag{2.6}$$

For convenience, we will often use the abbreviation $\mathcal{B}(\mathbf{y}) = \mathcal{B}[\cdot,\cdot](\mathbf{y})$ and suppress the dependence on $x \in D$ in writing $a(\mathbf{y}) = a(\cdot,\mathbf{y})$ and $u(\mathbf{y}) = u(\cdot,\mathbf{y})$. It follows from (A1) that $\mathcal{B}(\mathbf{y})$ is a symmetric, uniformly coercive, and continuous bilinear operator on $H^1_0(D)$, parameterized by $\mathbf{y} \in \Gamma$, and $\mathcal{B}(\mathbf{y})$ induces the norm $\|u\|^2_{\mathcal{B}(\mathbf{y})} := \int_D a(x,\mathbf{y})|\nabla u|^2 dx$. Also from (A1), the Lax-Milgram lemma ensures the existence and uniqueness of the solution u to (2.5) in $H^1_0(D) \otimes L^2_\varrho(\Gamma)$.

The convergence of the global stochastic polynomial methods used to approximate (2.1) exploits the uniform ellipticity of the coefficient $a(\mathbf{y})$ and depends on the regularity of $u(\mathbf{y})$ with respect to $\mathbf{y}$. By $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ we denote the real and imaginary parts of $z \in \mathbb{C}$, and for $0 < \delta < a_{\min}$ we define

$$U(a,\delta) = \{\mathbf{z} \in \mathbb{C}^N : \mathrm{Re}(a(x,\mathbf{z})) \geq \delta, \forall x \in \overline{D}\}. \tag{2.7}$$

If $U(a,\delta) \neq \emptyset$ for some $0 < \delta < a_{\min}$, we say that $a(x,\mathbf{z})$ is uniformly elliptic on the set $U(a,\delta)$ and we refer to $U(a,\delta)$ as its domain of uniform ellipticity. For

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ with $\gamma_i > 1\ \forall i$ we denote the polyellipse

$$\mathcal{E}_{\boldsymbol{\gamma}} = \bigotimes_{1 \le i \le N} \left\{ z_i \in \mathbb{C} : \mathrm{Re}(z_i) \le \frac{\gamma_i + \gamma_i^{-1}}{2} \cos\phi,\ \mathrm{Im}(z_i) \le \frac{\gamma_i - \gamma_i^{-1}}{2} \sin\phi,\ \phi \in [0, 2\pi) \right\}.$$

In [33] it was shown that if $a(\mathbf{y})$ satisfies (A1) and (A2), then for any $0 < \delta < a_{\min}$ there exists a $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ with $\gamma_i > 1\ \forall i$ such that $\mathcal{E}_{\boldsymbol{\gamma}} \subset U(a, \delta)$. We can also similarly define the polydisc $\mathcal{D}_{\boldsymbol{\gamma}} = \bigotimes_{1 \le i \le N} \{z_i \in \mathbb{C} : |z_i| \le \gamma_i\}$, though, for arbitrary $0 < \delta < a_{\min}$, it is not always possible to find a $\boldsymbol{\gamma}$ with $\gamma_i > 1\ \forall i$ such that $\mathcal{D}_{\boldsymbol{\gamma}} \subset U(a, \delta)$. Figure 1 provides an illustration of this fact for various one-dimensional coefficients $a(\mathbf{y})$, $\mathbf{y} \in \mathbb{C}$. Note that in the case of the 6th degree polynomial and exponential random variables, no disc of radius $\gamma > 1$ containing Γ can fit in the region. The following theorem, proved in [33], shows the regularity of the solution u with respect to the parameterization.

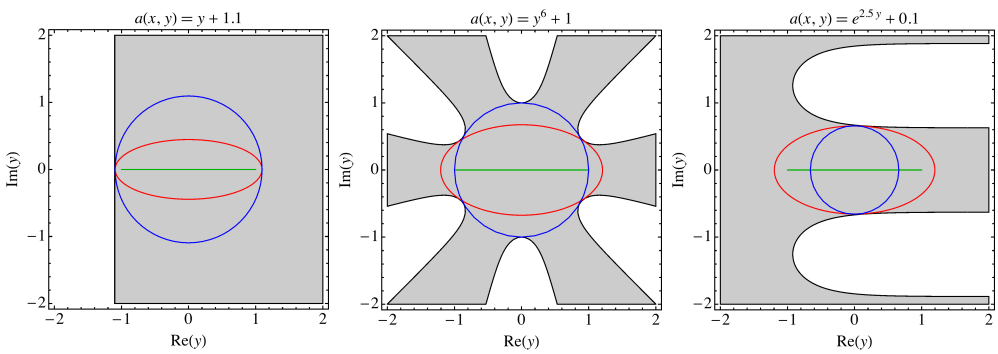

Figure 1: Domains of uniform ellipticity for some one-dimensional coefficients $a(x, y)$ are indicated by the gray regions in each plot. The blue and red curves represent the maximal discs and ellipses, respectively, that can be contained in those domains, and the green lines represent the interval $\Gamma = [-1, 1]$.

Theorem 2.4. *When the coefficient $a(x, \mathbf{y})$ satisfies* (A1) *and* (A2)*, so that for some $0 < \delta < a_{\min}$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)$ with $\gamma_i > 1\ \forall i$ we have $\mathcal{E}_{\boldsymbol{\gamma}} \subset U(a, \delta)$, then the function $\mathbf{z} \mapsto u(\mathbf{z})$ from* (2.1) *is holomorphic in an open neighborhood of $\mathcal{E}_{\boldsymbol{\gamma}}$.*

This result states that a direct consequence of the uniform ellipticity of the function $a(x, \mathbf{y})$ on the polyellipse $\mathcal{E}_{\boldsymbol{\gamma}} \subset U(a, \delta)$ is that the solution u of (2.1) has analytically smooth dependence on the parameterization $\mathbf{y}$. Theorem 2.4 is the key in motivating the construction of global stochastic Galerkin (SG) approximations to the solution u of (2.1), to be described in the following sections.

3. Stochastic Galerkin finite element method.

In this section we define the SGFEM for constructing fully discrete approximations to the solution u of problem (2.1). This discretization employs mixed Galerkin projections in the spatial and parameter domains. In particular we rely on the finite element method for the spatial

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

discretization, described in §3.1, and the stochastic Galerkin method for the parameter discretization, described in §3.2. In §3.3 we describe the linear systems that result from the SG discretization when Examples 2.1, 2.2, and 2.3 are used in problem (2.1). We then conclude in §3.4 with a discussion of the cost of solving the SG systems.

3.1. Parameterized finite element approximation. We briefly define the finite element method for obtaining a discretization of u from (2.1) over the spatial domain D. Let $\mathcal{T}_h$, be a triangulation of D with maximum mesh size $h > 0$, and $V_h(D) \subset H_0^1(D)$ a finite element space of piecewise continuous polynomials on $\mathcal{T}_h$ parameterized by $h \to 0$. Let $\{\phi_j(x)\}_{j=1}^{J_h}$ denote a finite basis of $V_h(D)$ of dimension J_h. We can write the *semi-discrete* problem as: *find* $u_h(\mathbf{y}) \in V_h(D)$ *such that* $\forall v \in V_h(D)$

$$\mathcal{B}[u_h(\mathbf{y}), v](\mathbf{y}) = F(v), \tag{3.1}$$

where $\mathcal{B}[\cdot,\cdot](\mathbf{y})$ and $F(\cdot)$ are defined in (2.6). For almost every $\mathbf{y} \in \Gamma$, problem (3.1) admits a unique solution of the form $u_h(x,\mathbf{y}) = \sum_{j=1}^{J_h} u_j(\mathbf{y})\phi_j(x)$. We discretize problem (3.1) by defining, for $i, j = 1, \ldots, J_h$,

$$[\mathbf{A}]_{i,j}(\mathbf{y}) = \mathcal{B}[\phi_j, \phi_i](\mathbf{y}), \qquad \mathbf{F}_i = F(\phi_i). \tag{3.2}$$

The coefficients $\mathbf{u}_h(\mathbf{y}) = [u_1(\mathbf{y}), u_2(\mathbf{y}), \ldots, u_{J_h}(\mathbf{y})]^{\mathrm{T}}$ of $u_h(x,\mathbf{y})$ are determined by solving the linear system

$$\mathbf{A}(\mathbf{y})\mathbf{u}_h(\mathbf{y}) = \mathbf{F}, \tag{3.3}$$

at fixed realizations of $\mathbf{y} \in \Gamma$. Here $\mathbf{A}(\mathbf{y})$ is symmetric and positive-definite so that (3.3) can be solved by iterative methods such as the conjugate gradient (CG) method.

3.2. Stochastic Galerkin approximation with an orthogonal basis. Based on the smoothness of the solution u to (2.1), characterized by Theorem 2.4, we now consider the construction of approximations to u in terms of global polynomials. Let $\Lambda_p \subset \mathbb{N}_0^N$ be a finite set of multi-indices, e.g., having dimension $\#\Lambda_p < \infty$, and define the space of polynomials $\mathcal{P}_{\Lambda_p}(\Gamma) = \mathrm{span}\{\mathbf{y}^{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \Lambda_p\}$. A general global polynomial approximation problem can be framed in terms of solving for the $\#\Lambda_p$ stochastic degrees of freedom (SDOF) $\{u_{\boldsymbol{p}}\}_{\boldsymbol{p}\in\Lambda_p}$. Depending on the type of approach, the resulting systems of equations are either coupled or decoupled finite element systems. Some isotropic examples of such index sets include

$$\begin{aligned} \Lambda_p^{\mathrm{TP}} &= \left\{ \boldsymbol{p} \in \mathbb{N}_0^N : \max_{1\le i\le N} p_i \le p \right\}, \quad \Lambda_p^{\mathrm{TD}} = \left\{ \boldsymbol{p} \in \mathbb{N}_0^N : \sum_{n=1}^{N} p_n \le p \right\}, \\ \Lambda_p^{\mathrm{SM}} &= \left\{ \boldsymbol{p} \in \mathbb{N}_0^N : \sum_{n=1}^{N} f(p_n) \le f(p) \right\}, \quad f(p) = \begin{cases} 0, & p = 0 \\ 1, & p = 1 \\ \lceil \log_2(p) \rceil, & p \ge 2 \end{cases} \end{aligned} \tag{3.4}$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

corresponding to the Tensor Products (TP), Total Degree (TD), and Smolyak (SM) polynomial spaces $\mathcal{P}_{\Lambda_p^{\mathrm{TP}}}(\Gamma)$, $\mathcal{P}_{\Lambda_p^{\mathrm{TD}}}(\Gamma)$, and $\mathcal{P}_{\Lambda_p^{\mathrm{SM}}}(\Gamma)$, respectively. When the solution u exhibits an anisotropic dependence on the parameters $\mathbf{y}$, anisotropic weighted versions of the index sets defined in (3.4) can be introduced to further reduce the number of SDOF needed to approximate u at a desired accuracy.

Remark 3.1 (Best M-term and quasi-optimal approximations)**.** *The optimal choice of Λ_p would be the set Λ of cardinality M such that the corresponding approximation provides maximum accuracy out of all sets of size M. Such approximations are referred to as best M-term approximations, and recent work has focussed on the construction of best M-term Taylor and Galerkin approximations [4, 6–8, 33]. These approaches construct Λ by utilizing the largest M coefficients $u_{\boldsymbol{p}}$ or sharp upper bounds of $u_{\boldsymbol{p}}$. However, in this effort we focus on analyzing the computational complexity of finding solutions to* (2.5) *in $\mathcal{P}_{\Lambda_p}(\Gamma)$ for a prescribed index set Λ_p.*

Given a specific choice of index set Λ_p, let $\boldsymbol{p} = (p_1, \ldots, p_N) \in \Lambda_p$ be a multi-index and ψ_{p_n} be a one dimensional orthonormal polynomial over Γ_n with respect to ϱ_n for each $n = 1, \ldots, N$. We define $\Psi_{\boldsymbol{p}}(\mathbf{y}) = \prod_{n=1}^{N} \psi_{p_n}(y_n)$, to be the multi-variate orthonormal polynomial over Γ with respect to $\varrho(\mathbf{y})$ corresponding to the multi-index $\boldsymbol{p}$. In the case that $\varrho = \frac{1}{2}$ for each $n = 1, \ldots, N$, ψ_{p_n} and $\Psi_{\boldsymbol{p}}$ are the univariate and multivariate Legendre polynomials respectively. It follows that $\{\Psi_{\boldsymbol{p}}\}_{\boldsymbol{p}\in\Lambda_p}$ forms a basis of $\mathcal{P}_{\Lambda_p}(\Gamma)$ with dimension $M_p = \dim(\mathcal{P}_{\Lambda_p}(\Gamma)) = \#\Lambda_p$. With $\{\phi_j\}_{j=1}^{J_h}$ as in §3.1 and $\{\Psi_{\boldsymbol{p}}\}_{\boldsymbol{p}\in\Lambda_p}$ as above, we can now write the fully discrete stochastic Galerkin (SG) approximation as

$$u_{h,p}(x, \mathbf{y}) = \sum_{\boldsymbol{p}\in\Lambda_p} \sum_{j=1}^{J_h} u_{j,\boldsymbol{p}}\, \phi_j(x) \Psi_{\boldsymbol{p}}(\mathbf{y}), \tag{3.5}$$

whose coefficients can be constructed by solving the following coupled problem: *find $u_{h,p} \in V_h(D) \otimes \mathcal{P}_{\Lambda_p}(\Gamma)$ such that for all $v \in V_h(D) \otimes \mathcal{P}_{\Lambda_p}(\Gamma)$*

$$\mathbb{E}\left[\mathcal{B}[u_{h,p}, v](\mathbf{y})\right] = \mathbb{E}\left[F(v)\right], \tag{3.6}$$

where $\mathcal{B}[\cdot,\cdot](\mathbf{y})$ and $F(\cdot)$ are defined in (2.6). To construct the linear system of equations resulting from the SG approximation given by (3.5), we let $\mathbf{u}_{h,\boldsymbol{p}} = [u_{1,\boldsymbol{p}}, \ldots, u_{J_h,\boldsymbol{p}}]^{\mathrm{T}}$ be the vector of nodal values of the finite element solution corresponding to the $\boldsymbol{p}$-th stochastic mode of $u_{h,p}$, and $\mathbf{u}_{h,p} = [\mathbf{u}_{h,\boldsymbol{p}}]^{\mathrm{T}}_{\boldsymbol{p}\in\Lambda_p}$. Observe that when f is deterministic $\langle \Psi_{\boldsymbol{p}} \mathbf{F}_i \rangle = \mathbf{F}_i \delta_{\mathbf{0},\boldsymbol{p}}$ for all $i = 1, \ldots, J_h$, where $\delta_{\mathbf{0},\boldsymbol{p}} = 1$ if $\boldsymbol{p} = \mathbf{0}$ and $\delta_{\mathbf{0},\boldsymbol{p}} = 0$ otherwise. Performing a Galerkin projection onto $\mathrm{span}\{\Psi_{\boldsymbol{p}}\}_{\boldsymbol{p}\in\Lambda_p}$ for the solution of (3.6) yields the following system: for each $\boldsymbol{p} \in \Lambda_p$

$$\sum_{\boldsymbol{q}\in\Lambda_p} \langle \Psi_{\boldsymbol{p}}(\mathbf{y}), \mathbf{A}(\mathbf{y}) \Psi_{\boldsymbol{q}}(\mathbf{y}) \rangle \mathbf{u}_{h,\boldsymbol{q}} = \langle \Psi_{\boldsymbol{p}}(\mathbf{y}), \mathbf{F} \rangle, \tag{3.7}$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

which can be written algebraically as a system of *fully coupled finite element problems*: for each $\boldsymbol{p} \in \Lambda_p$

$$\sum_{\boldsymbol{q}\in\Lambda_p} [\mathbf{K}]_{\boldsymbol{p},\boldsymbol{q}} \mathbf{u}_{h,\boldsymbol{q}} = \mathbf{F}\delta_{\mathbf{0},\boldsymbol{p}} \tag{3.8}$$

with $[\mathbf{K}]_{\boldsymbol{p},\boldsymbol{q}} = \langle \Psi_{\boldsymbol{p}}(\mathbf{y}), \mathbf{A}(\mathbf{y})\Psi_{\boldsymbol{q}}(\mathbf{y})\rangle$ and $\mathbf{A}(\mathbf{y})$ as given in (3.2).

Remark 3.2. *Typically matrix free methods are applied to solve* (3.8) *without ever explicitly forming* $\mathbf{K}$ *in memory, as described in [27]. When the resulting system is sparse, as a result of an affine coefficient* $a(x, \mathbf{y})$*, e.g., Example 2.1, this can lead to computationally efficient solution strategies. However, these implementations rely on the fact that the coefficient* $a(x, \mathbf{y})$ *can be written as a sum of separable functions of* x *and* $\mathbf{y}$*, e.g.,* $a(x, \mathbf{y}) = \sum_{j=1}^{N} b_j(x)c_j(\mathbf{y})$*. For the transcendental function* $a(x, \mathbf{y})$ *from Example 2.3, this may not be the case. Moreover, when* $\mathbf{K}$ *is block-dense, matrix-vector multiplications require approximately* $\mathcal{O}(J_h M_p^2)$ *floating point operations (FLOPs), so that when iterative methods are used, the solution of the fully coupled finite element problems given in* (3.8) *becomes unfeasible.*

3.3. Representations of $a(\mathbf{y})$ and the corresponding matrix $\mathbf{K}$. For a general coefficient $a(x, \mathbf{y})$, the matrix $\mathbf{K}$ in (3.8) requires the storage of at most M_p^2 block matrices of the size and sparsity of $\mathbf{A}(\mathbf{y})$, i.e., $\mathcal{O}(J_h M_p^2)$ elements. However, in several specific cases the actual block-sparsity of $\mathbf{K}$ is much less. We recall the coefficient from Example 2.1, where $\mathbf{K}$ can be rewritten

$$[\mathbf{K}]_{\boldsymbol{p},\boldsymbol{q}} = \langle \Psi_{\boldsymbol{p}}(\mathbf{y}), \Psi_{\boldsymbol{q}}(\mathbf{y})\rangle \mathbf{A}_0 + \sum_{k=1}^{N} \langle y_k \Psi_{\boldsymbol{p}}(\mathbf{y}), \Psi_{\boldsymbol{q}}(\mathbf{y})\rangle \mathbf{A}_k,$$

with $[\mathbf{A}_0]_{i,j} = \int_D a_0(x)\nabla\phi_j(x)\cdot\nabla\phi_i(x)dx$ and $[\mathbf{A}_k]_{i,j} = \int_D b_k(x)\nabla\phi_j(x)\cdot\nabla\phi_i(x)dx$. If we let $[\mathbf{G}_0]_{\boldsymbol{p},\boldsymbol{q}} = \langle \Psi_{\boldsymbol{p}}(\mathbf{y}), \Psi_{\boldsymbol{q}}(\mathbf{y})\rangle$ and $[\mathbf{G}_k]_{\boldsymbol{p},\boldsymbol{q}} = \langle y_k \Psi_{\boldsymbol{p}}(\mathbf{y}), \Psi_{\boldsymbol{q}}(\mathbf{y})\rangle$, then $\mathbf{K}$ has a matrix representation, given by,

$$\mathbf{K} = \mathbf{G}_0 \otimes \mathbf{A}_0 + \sum_{k=1}^{N} \mathbf{G}_k \otimes \mathbf{A}_k, \tag{3.9}$$

where $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$. We note that a similar construction can be optained for any coefficient $a(x, \mathbf{y})$ which can be written as a sum of separable functions of x and $\mathbf{y}$, such as the polynomial function of Example 2.2.

However, when $a(x, \mathbf{y})$ is not separable in x and $\mathbf{y}$, this construction is no longer valid, and the resulting matrix $\mathbf{K}$ may be block-dense if we simply carry out the Galerkin projections and compute $\mathbf{K}$ directly. For certain special cases, e.g., when the diffusion coefficient is given by a log-transformed random field, the problem can be reformulated as a convection-diffusion problem, and the resulting system can be solved much more efficiently than the original problem [35]. In general, this reformulation

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

is not applicable. Hence, for a general transcendental coefficient $a(x,\mathbf{y})$, as given in Example 2.3, we project the coefficient onto an additional subspace $\mathcal{P}_{\Lambda_r}(\Gamma)$, $r \in \mathbb{N}_0$, in order to obtain a separable representation. To see this, define $\{\Psi_{\boldsymbol{r}}(\mathbf{y})\}_{0\le|\boldsymbol{r}|}$ to be the (infinite) basis of orthonormal polynomials of $L^2_\varrho(\Gamma)$ as in §3.2. Then $a(x,\mathbf{y})$ can be written as an expansion such that $a(x,\mathbf{y}) = \sum_{0\le|\boldsymbol{r}|} a_{\boldsymbol{r}}(x)\Psi_{\boldsymbol{r}}(\mathbf{y})$, where the coefficients $a_{\boldsymbol{r}}(x) = \langle a(x,\mathbf{y}), \Psi_{\boldsymbol{r}}(\mathbf{y})\rangle$. Let Λ_r be an index set of the type described in §3.2. Since infinite series representations are not practical in computations, we seek a truncation

$$a^r(x,\mathbf{y}) := \sum_{\boldsymbol{r}\in\Lambda_r} a_{\boldsymbol{r}}(x)\Psi_{\boldsymbol{r}}(\mathbf{y}) \tag{3.10}$$

in the subspace $\mathcal{P}_{\Lambda_r}(\Gamma)$ for some $r \in \mathbb{N}_0$. When $a^r(x,\mathbf{y}) \neq a(x,\mathbf{y})$, e.g., in the case that the projection order r is chosen to minimize error independent of the SG discretization, we let $u^r_{h,p}$ denote the corresponding solution to the fully discrete SG approximation problem with $a(x,\mathbf{y})$ replaced with $a^r(x,\mathbf{y})$. By substituting $a^r(x,\mathbf{y})$ into (2.6) we obtain

$$\int_D \left(\sum_{\boldsymbol{r}\in\Lambda_r} a_{\boldsymbol{r}}(x)\Psi_{\boldsymbol{r}}(\mathbf{y})\right) \nabla\phi_j(x)\cdot\nabla\phi_i(x)dx = \sum_{\boldsymbol{r}\in\Lambda_r} [\mathbf{A}_{\boldsymbol{r}}]_{i,j}\Psi_{\boldsymbol{r}}(\mathbf{y}), \tag{3.11}$$

$$[\mathbf{A}_{\boldsymbol{r}}]_{i,j} = \int_D a_{\boldsymbol{r}}(x)\nabla\phi_j(x)\cdot\nabla\phi_i(x)dx. \tag{3.12}$$

Equation (3.11) represents an expansion of the stochastic finite element stiffness matrix $\mathbf{A}(\mathbf{y})$ and equation (3.12) represents the $\boldsymbol{r}$-th mode of the expansion. Let $\mathbf{u}^r_{h,\boldsymbol{p}} = [u^r_{1,\boldsymbol{p}},\ldots,u^r_{J_h,\boldsymbol{p}}]^{\mathrm{T}}$ denote the vector of nodal values of the finite element solution corresponding to the $\boldsymbol{p}$-th stochastic mode of $u^r_{h,p}$, and $\mathbf{u}^r_{h,p} = [\mathbf{u}^r_{h,\boldsymbol{p}}]^{\mathrm{T}}_{\boldsymbol{p}\in\Lambda_p}$. We substitute the expansion of $\mathbf{A}(\mathbf{y})$ into the Galerkin equations (3.7), to obtain the coupled system: for each $\boldsymbol{p}\in\Lambda_p$

$$\sum_{\boldsymbol{r}\in\Lambda_r}\sum_{\boldsymbol{q}\in\Lambda_p} [\mathbf{G}_{\boldsymbol{r}}]_{\boldsymbol{p},\boldsymbol{q}}\mathbf{A}_{\boldsymbol{r}}\mathbf{u}^r_{h,\boldsymbol{q}} = \langle\Psi_{\boldsymbol{p}},\mathbf{F}\rangle, \qquad [\mathbf{G}_{\boldsymbol{r}}]_{\boldsymbol{p},\boldsymbol{q}} = \langle\Psi_{\boldsymbol{p}}\Psi_{\boldsymbol{q}}\Psi_{\boldsymbol{r}}\rangle. \tag{3.13}$$

Alternatively, similar to (3.9), we may define $\mathbf{K}_r = \sum_{\boldsymbol{r}\in\Lambda_r}\mathbf{G}_{\boldsymbol{r}}\otimes\mathbf{A}_{\boldsymbol{r}}$ to again obtain the coupled system of finite element problems: for all $\boldsymbol{p}\in\Lambda_p$

$$\sum_{\boldsymbol{q}\in\Lambda_p}[\mathbf{K}_r]_{\boldsymbol{p},\boldsymbol{q}}\mathbf{u}^r_{h,\boldsymbol{q}} = \mathbf{F}\delta_{\mathbf{0},\boldsymbol{p}} \qquad \forall\boldsymbol{p}\in\Lambda_p. \tag{3.14}$$

Note that in forming $\mathbf{K}_r$, we now need only store the matrices $\{\mathbf{G}_{\boldsymbol{r}}\}_{\boldsymbol{r}\in\Lambda_r}$ and $\{\mathbf{A}_{\boldsymbol{r}}\}_{\boldsymbol{r}\in\Lambda_r}$, so that the required storage is minimized and efficient, matrix-free, solution strategies discussed in Remark 3.2 can be applied.

Remark 3.3 (Projection and well-posedness)**.** *In the case that the coefficient is a transcendental function of the random variables, as in Example 2.3, there does not*

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

exist a $r \in \mathbb{N}_0$ *such that the projection* (3.10) *is exact. Due to the orthogonality of the basis, setting* $r = 2p$ *in the construction of* $\mathbf{K}_r$ *yields an entirely block-dense system [23] that is equivalent to* (3.8)*, and computationally infeasible to solve. A more practical approach is to choose the expansion order* $0 \leq r \leq 2p$*, based on a-priori estimates of the error in the solution introduced by the truncation, so that the error when using the truncated expansion does not exceed that of the SG approximation. In this approach however, it becomes important to consider whether the truncated projection violates the well-posedness of* (2.1) *by failing to satisfy assumption* (A1)*. One way to guarantee this is to choose* $\tilde{r} \leq r \leq 2p$ *such that*

$$\tilde{r} := \min\{r \in \mathbb{N}_0 : \|a - a_\nu\|_{L^\infty(\Gamma \times D)} \leq a_{\min},\ \forall \nu \in \mathbb{N}_0,\ \nu \geq r\}. \tag{3.15}$$

An example of this problem can be seen in Figure 2 where for the function $a(x, y) = 0.1 + \exp(2.5y)$*, uniform ellipticity of the truncated projection* $a^r(x, y)$ *does not hold on* Γ *for* $r < 4$*.*

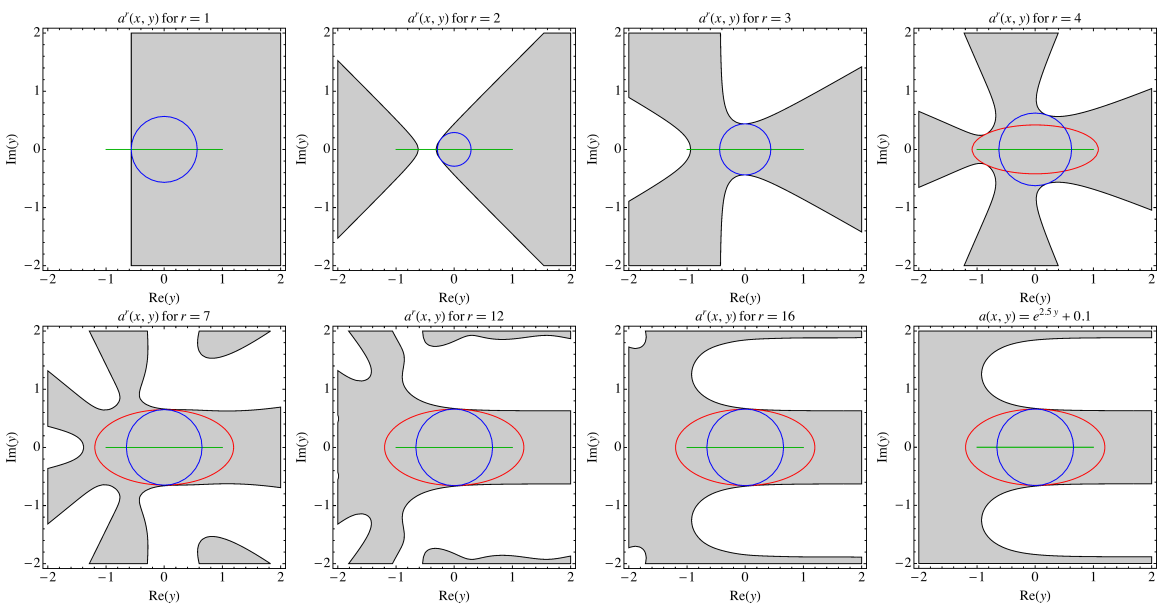

Figure 2: Domains of uniform ellipticity for the total degree orthogonal expansions of order r of the one-dimensional coefficient $a(x, y) = 0.1 + \exp(2.5y)$, for $y \in \Gamma = [-1, 1] \subset \mathbb{R}^1$, are indicated by the gray regions in each plot. The last plot shows the domain of uniform ellipticity of the original function $a(x, y)$. The blue and red curves represent the maximal discs and ellipses, respectively, that can be contained in those domains, and the green lines represent the interval Γ.

3.4. Cost of solving the generalized SG system. Without loss of generality, to solve the stochastic Galerkin system (3.14) for $\mathbf{u}^r_{h,p}$, we use the precondtioned conjugate gradient (PCG) method, wherein, for the unpreconditioned CG method,

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

we have the estimate

$$\|\mathbf{u}_{h,p}^{r} - \mathbf{u}_{h,p}^{r,(k)}\|_{\mathbf{K}_r} \leq 2\left(\frac{\sqrt{\boldsymbol{\kappa}_r}-1}{\sqrt{\boldsymbol{\kappa}_r}+1}\right)^k \|\mathbf{u}_{h,p}^{r} - \mathbf{u}_{h,p}^{r,(0)}\|_{\mathbf{K}_r}. \tag{3.16}$$

Here $\boldsymbol{\kappa}_r$ is the condition number of $\mathbf{K}_r$, $\mathbf{u}_{h,p}^{r,(0)}$ is the vector of the initial guess, and $\mathbf{u}_{h,p}^{r,(k)}$ is the output of the k-th iteration of the CG solver. The CG method is highly dependent on the conditioning of the system, and when $\boldsymbol{\kappa}_r$ is large, the number of iterations needed to reduce the error in $\mathbf{u}_{h,p}^{r,(k)}$ will also be significant. Hence we introduce the mean-based block-diagonal preconditioner (see, e.g., [27, 29]),

$$\mathbf{P} := \mathbf{G_0} \otimes \mathbf{A_0}, \tag{3.17}$$

with $\mathbf{A_0}$ and $\mathbf{G_0}$ the matrices defined in (3.12) and (3.13) for $\boldsymbol{r} = \mathbf{0}$, respectively.

For $\boldsymbol{r} \in \Lambda_r$, at every iteration of the CG method, or any iterative approach, each nonzero entry in each matrix $\mathbf{G}_{\boldsymbol{r}}$ implies a matrix-vector product of the form $\langle \Psi_{\boldsymbol{p}}\Psi_{\boldsymbol{q}}\Psi_{\boldsymbol{r}}\rangle \mathbf{A}_{\boldsymbol{r}}\mathbf{p}_{\boldsymbol{q}}^{(k)}$, where $\langle \Psi_{\boldsymbol{p}}\Psi_{\boldsymbol{q}}\Psi_{\boldsymbol{r}}\rangle$ is a scalar quantity. Let $\mathrm{nnz}(\mathbf{A})$ denote the number of nonzeros of a matrix $\mathbf{A}$, and define

$$\mathcal{M}(p,r) = \sum_{\boldsymbol{r}\in\Lambda_r} \mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}}) \tag{3.18}$$

to be the total number of nonzeros in all of the matrices $\{\mathbf{G}_{\boldsymbol{r}}\}_{\boldsymbol{r}\in\Lambda_r}$ at order p. With this in mind, an upper bound for the work in floating point operations (FLOPs) of solving (3.14) is given by

$$W^{\mathrm{SG}} \approx \mathcal{O}(J_h * \mathcal{M}(p,r) * N_{\mathrm{iter}}^{\mathrm{SG}}), \tag{3.19}$$

where the term $\mathcal{O}(J_h)$ corresponds to the cost of a single finite element matrix-vector product, and $N_{\mathrm{iter}}^{\mathrm{SG}}$ is the number of iterations of the CG solver without a preconditioner. If we apply a preconditioner, in hopes to minimize $N_{\mathrm{iter}}^{\mathrm{SG}}$, we must also account for the added cost of applying the preconditioner at each iteration. With the mean-based preconditioner from (3.17), at each iteration we multiply an additional matrix of size $J_hM_p \times J_hM_p$, but the matrix consists only of M_p diagonal blocks. Hence, for each iteration we require M_p additional matrix-vector products of the size of the original finite element system, so the work estimate in FLOPs for the case of this preconditioner is given by

$$W^{\mathrm{pSG}} \approx \mathcal{O}\left(J_h * (M_p + \mathcal{M}(p,r)) * N_{\mathrm{iter}}^{\mathrm{pSG}}\right), \tag{3.20}$$

where $N_{\mathrm{iter}}^{\mathrm{pSG}}$ is the number of iterations needed by the PCG method. Other preconditioners, such as the Kronecker product preconditioner suggested in [34] would require a different form of (3.20).

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Figure 3 displays the effect of fixing the projection order of the solution but increasing the order of the projection of the coefficient. In order to minimize the error of the projection, such a situation would be required if the coefficient is highly nonlinear and reflects the importance of considering $\mathcal{M}(p,r)$ in the computational cost of the SGFEM.

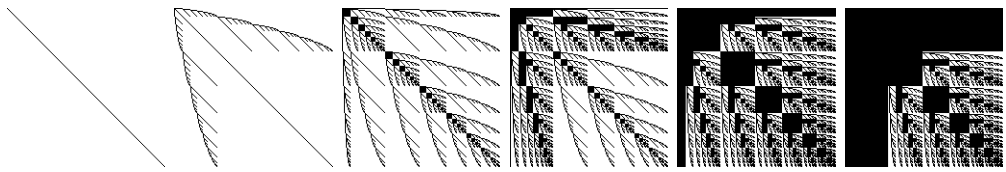

Figure 3: Visualization of the number of nonzeros of a 165×165 SG matrix with elements $[\mathbf{K}_r]_{\boldsymbol{p},\boldsymbol{q}} = \sum_{\boldsymbol{r}\in\Lambda_r} [\mathbf{G}_{\boldsymbol{r}}]_{\boldsymbol{p},\boldsymbol{q}} * \mathbf{A}_{\boldsymbol{r}}$. Each pixel represents a block finite element system when using a total degree projection of the solution of fixed degreee $p = 3$, and increasing the total degree of the projection of the coefficient, i.e., $r = 0, 1, 2, 3, 4, 5$. At $r = 6$, the matrix is entirely block-dense.

4. Explicit cost bounds for the SGFEM. The primary goal of this section is to estimate the algorithmic complexity required by the SGFEM to construct an approximation to (2.1) within a prescribed tolerance $\varepsilon > 0$. We assume $a(x, \mathbf{y})$ is a general non-affine coefficient, as in Examples 2.2 and 2.3, satisfying assumptions (A1) and (A2). Let $a^r(x, \mathbf{y})$ be the orthogonal expansion of $a(x, \mathbf{y})$, given by (3.10), of total degree r, i.e., $a^r(x, \mathbf{y}) \in \mathcal{P}_{\Lambda_r}(\Gamma)$ with $\Lambda_r = \Lambda_r^{\mathrm{TD}}$ from (3.4). We further assume that $\tilde{r} \leq r \leq 2p$, with $\tilde{r}$ given in (3.15), so that $a^r(x, \mathbf{y})$ also satisfies (A1) and (A2). We will focus on the complexity of solving (3.14), when the stochastic discretization to (2.1) is performed in $\mathcal{P}_{\Lambda_p}(\Gamma)$ with $\Lambda_p = \Lambda_p^{\mathrm{TD}}$ from (3.4), i.e., in the space of total degree polynomials of order p, and the physical discretization is performed with the finite element method. These results are presented in the context of solving the linear system (3.14) with a PCG method when a zero initial vector is used to seed the solver. The results, however, can be generalized to other methods, such as preconditioned GMRES, Krylov methods, etc.

The results are organized as follows. In §4.1 we discuss the overall complexity of the matrix-vector products associated with solving (3.14) when using the SG matrix $\mathbf{K}_r = \sum_{\boldsymbol{r}\in\Lambda_r} \mathbf{G}_{\boldsymbol{r}} \otimes \mathbf{A}_{\boldsymbol{r}}$. Our analysis extends the results of [14] in order to provide a bound on the block-sparsity of the SG system $\mathbf{K}_r$ in the more general setting of a non-affine coefficient $a^r(x, \mathbf{y})$, as given in Examples 2.2 and 2.3. In particular, for $\mathbf{G}_{\boldsymbol{r}}$ given in (3.13), we show that $\mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}}) = \mathcal{O}(\min\{2^{|\boldsymbol{r}|}, M_{\lceil|\boldsymbol{r}|/2\rceil}\} M_{p-\lceil|\boldsymbol{r}|/2\rceil})$ for every $\boldsymbol{r} \in \Lambda_r$, where $M_r = \binom{N+r}{N}$ for $r \in \mathbb{N}_0$, when solving (3.14), so that the total complexity of the matrix-vector products with the Galerkin system is $\mathcal{O}(J_h M_p M_r \min\{2^r, M_{\lceil r/2\rceil}\})$. In §4.2, we perform an ε-complexity analysis to derive the explicit cost bounds of the SGFEM using PCG, in terms of FLOPs as the tolerance $\varepsilon \to 0$. Finally, in section 4.3 we discuss issues related to the conditioning of the SG system.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

4.1. Complexity of matrix-vector multiplications in the SG approximation. In this section we provide a rigorous counting argument for the number of nonzeros in the Galerkin matrices $\mathbf{G}_{\boldsymbol{r}}$ from (3.13), for arbitrary $\boldsymbol{r} \in \Lambda_r$ with $0 \le r \le 2p$, and $p \in \mathbb{N}_0$. Our main result, given by Theorem 4.1, is an extension of the estimates in [14, Lemma 28], where an upper bound on the sparsity of $\mathbf{G}_{\boldsymbol{r}}$ was shown in the case $|\boldsymbol{r}| = 1$ when the integrals $[\mathbf{G}_{\boldsymbol{r}}]_{\boldsymbol{p},\boldsymbol{q}} = \langle \Psi_{\boldsymbol{r}} \Psi_{\boldsymbol{p}} \Psi_{\boldsymbol{q}} \rangle$ are defined in terms of even weight functions ϱ to the generalized case $|\boldsymbol{r}| \in \mathbb{N}$. We then provide upper bounds on the total number of nonzeros blocks $\mathcal{M}(p,r) = \sum_{\boldsymbol{r} \in \Lambda_r} \mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}})$ of the matrix $\mathbf{K}_r$ from (3.14), both in the cases that $a(x,\mathbf{y})$ is a finite order polynomial as in Example 2.2 and the case that $a(x,\mathbf{y})$ is a transcendental function of the random variables, as in Example 2.3. Our first major result is summarized in the following Theorem:

Theorem 4.1. *Let Λ_p and Λ_r be the isotropic total degree index sets corresponding to the solution and coefficient, respectively, with $p, r \in \mathbb{N}_0$, and $0 \le r \le 2p$. If $\boldsymbol{r} \in \Lambda_r$, and ϱ_i are even for all $i = 1, \dots, N$, then for the matrix $\mathbf{G}_{\boldsymbol{r}}$ from* (3.13) *we have*

$$\begin{aligned} \mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}}) &= \sum_{\ell=\lceil |\boldsymbol{r}|/2 \rceil}^{|\boldsymbol{r}|} c(\boldsymbol{r},\ell) \binom{N+p-\ell}{p-\ell}, \\ c(\boldsymbol{r},\ell) &= \begin{cases} \#\boldsymbol{S}(\boldsymbol{r},\ell) & |\boldsymbol{r}| \text{ even, } \ell = |\boldsymbol{r}|/2, \\ 2\#\boldsymbol{S}(\boldsymbol{r},\ell) & \text{otherwise,} \end{cases} \end{aligned} \tag{4.1}$$

with $\boldsymbol{S}(\boldsymbol{r},\ell) = \left\{ \boldsymbol{s} \in \mathbb{N}_0^N : |\boldsymbol{s}| = \ell,\ \boldsymbol{s} \le \boldsymbol{r} \right\}$, so that $\#\boldsymbol{S}(\boldsymbol{r},\ell)$ is equal to the coefficient of t^ℓ in the polynomial $P_{\boldsymbol{r}}(t) = \prod_{i=1}^N \sum_{j=0}^{r_i} t^j$. Moreover, we have the following bound for $\mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}})$, *i.e.,*

$$\mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}}) \le 2 \min \left\{ 2^{|\boldsymbol{r}|}, \binom{N + \lceil |\boldsymbol{r}|/2 \rceil}{N} \right\} \binom{N+p-\lceil |\boldsymbol{r}|/2 \rceil}{N}, \tag{4.2}$$

so that

$$\mathcal{M}(p,r) \le 2 \sum_{j=0}^{r} \min \left\{ 2^j, \binom{N+\lceil j/2 \rceil}{N} \right\} \binom{N-1+j}{N-1} \binom{N+p-\lceil j/2 \rceil}{N}. \tag{4.3}$$

Proof. For a given $\boldsymbol{r} \in \Lambda_r$, we estimate the number of pairs $(\boldsymbol{p},\boldsymbol{q}) \in \Lambda_p \times \Lambda_p$ such that $\langle \Psi_{\boldsymbol{r}} \Psi_{\boldsymbol{p}} \Psi_{\boldsymbol{q}} \rangle = \prod_{i=1}^N \langle \psi_{r_i} \psi_{p_i} \psi_{q_i} \rangle \neq 0$. To do this, we extend the result of [14, Lemma 28] to a general matrix $\mathbf{G}_{\boldsymbol{r}}$ with $|\boldsymbol{r}| \in \mathbb{N}$. Since $\{\Psi_{\boldsymbol{r}}\}_{\boldsymbol{r} \in \Lambda_r}$ are orthonormal with respect to the even weight function $\rho(\mathbf{y}) = \prod_{i=1}^N \rho_i(y_i)$, we see that $\langle \Psi_{\boldsymbol{r}} \Psi_{\boldsymbol{p}} \Psi_{\boldsymbol{q}} \rangle \neq 0$ if and only if $(\boldsymbol{p},\boldsymbol{q}) \in \Theta_{\boldsymbol{r}}$, where

$$\begin{aligned} \Theta_{\boldsymbol{r}} = \{ (\boldsymbol{p},\boldsymbol{q}) \in \Lambda_p \times \Lambda_p : |p_i - q_i| \le r_i \le p_i + q_i, \\ \text{and } p_i + q_i + r_i \text{ is even } \forall i = 1, \dots, N \}. \end{aligned}$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Therefore, to estimate the number of nonzeros in the matrix $\mathbf{G}_{\boldsymbol{r}}$, we must estimate $\#\Theta_{\boldsymbol{r}}$. However, $\Theta_{\boldsymbol{r}}$ is different for each $\boldsymbol{r} \in \Lambda_r$. Even when $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \Lambda_r$ are such that $|\boldsymbol{r}_1| = |\boldsymbol{r}_2|$, in general we do not have that $\#\Theta_{\boldsymbol{r}_1} = \#\Theta_{\boldsymbol{r}_2}$. On the other hand, if $\boldsymbol{r}_2$ is a permutation of $\boldsymbol{r}_1$, then it is easy to see that $\#\Theta_{\boldsymbol{r}_1} = \#\Theta_{\boldsymbol{r}_2}$ since Λ_p is the isotropic total degree set. Also note that $\langle \Psi_{\boldsymbol{r}} \Psi_{\boldsymbol{p}} \Psi_{\boldsymbol{q}} \rangle = \langle \Psi_{\boldsymbol{r}} \Psi_{\boldsymbol{q}} \Psi_{\boldsymbol{p}} \rangle$ so if $(\boldsymbol{p}, \boldsymbol{q}) \in \Theta_{\boldsymbol{r}}$ then $(\boldsymbol{q}, \boldsymbol{p}) \in \Theta_{\boldsymbol{r}}$ as well. Note that $\Theta_{\boldsymbol{r}}$ can be rewritten

$$\begin{aligned}\Theta_{\boldsymbol{r}} = \{(\boldsymbol{p}, \boldsymbol{q}) \in \Lambda_p \times \Lambda_p : |p_i - q_i| \le r_i \le p_i + q_i, \\ \text{and } |p_i - q_i| + r_i \text{ is even } \forall i = 1, \ldots, N\}.\end{aligned}$$

Hence if $(\boldsymbol{p}, \boldsymbol{q}) \in \Theta_{\boldsymbol{r}}$, then we see that $(\boldsymbol{p}, \boldsymbol{q})$ must satisfy

(i) $|p_i - q_i| \le r_i$ for all $i = 1, \ldots, N$,

(ii) $r_i \le p_i + q_i$ for all $i = 1, \ldots, N$, and

(iii) $|p_i - q_i| + r_i$ is even for all $i = 1, \ldots, N$.

Note that when $r_i = 0$, we see that $p_i = q_i \le p$, and when $r_i > 0$ we see that (i) and (iii) imply $|p_i - q_i| \in \{0, 2, 4, \ldots, r_i\}$ for r_i even, and $|p_i - q_i| \in \{1, 3, 5, \ldots, r_i\}$ for r_i odd. For each $i = 1, \ldots, N$, let $\{k_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ be the sequence defined by

$$k_i^{(n)} = \begin{cases} 2n+1 & r_i \text{ odd}, \\ 2n & r_i \text{ even}, \end{cases}$$

so that fixing $|p_i - q_i| = k_i^{(n)}$ implies that conditions (i) and (iii) are met.

To satisfy (ii) we must have $r_i \le p_i + q_i$ and to satisfy (i) and (iii) we must have $|p_i - q_i| = k_i^{(n)}$. To avoid overcounting due to symmetry, we first fix possible values of p_i and consider what q_i must be. Let $\{s_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ be the sequence defined by

$$s_i^{(n)} = \frac{r_i + k_i^{(n)}}{2},$$

which we will refer to as the sequence of *starting points* for p_i corresponding to $k_i^{(n)}$. Note that the starting points $\{s_i^{(n)}\}_{n=0}^{\lfloor r_i/2 \rfloor}$ enumerate the integers between $\lceil r_i/2 \rceil$ and r_i. Picking $p_i \in \{s_i^{(n)}, s_i^{(n)} + 1, \ldots, p\}$ and $q_i = p_i - k_i^{(n)}$ we have

$$p_i + q_i = 2p_i - k_i^{(n)} \ge 2s_i^{(n)} - k_i^{(n)} = 2\left(\frac{r_i + k_i^{(n)}}{2}\right) - k_i^{(n)} = r_i$$

or $p_i + q_i \ge r_i$, so that (ii) is satisfied.

Since (i), (ii), and (iii) are satisfied by setting $p_i \in \{s_i^{(n)}, s_i^{(n)} + 1, \ldots, p\}$ and $q_i = p_i - k_i^{(n)}$ for a fixed $0 \le n \le \lfloor r_i/2 \rfloor$, we count the number of admissible pairs for

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

these choices. In $N-1$ variables, the number of polynomials of total degree less than or equal to $p-p_i$ is given by

$$\binom{N-1+p-p_i}{p-p_i},$$

where $\binom{n}{k}=0$ if $n<k$ or $k<0$. To simplify notation, pick $s_i=s_i^{(n)}$ (one of the starting points in the i-th direction) and $k_i=k_i^{(n)}$ (its associated distance), where $0\le n\le\lfloor r_i/2\rfloor$ is fixed. To count the number of admissible pairs associated with the difference k_i and starting point s_i, we compute

$$\sum_{p_i=s_i}^{p}\binom{N-1+p-p_i}{p-p_i}=\sum_{j=0}^{p-s_i}\binom{N-1+j}{j}=\binom{N+p-s_i}{p-s_i}.$$

Define $\boldsymbol{s}\in\mathbb{N}_0^N$ with the s_i as above, then $\boldsymbol{s}$ corresponds to a possible combination of starting points in each direction. To estimate the number of polynomials associated with the starting point $\boldsymbol{s}$, we compute

$$\sum_{p_1=s_1}^{p}\sum_{p_2=s_2}^{p-p_1}\cdots\sum_{p_N=s_N}^{p-p_1-\cdots-p_{N-1}}\binom{p-p_1-\cdots-p_N}{p-p_1-\cdots-p_N}=\binom{N+p-|\boldsymbol{s}|}{p-|\boldsymbol{s}|},\tag{4.4}$$

where the sum easily follows by an induction argument and Pascal's rule.

Enumerating all of the pairs $(\boldsymbol{p},\boldsymbol{q})\in\Theta_{\boldsymbol{r}}$ thus reduces to counting the number of possible combinations of starting points. Hence, in N dimensions we consider all such multi-indices of the $\{s_i^{(n)}\}_{n=0}^{\lfloor r_i/2\rfloor}$ whose components sum to some integer $\lceil|\boldsymbol{r}|/2\rceil\le\ell\le|\boldsymbol{r}|$. For two multi-indicies $\boldsymbol{s},\boldsymbol{r}\in\mathbb{N}_0^N$, we say $\boldsymbol{s}\le\boldsymbol{r}$ if and only if $s_i\le r_i$ for all $i=1,\dots,N$. Define the set $\boldsymbol{S}(\boldsymbol{r},\ell)=\left\{\boldsymbol{s}\in\mathbb{N}_0^N:|\boldsymbol{s}|=\ell,\ \boldsymbol{s}\le\boldsymbol{r}\right\}$, which corresponds to a particular slice of the desired set of starting points. To estimate $\#\boldsymbol{S}(\boldsymbol{r},\ell)$, we consider the familiar counting argument of placing N bars among ℓ stars with the added restriction that the numer of stars in the i-th bin not exceed r_i. Such a problem can be reframed in terms of finding the coefficient $c(\boldsymbol{r},\ell)$ of t^ℓ in the generating function $P_{\boldsymbol{r}}(t)=\prod_{i=1}^N\sum_{j=0}^{r_i}t^j$. Combining (4.4) and summing over ℓ between $\lceil|\boldsymbol{r}|/2\rceil\le\ell\le|\boldsymbol{r}|$ we arrive at (4.1), where the coefficients $c(\boldsymbol{r},\ell)=\#\boldsymbol{S}(\boldsymbol{r},\ell)$ when $|\boldsymbol{r}|$ is even and $\ell=|\boldsymbol{r}|/2$ (in this case the roles of p_i and q_i can not be reversed) and $c(\boldsymbol{r},\ell)=2\#\boldsymbol{S}(\boldsymbol{r},\ell)$ otherwise.

Noting that $\cup_{\ell=\lceil|\boldsymbol{r}|/2\rceil}^{|\boldsymbol{r}|}\boldsymbol{S}(\boldsymbol{r},\ell)$ is a change of coordinates of a total degree index set of order $\lceil|\boldsymbol{r}|/2\rceil$ intersected with the hyperrectangle $\{\boldsymbol{s}\in\mathbb{N}_0^N:\boldsymbol{s}\le\boldsymbol{r}\}$ yields the bound

$$\sum_{\ell=\lceil|\boldsymbol{r}|/2\rceil}^{|\boldsymbol{r}|}c(\boldsymbol{r},\ell)\le 2\sum_{\ell=\lceil|\boldsymbol{r}|/2\rceil}^{|\boldsymbol{r}|}\#\boldsymbol{S}(\boldsymbol{r},\ell)\le 2\binom{N+\lceil|\boldsymbol{r}|/2\rceil}{N}.$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

On the other hand, from the generating function $P_{\boldsymbol{r}}(t)$ we see that $c(\boldsymbol{r},\ell)$ is bounded by $\binom{|\boldsymbol{r}|}{\ell}$ when $|\boldsymbol{r}|$ is even and $\ell = |\boldsymbol{r}|/2$ and $2\binom{|\boldsymbol{r}|}{\ell}$ otherwise. This follows from the fact that when $\boldsymbol{k}$ is the multi-index having $|\boldsymbol{r}|$ ones and the rest zeros, since $\ell \le |\boldsymbol{r}|$, we have that $\#\boldsymbol{S}(\boldsymbol{r},\ell) \le \#\boldsymbol{T}(\boldsymbol{k},\ell)$ where $\boldsymbol{T}(\boldsymbol{k},\ell) = \{\boldsymbol{s} \in \mathbb{N}_0^N : |\boldsymbol{s}| = \ell, \boldsymbol{s} \le \boldsymbol{k}\}$ and $\#\boldsymbol{T}(\boldsymbol{k},\ell)$ is given by the coefficient of t^ℓ in $P_{\boldsymbol{k}}(t) = (1+t)^{|\boldsymbol{r}|} = \sum_{\ell=0}^{|\boldsymbol{r}|}\binom{|\boldsymbol{r}|}{\ell}t^\ell$ from the binomial theorem. Then

$$\sum_{\ell=\lceil|\boldsymbol{r}|/2\rceil}^{|\boldsymbol{r}|} c(\boldsymbol{r},\ell) \le 2\sum_{\ell=0}^{|\boldsymbol{r}|} c(\boldsymbol{r},\ell) = 2^{|\boldsymbol{r}|+1},$$

so that

$$\begin{aligned}\operatorname{nnz}(\mathbf{G}_{\boldsymbol{r}}) &= \sum_{\ell=\lceil|\boldsymbol{r}|/2\rceil}^{|\boldsymbol{r}|} c(\boldsymbol{r},\ell)\binom{N+p-\ell}{p-\ell} \\ &\le 2\min\left\{2^{|\boldsymbol{r}|}, \binom{N+\lceil|\boldsymbol{r}|/2\rceil}{N}\right\}\binom{N+p-\lceil|\boldsymbol{r}|/2\rceil}{N},\end{aligned}$$

showing (4.2). Substituting (4.2) into (3.18) shows the bound of $\mathcal{M}(p,r)$ from (4.3). □

We note that the bound of $\mathcal{M}(p,r)$ from (4.3) is an overestimate due to the particular form of (4.1), which is different for each $\boldsymbol{r} \in \Lambda_r$. As a consequence, we see that $\operatorname{nnz}(\mathbf{G}_{\boldsymbol{r}}) = \mathcal{O}(\min\{2^{|\boldsymbol{r}|}, M_{\lceil|\boldsymbol{r}|/2\rceil}\}M_{p-\lceil|\boldsymbol{r}|/2\rceil})$ for $\boldsymbol{r} \in \Lambda_r$. For large N and small r, 2^r is smaller than $M_{\lceil r/2\rceil}$, however, as $r \to \infty$ the bound $M_{\lceil r/2\rceil}$ is sharper. For the ε-complexity analysis in the next section, we note that

$$\mathcal{M}(p,r) \le 2\min\left\{2^r, \binom{N+\lceil r/2\rceil}{N}\right\}\binom{N+r}{N}\binom{N+p}{N}, \tag{4.5}$$

which can be seen by bounding the elements of the sum of (4.3) above by their largest values. Figure 4 plots how sharply $\mathcal{M}(p,r)$ is bounded by (4.3) and (4.5).

Corollary 4.2. *Under the same conditions in Theorem 4.1, when $\boldsymbol{r} \in \Lambda_r$ is such that $|\boldsymbol{r}| = 1$, we have*

$$\operatorname{nnz}(\mathbf{G}_{\boldsymbol{r}}) = \sum_{\ell=\lceil|\boldsymbol{r}|/2\rceil}^{|\boldsymbol{r}|} c(\boldsymbol{r},\ell)\binom{N+p-\ell}{p-\ell} = 2\binom{N+p-1}{p-1}. \tag{4.6}$$

Corollary 4.2 is the result of [14, Lemma 28], and follows from Theorem 4.1, since in this case we can use the exact formula for $\operatorname{nnz}(\mathbf{G}_{\boldsymbol{r}})$ from (4.1). Here $|\boldsymbol{r}| = \lceil|\boldsymbol{r}|/2\rceil = 1$ is odd and $\boldsymbol{S}(\boldsymbol{r},1)$ has only one element $\boldsymbol{S}(\boldsymbol{r},1) = \{\boldsymbol{s} \in \mathbb{N}_0^N : |\boldsymbol{s}| = 1, \boldsymbol{s} \le \boldsymbol{r}\} = \{\boldsymbol{r}\}$. Hence $c(\boldsymbol{r},1) = 2\#\boldsymbol{S}(\boldsymbol{r},1) = 2$, and (4.6) is shown. In the remarks that follow, we make a distinction between the cases that $a(x,\mathbf{y})$ is a polynomial of fixed

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

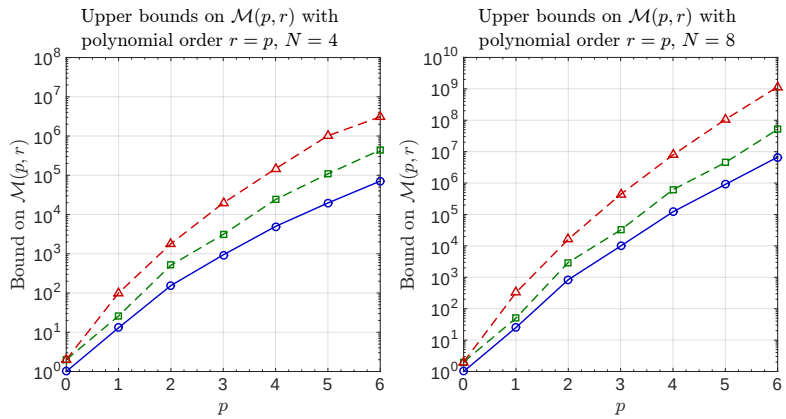

Figure 4: For $r = p$ with p ranging from $0, 1, \ldots, 6$ we plot for $N = 4$ (left) and $N = 8$ (right) the actual sparsity $\mathcal{M}(p, r)$ given by (3.18) of the Galerkin system $\mathbf{K}_r$ from (3.14) (blue), the bound on the sparsity from (4.3) (green), and the bound on the sparsity from (4.5) (red).

degree $\overline{r} < \infty$, e.g., the coefficients from Examples 2.1 and 2.2, and that $a(x, \mathbf{y})$ is a transcendental function of the random variables, e.g., the coefficient from Example 2.3.

Remark 4.3. (Complexity of matrix-vector products for polynomial coefficients, see e.g., Examples 2.1 and 2.2) *From Corollary 4.2 and the work estimate* (3.20) *when using* (3.17) *as a preconditioner, we see that for coefficients that are affine functions of the random variables, e.g., Example 2.1, the complexity of a single PCG iteration is of the order* $\mathcal{O}(J_h(2M_p + 2NM_{p-1})) = \mathcal{O}(J_h M_p)$, *where* $M_p = \#\Lambda_p^{\mathrm{TD}} = \binom{N+p}{N}$. *On the other hand, when the coefficient* $a(x, \mathbf{y})$ *is a polynomial function of the random variables, e.g., Example 2.2, having fixed order* $\overline{r} \in \mathbb{N}$, $\overline{r} < \infty$, *we use Theorem 4.1 to obtain a different estimate. Since* $\{\Psi_{\boldsymbol{r}}\}_{\boldsymbol{r}\in\Lambda_{\overline{r}}}$ *is a basis for the space* $\mathcal{P}_{\Lambda_{\overline{r}}}(\Gamma)$, *there exits coefficients* $\{a_{\boldsymbol{r}}(x)\}_{\boldsymbol{r}\in\Lambda_{\overline{r}}}$ *such that* $a(x, \mathbf{y}) = a^{\overline{r}}(x, \mathbf{y}) = \sum_{\boldsymbol{r}\in\Lambda_{\overline{r}}} a_{\boldsymbol{r}}(x)\Psi_{\boldsymbol{r}}(\mathbf{y})$. *With this representation, it is clear to see that substituting* $a(x, \mathbf{y})$ *into* (3.11) *yields* $\mathbf{K}_{\overline{r}}$ *from* (3.14), *and* $\mathbf{K}_{\overline{r}} = \mathbf{K}$ *from* (3.8). *However, it is not clear how many of the coefficients* $a_{\boldsymbol{r}}(x)$ *are identically zero. In this case, we can provide an upper bound on the block-sparsity of* $\mathbf{K}_{\overline{r}}$ *under the assumption that* $a_{\boldsymbol{r}}(x) \not\equiv 0\ \forall \boldsymbol{r} \in \Lambda_{\overline{r}}$. *Using the bound of* (4.5), *the complexity of a single matrix-vector product of* $\mathbf{K}_{\overline{r}}$ *is of the order* $\mathcal{O}(J_h M_p M_{\overline{r}} \min\{2^{\overline{r}}, M_{\lceil \overline{r}/2 \rceil}\})$. *Thus, when* $\overline{r}$ *is fixed,* $M_{\overline{r}} \min\{2^{\overline{r}}, M_{\lceil \overline{r}/2 \rceil}\}$ *is a constant, and this estimate has the same asymptotic complexity as* $\mathcal{O}(J_h M_p)$.

Remark 4.4. (Complexity of matrix-vector products in the trascendental case, see e.g., Example 2.3) *We recall the discussion of [34, Section 3.4]. There, the complexity of matrix-vector products with the SG system was estimated when a full orthogonal expansion is substituted into the SG discretization. This case corresponds to fixing the expansion order* $r = 2p$ *following Remark 3.3. Assuming that* $\mathrm{nnz}(\mathbf{G}_{\boldsymbol{r}}) = \mathcal{O}(M_p)$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

or $\mathcal{O}(M_p^2)$, it was estimated in [35] that the cost of matrix-vector products involving $\mathbf{K}_r$ is be between $\mathcal{O}(J_h M_p^2)$ and $\mathcal{O}(J_h M_p^3)$. However, the use of Theorem 4.1 allows us to consider the complexity in the case of truncating the expansion, where a sharper estimate can be obtained. Let $T_r := \prod_{k=\lceil r/2\rceil+1}^{r} \frac{N+k}{k} \ll M_{\lceil r/2\rceil} = \binom{N+\lceil r/2\rceil}{\lceil r/2\rceil}$, which is bounded independent of r, i.e.,

$$T_r \leq \left(\frac{N+\lceil r/2\rceil+1}{\lceil r/2\rceil+1}\right)^{\lceil r/2\rceil+1} \to e^N \quad as \quad r\to\infty,$$

so that $M_r = T_r M_{\lceil r/2\rceil} \leq e^N M_{\lceil r/2\rceil}$. From (4.5), we see that $\mathcal{M}(p,r)$ is of the order $\mathcal{O}(M_p M_r M_{\lceil r/2\rceil})$ as $p,r\to\infty$, since $\min\{2^r, M_{\lceil r/2\rceil}\} \to M_{\lceil r/2\rceil}$ as $r\to\infty$. When $r=2p$, this implies the complexity of matrix-vector multiplications involving $\mathbf{K}_r$ is of the order $\mathcal{O}(J_h M_p M_r M_{\lceil r/2\rceil}) = \mathcal{O}(J_h M_p T_r M_{\lceil r/2\rceil}^2) = \mathcal{O}(J_h M_p^3)$. On the other hand, when $r=p$, we see that the complexity of matrix-vector products with $\mathbf{K}_r$ is order $\mathcal{O}(J_h M_p M_r M_{\lceil r/2\rceil}) = \mathcal{O}(J_h T_r^2 M_{\lceil p/2\rceil}^3) = \mathcal{O}(J_h M_{\lceil p/2\rceil}^3)$.

4.2. ε-complexity analysis of the SGFEM. An estimate of the total complexity to obtain a fully discrete approximation of tolerance $\varepsilon > 0$ with the SGFEM and PCG solver can be shown in four steps:

1. Estimate the maximum mesh size $h_{\max}$ and minimum polynomial order $p_{\min}$ necessary in the finite element and SG discretizations, respectively,
2. If projection of the coefficient is necessary, estimate the minimum projection order $r_{\min}$, otherwise set $r_{\min} = \overline{r}$ where $\overline{r} < \infty$ is the order of the coefficient,
3. Estimate the minimum number of iterations $k_{\min}$ needed by the PCG solver,
4. Substitute $h_{\max}$, $p_{\min}$, $r_{\min}$, and $k_{\min}$ into the cost (3.20).

We proceed to estimate these parameters as follows. Denote by u^r the corresponding solution of (2.1) when $a^r(x,\mathbf{y})$ is substituted in place of $a(x,\mathbf{y})$, and let $\tilde{u}_{h,p}^r$ be the approximation to $u_{h,p}^r$ found by PCG. Then the total error for the SG approximation satisfies the following bound:

$$\left\|u-\tilde{u}_{h,p}^r\right\|_{\mathcal{H}_\varrho^2} \leq \underbrace{\|u-u^r\|_{\mathcal{H}_\varrho^2}}_{\text{SG(I)}} + \underbrace{\|u^r-u_h^r\|_{\mathcal{H}_\varrho^2}}_{\text{SG(II)}} + \underbrace{\left\|u_h^r-u_{h,p}^r\right\|_{\mathcal{H}_\varrho^2}}_{\text{SG(III)}} + \underbrace{\left\|u_{h,p}^r-\tilde{u}_{h,p}^r\right\|_{\mathcal{H}_\varrho^2}}_{\text{SG(IV)}}. \quad (4.7)$$

In this setting SG(I) is the approximation error using a truncated expansion of $a(x,\mathbf{y})$, SG(II) is the discretization error induced by the finite element method, SG(III) is the SG error coming from the orthogonal expansion (3.5), and SG(IV) is the solver error resulting from the PCG method. We note that when the projection of the coefficient is exact, as discussed in Remark 3.3, the approximation error SG(I) is no longer present and $u_{h,p}^r \equiv u_{h,p}$.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

We start with bounding SG(III). Without loss of generality, it is reasonable to assume that since u^r has a holomorphic dependence on $\mathbf{z} \in \mathbb{C}^N$ in an open neighborhood of the polyellips $\mathcal{E}_{\boldsymbol{\gamma}}$ from Theorem 2.4, then u_h^r does as well. Then, the following result, whose proof is found in [32], and immediately follows from classical spectral convergence results [9,31], describes the convergence rate of the fully discrete solutions obtained by the SG method using a total degree approximation in $\mathcal{P}_{\Lambda_p^{\mathrm{TD}}}(\Gamma)$:

Proposition 4.5 (Convergence rate for the SG method)**.** *If Theorem 2.4 holds for the solution u_h^r to* (3.1) *with coefficient $a^r(x,\mathbf{y})$, and $u_{h,p}^r$ is the solution to* (3.6) *with Λ_p the order p total degree index set, then*

$$\|u_h^r - u_{h,p}^r\|_{\mathcal{H}^\infty} \leq C_1 \exp(-C_2 p) \qquad \forall p \in \mathbb{N},$$

for some constants $C_1, C_2 > 0$ independent of p.

To investigate the error in SG(I), we note that since $a(x,\mathbf{y})$ satisfies assumption (A2), the projection error in $\mathcal{P}_{\Lambda_r^{\mathrm{TD}}}(\Gamma)$ can be similarly estimated as

$$\|a - a^r\|_{L^2_\varrho(\Gamma)\otimes L^\infty(D)} \leq C_3 \exp(-C_4 r) \qquad \forall r \in \mathbb{N}, \tag{4.8}$$

for some constants $C_3, C_4 > 0$ independent of r. Hence, $\forall r \in \mathbb{N}$,

$$\|u - u^r\|_{\mathcal{H}^2_\varrho} \leq \frac{\|f\|_{H^{-1}}}{a_{\min}^2}\|a - a^r\|_{L^2_\varrho(\Gamma)\otimes L^\infty(D)} \leq \frac{\|f\|_{H^{-1}}}{a_{\min}^2} C_3 \exp(-C_4 r) \tag{4.9}$$

providing a bound for SG(I). For a bound of SG(II), we present the following convergence result reguarding solutions to the parameterized finite element problem, whose proof can be found in a number of standard texts on the theory of finite element methods, e.g., [2, 20]:

Lemma 4.6. *Let $\mathcal{T}_h$ be a uniform finite element mesh over D with $J_h = \mathcal{O}(h^{-d})$ degrees of freedom and $h > 0$. For the elliptic PDE* (2.1) *and $\mathbf{y} \in \Gamma$, when $u^r(\mathbf{y}) \in H_0^1(D) \cap H^{s+1}(D)$, the error from the finite element approximation is bounded by*

$$\|u^r(\mathbf{y}) - u_h^r(\mathbf{y})\|_{H_0^1(D)} \leq C_{\mathrm{FEM}} h^s,$$

where the constant $C_{\mathrm{FEM}} > 0$ is independent of h and $\mathbf{y}$.

For the treatment of SG(IV), we begin by defining $\mathcal{B}^r(\mathbf{y})$ to be the corresponding bilinear operator in (2.6) with $a(x,\mathbf{y})$ replaced with $a^r(x,\mathbf{y})$. Since both $\mathcal{B}(\mathbf{y})$ and $\mathcal{B}^r(\mathbf{y})$ are symmetric, uniformly coercive and continuous bilinear operators on $H_0^1(D)$, there exist $\alpha, \beta > 0$ independent of $\mathbf{y}$ such that for every $u, v \in H_0^1(D)$

$$|\mathcal{B}^r[u,v](\mathbf{y})| = \left|\int_D a^r(x,\mathbf{y})\nabla u \cdot \nabla v dx\right| \leq \alpha \|u\|_{H_0^1(D)}\|v\|_{H_0^1(D)},$$
$$\beta\|u\|^2_{H_0^1(D)} \leq \int_D a^r(x,\mathbf{y})|\nabla u|^2 dx = \|u\|^2_{\mathcal{B}^r(\mathbf{y})},$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

and similarly for $\mathcal{B}_r(\mathbf{y})$ with the same α, β, e.g., taking α to be the maximum and β to be the minimum in each case. Recall $\mathbf{u}^r_{h,\boldsymbol{p}} = [u^r_{1,\boldsymbol{p}}, \ldots, u^r_{J_h,\boldsymbol{p}}]^{\mathrm{T}}$, the vector of nodal values of the finite element solution corresponding to the $\boldsymbol{p}$-th stochastic mode of $u^r_{h,p}$, and $\mathbf{u}^r_{h,p} = [\mathbf{u}^r_{h,\boldsymbol{p}}]^{\mathrm{T}}_{\boldsymbol{p}\in\Lambda_p}$. Then we have the following estimates expressing

$$\text{Continuity:} \qquad \left\|\mathbf{u}^r_{h,p}\right\|_{\mathbf{K}_r} = \left\|u^r_{h,p}\right\|_{\mathbb{E}[\mathcal{B}^r(\mathbf{y})]} \leq \sqrt{\alpha}\left\|u^r_{h,p}\right\|_{\mathcal{H}^2_\varrho}, \quad \text{and} \tag{4.10}$$

$$\text{Ellipticity:} \qquad \sqrt{\beta}\left\|u^r_{h,p}\right\|_{\mathcal{H}^2_\varrho} \leq \left\|u^r_{h,p}\right\|_{\mathbb{E}[\mathcal{B}^r(\mathbf{y})]} = \left\|\mathbf{u}^r_{h,p}\right\|_{\mathbf{K}_r}, \tag{4.11}$$

where $\|\mathbf{u}^r_{h,p}\|^2_{\mathbf{K}_r} = (\mathbf{u}^r_{h,p})^{\mathrm{T}}\mathbf{K}_r\mathbf{u}^r_{h,p}$ is the $\mathbf{K}_r$ matrix norm. Given Proposition 4.5, Lemma 4.6, and the estimates from (4.9), (4.10), and (4.11), we can now provide the minimal projection orders $p, r \in \mathbb{N}$ for the SG approximation (3.5) and the coefficient (3.10), respectively, the maximum mesh size h for finite element method, and the minimum number of PCG iterations k necessary to ensure that the error in the SGFEM solution $\tilde{u}^r_{h,p}$ is less than the tolerance $\varepsilon > 0$.

Lemma 4.7. *Let $u \in L^2_\varrho(\Gamma; H^1_0(D) \cap H^{s+1}(D))$ be the solution to* (2.1)*, $u^r_{h,p}$ be the solution to* (3.6) *with the coefficient $a^r(x,\mathbf{y})$, and $\tilde{u}^r_{h,p}$ be the approximation of $u^r_{h,p}$ found by PCG with a zero initial guess. Then, for $\varepsilon > 0$, to ensure that $\|u - \tilde{u}^r_{h,p}\|_{\mathcal{H}^2_\varrho} \leq \varepsilon$ we must choose $h \leq h_{\max}$, $r \geq r_{\min}$, $p \geq p_{\min}$, and $k \geq k_{\min}$, where:*

$$h_{\max} = \left(\frac{\varepsilon}{4C_{\mathrm{FEM}}}\right)^{\frac{1}{s}}, \qquad r_{\min} = \log\left[\left(\frac{4C_5}{\varepsilon}\right)^{\frac{1}{C_4}}\right],$$

$$p_{\min} = \log\left[\left(\frac{4C_1}{\varepsilon}\right)^{\frac{1}{C_2}}\right], \qquad k_{\min} = \frac{\log\left(\frac{4C_6}{\varepsilon}\right)}{\log\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}_r}+1}{\sqrt{\tilde{\boldsymbol{\kappa}}_r}-1}\right)},$$

with $C_{\mathrm{FEM}} > 0$ the constant from Lemma 4.6, $C_1, C_2, C_3, C_4 > 0$ the constants from Proposition 4.5 and (4.8)*, and, for $\alpha, \beta > 0$ from* (4.10) *and* (4.11)

$$C_5 = C_3\frac{\|f\|_{H^{-1}}}{a^2_{\min}}, \qquad C_6 = 2\sqrt{\frac{\alpha}{\beta}}\left\|u^r_{h,p}\right\|_{L^\infty},$$

with $\tilde{\boldsymbol{\kappa}}_r$ is the condition number of $\mathbf{P}^{-1}\mathbf{K}_r$ with $\mathbf{P}$ the mean-based preconditioner from (3.17)*.*

Proof. Without loss of generality, we seek to bound the quantities SG(I)-SG(IV) from (4.7) each by $\varepsilon/4$. For the error SG(I) we recall estimate (4.9) and solve for r. From Lemma 4.6, when $u \in L^2_\varrho(\Gamma; H^1_0(D) \cap H^{s+1}(D))$ we have that $\|u^r - u^r_h\|_{\mathcal{H}^2_\varrho} \leq C_{\mathrm{FEM}}h^s$ $\forall h > 0$, and from Proposition 4.5 we have that $\|u^r_h - u^r_{h,p}\|_{L^\infty} \leq C_1\exp(-C_2 p)$ $\forall p \in \mathbb{N}$, so that solving for h and p gives the desired maximum mesh size $h_{\max}$ and minimum polynomial order $p_{\min}$ to bound SG(II) and SG(III) by $\varepsilon/4$. Let $\mathbf{u}^r_{h,p}$ and $\mathbf{u}^{r,(k)}_{h,p}$ be the coefficients of the exact SG solution $u^r_{h,p}$ and the approximate SG solution $\tilde{u}^r_{h,p}$ after

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

k PCG iterations, respectively. Then from (3.16) and (4.11) we see that

$$\left\|u_{h,p}^r - \tilde{u}_{h,p}^r\right\|_{\mathcal{H}_\varrho^2} \leq \frac{1}{\sqrt{\beta}}\|\mathbf{u}_{h,p}^r - \mathbf{u}_{h,p}^{r,(k)}\|_{\mathbf{K}_r} \leq \frac{2}{\sqrt{\beta}}\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}_r}-1}{\sqrt{\tilde{\boldsymbol{\kappa}}_r}+1}\right)^k \|\mathbf{u}_{h,p}^r - \mathbf{u}_{h,p}^{r,(0)}\|_{\mathbf{K}_r},$$

where $\mathbf{u}_{h,p}^{r,(0)}$ is the initial guess used in CG and $\tilde{\boldsymbol{\kappa}}_r = \mathrm{cond}(\mathbf{P}^{-1}\mathbf{K}_r)$ with mean based preconditioner $\mathbf{P}$ from (3.17). If we use the zero vector as the initial iteration in PCG, we have from (4.10)

$$\begin{aligned}\left\|u_{h,p}^r - \tilde{u}_{h,p}^r\right\|_{\mathcal{H}_\varrho^2} &\leq \frac{2}{\sqrt{\beta}}\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}_r}-1}{\sqrt{\tilde{\boldsymbol{\kappa}}_r}+1}\right)^k \|\mathbf{u}_{h,p}^r\|_{\mathbf{K}_r} \\ &\leq 2\sqrt{\frac{\alpha}{\beta}}\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}_r}-1}{\sqrt{\tilde{\boldsymbol{\kappa}}_r}+1}\right)^k \left\|u_{h,p}^r\right\|_{\mathcal{H}_\varrho^2}.\end{aligned} \tag{4.12}$$

Solving for k gives the minimum number of iterations $k_{\min}$ required to ensure SG(IV) is bounded by $\varepsilon/4$. □

Given the necessary parameters from Lemma 4.7 to achieve $\|u - \tilde{u}_{h,p}^r\|_{\mathcal{H}_\varrho^2} \leq \varepsilon$, and the estimates on the computational complexity of one iteration in the PCG method from §4.1, we provide a bound on the minimal number of FLOPs required by the SGFEM when approximating (2.1). We split these results into the cases that the stochastic coefficient $a(x, \mathbf{y})$ from (2.1) is:

(i) an affine function of the random parameters, e.g., $a(x, \mathbf{y}) \in \mathcal{P}_{\Lambda_{\overline{r}}}(\Gamma)$ with $\overline{r} = 1$, as in Example 2.1,

(ii) a non-affine polynomial of the random parameters, e.g., $a(x, \mathbf{y}) \in \mathcal{P}_{\Lambda_{\overline{r}}}(\Gamma)$ for some $1 < \overline{r} < \infty$, as in Example 2.2,

(iii) a non-affine, transcendental function of the random parameters, e.g., $a(x, \mathbf{y}) \notin \mathcal{P}_{\Lambda_r}(\Gamma)$ for any $r \in \mathbb{N}$, as in Example 2.3, so that r must be chosen to satisfy $r \geq r_{\min}$ from Lemma 4.7.

The results are summarized in Theorems 4.8 and 4.9 next.

Theorem 4.8. *Let $u \in L_\varrho^2(\Gamma; H_0^1(D) \cap H^{s+1}(D))$ be the solution to* (2.1)*, and $\overline{r}$ be the smallest natural number such that $a(x, \mathbf{y}) \in \mathcal{P}_{\Lambda_{\overline{r}}}(\Gamma)$. When $\overline{r} = 1$, the minimum work* (3.20) *of solving* (3.14) *with PCG to a tolerance $\varepsilon > 0$ can be bounded by*

$$\begin{aligned}W_{\mathrm{A}}^{\mathrm{pSG}} \leq\ & C_7\left(\frac{3C_{\mathrm{FEM}}}{\varepsilon}\right)^{\frac{d}{s}} 2e^N(1+N) \\ & \times\left(1+\log\left[\left(\frac{3C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N \left(\frac{\log\left(\frac{3C_6}{\varepsilon}\right)}{\log\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}}+1}{\sqrt{\tilde{\boldsymbol{\kappa}}}-1}\right)}\right),\end{aligned} \tag{4.13}$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

and when $\overline{r} > 1$*, the minimum work* (3.20) *of solving* (3.14) *with PCG to a tolerance* $\varepsilon > 0$ *can be bounded by*

$$W_{\text{NAP}}^{\text{pSG}} \leq C_7 \left(\frac{3C_{\text{FEM}}}{\varepsilon}\right)^{\frac{d}{s}} \left(\frac{\log\left(\frac{3C_6}{\varepsilon}\right)}{\log\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}}+1}{\sqrt{\tilde{\boldsymbol{\kappa}}}-1}\right)}\right) e^N \left[\left(1 + \log\left[\left(\frac{3C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N \right.$$
$$\left. +2\sum_{j=0}^{\overline{r}} \binom{N-1+j}{N-1} \min\left\{2^j, \binom{N+\lceil j/2\rceil}{N}\right\} \left(1 - \frac{\lceil j/2\rceil}{N} + \log\left[\left(\frac{3C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N\right], \tag{4.14}$$

with $C_{\text{FEM}}, C_1, C_2, C_6$ *as in Lemma 4.7,* $C_7 > 0$ *independent of* ε*, and* $\tilde{\boldsymbol{\kappa}}$ *the condition number of the preconditioned system* $\mathbf{P}^{-1}\mathbf{K}_{\overline{r}} = \mathbf{P}^{-1}\mathbf{K}$*, using the mean-based preconditioner from* (3.17)*.*

Proof. When $a(x, \mathbf{y}) \in \mathcal{P}_{\Lambda_{\overline{r}}}(\Gamma)$ we do not need to consider SG(I) from (4.7), and bound SG(II), SG(III), and SG(IV) by $\varepsilon/3$. Hence, to minimize the error of the SG discretization, we choose $p \geq p_{\min} = \log[(3C_1/\varepsilon)^{1/C_2}]$ which differs from the $p_{\min}$ stated in Lemma 4.7. For a uniform triangulation $\mathcal{T}_h$, $J_h = \mathcal{O}(h^{-d})$ so that

$$J_{h_{\max}} = C_7 \left[\left(\frac{\varepsilon}{3C_{\text{FEM}}}\right)^{\frac{1}{s}}\right]^{-d} = C_7 \left(\frac{3C_{\text{FEM}}}{\varepsilon}\right)^{\frac{d}{s}} \tag{4.15}$$

for some constant $C_7 > 0$ depending on the connectivity of the finite element mesh, but independent of ε. In the case that $\overline{r} = 1$, we substitute $p_{\min}$ into (4.1) for the matrices $\mathbf{G}_{\boldsymbol{r}}$ having $0 \leq |\boldsymbol{r}| \leq 1$, and apply Stirling's approximation to obtain

$$\begin{aligned} M_{p_{\min}} + \mathcal{M}(p_{\min}, 1) &= 2\binom{N+p_{\min}}{N} + 2N\binom{N+p_{\min}-1}{N} \\ &\leq 2e^N(1+N)\left(1 + \log\left[\left(\frac{3C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N, \end{aligned}$$

Similarly, when $\overline{r} > 1$ we use the bound from (4.3) and Stirling's approximation to obtain

$$M_{p_{\min}} + \mathcal{M}(p_{\min}, \overline{r}) \leq e^N \left[\left(1 + \log\left[\left(\frac{3C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N \right.$$
$$\left. +2\sum_{j=0}^{\overline{r}} \min\left\{2^j, \binom{N+\lceil j/2\rceil}{N}\right\} \binom{N-1+j}{N-1} \left(1 - \frac{\lceil j/2\rceil}{N} + \log\left[\left(\frac{3C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N\right].$$

Substituting $J_{h_{\max}}$ for J_h, $k_{\min}$ for $N_{\text{iter}}^{\text{SG}}$ from Lemma 4.7, and the bounds for $M_{p_{\min}} +$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

$N_K^{(p_{\min},\overline{r})}$ into the work estimate (3.20), in the cases $\overline{r}=1$ and $\overline{r}>1$ above, we obtain the desired results. □

Theorem 4.9. *Let $u \in L^2_\varrho(\Gamma; H^1_0(D)\cap H^{s+1}(D))$ be the solution to* (2.1)*, and suppose that $a(x,\mathbf{y}) \notin \mathcal{P}_{\Lambda_r}(\Gamma)$ for any $r\in\mathbb{N}$. In this case r must be chosen to satisfy $r \geq r_{\min}$ from Lemma 4.7. Then the minimum work* (3.20) *of solving* (3.14) *with PCG to a tolerance $\varepsilon>0$ can be bounded by*

$$W_{\mathrm{NAT}}^{\mathrm{pSG}} \leq C_7 \left(\frac{4C_{\mathrm{FEM}}}{\varepsilon}\right)^{\frac{d}{s}} e^N \left(1+\log\left[\left(\frac{4C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N \tag{4.16}$$
$$\left[1+2e^{2N}\left(1+\log\left[\left(\frac{4C_5}{\varepsilon}\right)^{\frac{1}{C_4 N}}\right]\right)^N \left(1+\frac{1}{N}+\log\left[\left(\frac{4C_5}{\varepsilon}\right)^{\frac{1}{2C_4 N}}\right]\right)^N\right]\left(\frac{\log\left(\frac{4C_6}{\varepsilon}\right)}{\log\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}_r}+1}{\sqrt{\tilde{\boldsymbol{\kappa}}_r}-1}\right)}\right),$$

with $C_{\mathrm{FEM}}, C_1, C_2, C_4, C_5, C_6$ as in Lemma 4.7, $C_7>0$ independent of ε, and $\tilde{\boldsymbol{\kappa}}_r$ the condition number of the preconditioned system $\mathbf{P}^{-1}\mathbf{K}_r$, using the mean-based preconditioner from (3.17)*.*

Proof. In this setting r must be chosen to satisfy $r \geq r_{\min} = \log[(4C_5/\varepsilon)^{1/C_4}]$ from Lemma 4.7 and, therefore, we must bound the sum from (4.3) which now depends on $r_{\min}$, and hence on ε. Thus, we use the bound (4.5) for $\mathcal{M}(p,r)$, noting that as $\varepsilon\to 0$, $r_{\min}\to\infty$ so that

$$\min\left\{2^{r_{\min}}, \binom{N+\lceil r_{\min}/2\rceil}{N}\right\} = \binom{N+\lceil r_{\min}/2\rceil}{N}.$$

Substituting $p_{\min}$ and $r_{\min}$ from Lemma 4.7 into (4.5) and applying Stirling's approximation, we obtain

$$M_{p_{\min}} + \mathcal{M}(p_{\min}, r_{\min}) \leq e^N \left(1+\log\left[\left(\frac{4C_1}{\varepsilon}\right)^{\frac{1}{C_2 N}}\right]\right)^N$$
$$\left[1+2e^{2N}\left(1+\log\left[\left(\frac{4C_5}{\varepsilon}\right)^{\frac{1}{C_4 N}}\right]\right)^N \left(1+\frac{1}{N}+\log\left[\left(\frac{4C_5}{\varepsilon}\right)^{\frac{1}{2C_4 N}}\right]\right)^N\right].$$

As in the proof of Theorem 4.8, we substitute $J_{h_{\max}}$ for J_h from (4.15), $k_{\min}$ for $N_{\mathrm{iter}}^{\mathrm{SG}}$, and the bound for $M_{p_{\min}}+\mathcal{M}(p_{\min},r_{\min})$ with $p_{\min}$ and $k_{\min}$ from Lemma 4.7 into the cost (3.20) to complete the proof. □

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Given Theorems 4.8 and 4.9 we see that the work of obtaining the fully discrete approximation using the SGFEM, with PCG as a solver, is asymptotically given by:

$$\underbrace{\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{\frac{d}{s}}}_{\text{(SG.1)}}\underbrace{\left[\log\left(\frac{1}{\varepsilon}\right)\right]^{g(N)}}_{\text{(SG.2)}}\underbrace{\left(\frac{\log\left(\frac{1}{\varepsilon}\right)}{\log\left(\frac{\sqrt{\tilde{\boldsymbol{\kappa}}_r}+1}{\sqrt{\tilde{\boldsymbol{\kappa}}_r}-1}\right)}\right)}_{\text{(SG.3)}}, \tag{4.17}$$

where $g(N) = N$ and $\tilde{\boldsymbol{\kappa}}_r = \tilde{\boldsymbol{\kappa}}$ if $a(x, \mathbf{y})$ is an affine or non-affine, polynomial function of the random parameters of fixed order $\overline{r} < \infty$, e.g., Examples 2.1 and 2.2, and $g(N) = 3N$ when $a(x, \mathbf{y})$ is a non-affine, transcendental function of the random parameters, e.g., Example 2.3, requiring a total degree orthogonal expansion of order $r \geq r_{\min}$ depending on ε. Here, (SG.1), (SG.2), and (SG.3) correspond to the work required by the finite element, SG, and PCG methods, respectively. In particular, (SG.2) corresponds to the estimates for the sparsity of the Galerkin system $\mathbf{K}_r$ from (3.14), and represents the number of coupled finite element systems that must be solved simultaneously by the PCG method. However, due to the bound (4.3), the asymptotic complexity in the cases that $a(x, \mathbf{y})$ is affine or polynomial in $\mathbf{y}$ are the same. This does not imply that there is no need to consider the work estimates in these cases separately. Indeed, if $a(x, \mathbf{y})$ is a polynomial having the representation $\sum_{\boldsymbol{r}\in\Lambda_{\overline{r}}} a_{\boldsymbol{r}}(x)\Psi_{\boldsymbol{r}}(\mathbf{y})$ where $a_{\boldsymbol{r}}(x) \neq 0$ for all $\boldsymbol{r} \in \Lambda_{\overline{r}}$, then the complexity of matrix-vector multiplications with $\mathbf{K}_r$ is of the order $\mathcal{O}(J_h M_p M_{\overline{r}} \min\{2^{\overline{r}}, M_{\lceil \overline{r}/2 \rceil}\})$. Here, the constant $M_{\overline{r}} \min\{2^{\overline{r}}, M_{\lceil \overline{r}/2 \rceil}\}$ grows rapidly with $\overline{r}$, suggesting that higher order polynomial functions of $\mathbf{y}$ require additional cost.

4.3. Conditioning of the generalized SG system. In this section, we discuss issues related to the conditioning of the linear system that results from the SGFEM discretization. We first recall [14, Theorem 10]: the eigenvalues of the matrices $\{\mathbf{G}_{\boldsymbol{r}}\}_{\boldsymbol{r}\in\Lambda_r}$ from (3.13) lie in the interval $[\xi_{\boldsymbol{r}}, \Xi_{\boldsymbol{r}}]$, where

$$\xi_{\boldsymbol{r}} := \min\{\Psi_{\boldsymbol{r}}(\mathbf{y}) : \mathbf{y} \in \mathcal{G}^{\boldsymbol{m}(\mathbf{l})}\}, \qquad \Xi_{\boldsymbol{r}} := \max\{\Psi_{\boldsymbol{r}}(\mathbf{y}) : \mathbf{y} \in \mathcal{G}^{\boldsymbol{m}(\mathbf{l})}\}, \tag{4.18}$$

$\mathcal{G}^{\boldsymbol{m}(\mathbf{l})}$ is a tensor product grid of Gauss-Legendre quadrature points having $\boldsymbol{m}(\mathbf{l}) = (m(l_1), \ldots, m(l_n))$ points in each direction, and $\mathbf{l}$ is such that $m(l_n) := p + \lceil \frac{k_n+1}{2} \rceil$, $n = 1, \ldots, N$. Since $a^r(x, \mathbf{y})$ satisfies (A1), the analysis of [29, Theorem 3.8] shows that the eigenvalues for the preconditioned system $\mathbf{P}^{-1}\mathbf{K}_r$ lie in the interval $[1 - \underline{\tau}_r, 1 + \overline{\tau}_r]$ where

$$\underline{\tau}_r = \frac{1}{a_{\min}} \sum_{\substack{\boldsymbol{r}\in\Lambda_r \\ |\boldsymbol{r}|\neq 0}} \xi_{\boldsymbol{r}} \|a_{\boldsymbol{r}}(x)\|_{L^\infty(D)}, \qquad \overline{\tau}_r = \frac{1}{a_{\min}} \sum_{\substack{\boldsymbol{r}\in\Lambda_r \\ |\boldsymbol{r}|\neq 0}} \Xi_{\boldsymbol{r}} \|a_{\boldsymbol{r}}(x)\|_{L^\infty(D)}. \tag{4.19}$$

As a result of (4.19), we see that in the case that the projection order r of the coefficient $a^r(x, \mathbf{y})$ depends on ε, the condition number of the preconditioned system

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

$\mathbf{P}^{-1}\mathbf{K}_r$ does as well through the number of terms in $\underline{\tau}_r$ and $\overline{\tau}_r$. This should come as no surprise since even in the case of the Karhunen-Loève expansion, the condition number of $\mathbf{P}^{-1}\mathbf{K}$ depends on the number of terms in the truncated Karhunen-Loève expansion which is chosen a-priori to minimize the error.

5. Comparison with the SCFEM. In this section we compare our explicit cost bounds for the SGFEM with the complexity estimates for the SCFEM developed in [16], when solving (2.1). The basic idea behind the SCFEM is to construct a fully discrete approximation in a subspace of $V_h(D) \otimes L^2_\varrho(\Gamma)$ by collocating semi-discrete solutions u_h from (3.1) on a deterministic set of points to obtain solutions $\{u_h(\cdot, \mathbf{y}_k)\}_{k=1}^{M_L} \in V_h(D)$.

5.1. A generalized SCFEM using Lagrange interpolation. To construct the stochastic collocation (SC) approximation, we consider a class of multi-index sets defined in terms of increasing functions $\boldsymbol{m} : \mathbb{N}_+^N \to \mathbb{N}_+^N$ and $g : \mathbb{N}_+^N \to \mathbb{N}_+$. By $\boldsymbol{m}$ we specify the multivariate function $\boldsymbol{m}(\mathbf{l}) := (m_1(l_1), \cdots, m_N(l_N))$ where each $m_n : \mathbb{N}_+ \to \mathbb{N}_+$ is an increasing function, possibly different for each $n = 1, \ldots, N$. Here the m_n are referred to as growth functions, specifying how the number of points grows in the direction n. Associated with m_n we define the *left-inverse* $m_n^\dagger : \mathbb{N}_+ \to \mathbb{N}_+$ by $m_n^\dagger(q) = \min\{k \in \mathbb{N}_+ : m_n(k) \geq q\}$, and let $\boldsymbol{m}^\dagger(\boldsymbol{q}) = (m_1^\dagger(q_1), \ldots, m_N^\dagger(q_N))$. In this case, we note that $m_n^\dagger(m_n(k)) = k$ and $m_n(m_n^\dagger(k)) \geq k$ for each $k \in \mathbb{N}_+$ and $n = 1, \ldots, N$. Given $\boldsymbol{m}$ and g we can define the multi-index set

$$\Lambda_L^{\boldsymbol{m},g} = \left\{\boldsymbol{q} \in \mathbb{N}_+^N : g(\boldsymbol{m}^\dagger(\boldsymbol{q}+\mathbf{1})) \leq L\right\}, \tag{5.1}$$

to be used in constructing polynomial approximations. In particular, setting $m_n(j) = j$ for all $j \in \mathbb{N}_+$ and $n = 1, \ldots, N$, and defining

$$g_{\mathrm{TP}}(\boldsymbol{p}) = \max_{1\leq n\leq N} p_n, \qquad g_{\mathrm{TD}}(\boldsymbol{p}) = \sum_{n=1}^N (p_n - 1), \qquad g_{\mathrm{SM}}(\boldsymbol{p}) = \sum_{n=1}^N f(p_n), \tag{5.2}$$

where $f(p)$ is given in (3.4), and using the definition of $\Lambda_L^{\boldsymbol{m},g}$ from (5.1), we obtain the TP, TD, and SM index sets Λ_L^{TP}, Λ_L^{TD}, and Λ_L^{SM}, respectively, given in (3.4).

We introduce a sequence of one-dimensional Lagrange interpolation operators $\mathcal{U}^{m_n(l_n)} : C^0(\Gamma_n) \to \mathcal{P}_{m_n(l_n)-1}(\Gamma_n)$. Then for $v \in C^0(\Gamma)$ the generalized multi-dimensional approximation operator $\mathcal{I}_L^{\boldsymbol{m},g} : C^0(\Gamma) \to \mathcal{P}_{\Lambda_L^{\boldsymbol{m},g}}(\Gamma)$ is given by

$$\mathcal{I}_L^{\boldsymbol{m},g}[v](\mathbf{y}) = \sum_{g(\mathbf{l})\leq L} \sum_{\mathbf{i}\in\{0,1\}^N} (-1)^{|\mathbf{i}|} \left(\bigotimes_{n=1}^N \mathcal{U}_n^{m_n(l_n-i_n)}\right)[v](\mathbf{y}). \tag{5.3}$$

Construction of the approximation $\mathcal{I}_L^{\boldsymbol{m},g}[v](\mathbf{y})$ requires the independent evaluation of samples $v(\mathbf{y})$ on a deterministic set of distinct collocation points $\mathcal{G}_L^{\boldsymbol{m},g}$ having cardinality $M_L = \#\mathcal{G}_L^{\boldsymbol{m},g}$. Applying $\mathcal{I}_L^{\boldsymbol{m},g}[\cdot]$ from (5.3) to the semi-discrete solution

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

$u_h(x, \mathbf{y})$ of problem (3.1), we obtain the fully discrete SC approximation

$$u_{h,L}(x, \mathbf{y}) = \mathcal{I}_L^{\boldsymbol{m},g}[u_h](x, \mathbf{y}). \tag{5.4}$$

One-dimensional abscissas In this effort, we use three examples for constructing the fully discrete approximation. The first is that of a fully-nested rule constructed on the Clenshaw-Curtis choice of abscissas [5] with function $g_{\mathrm{TD}}(\boldsymbol{p})$ from (5.2) and an isotropic growth rule $\boldsymbol{m} = (m, \ldots, m)$ with m given by

$$m(1) = 1, \quad m(l_n) = 2^{l_n - 1} + 1 \quad \text{for} \quad l_n > 1, \tag{5.5}$$

This is the classical Smolyak sparse-tensorization construction [30], and here the choice of $\boldsymbol{m}$ corresponds to a doubling growth rule that leads to a nested sequence of multi-dimensional grids, e.g., $\mathcal{G}_L^{\boldsymbol{m},g} \subset \mathcal{G}_{L+1}^{\boldsymbol{m},g}$. On the other hand, we can construct a sparse-Smolyak approximation on the Gauss-Legendre abscissas corresponding to the zeros of the Legendre polynomials $\{\Psi_{\boldsymbol{p}}\}$, as defined in §3. When the points are grown isotropically according to the linear growth rule with $\boldsymbol{m} = (m, \ldots, m)$ and m defined as

$$m(l_n) = l_n \qquad \text{for} \quad l_n \in \mathbb{N}, \tag{5.6}$$

and $g_{\mathrm{TD}}(\boldsymbol{p})$ from (5.2), we obtain a grid that is not nested. Another construction that yields a sequence of nested grids is that based on the Leja points, defined as the sequence of points satisfying $y_{k+1} := \operatorname{argmax}_{y \in \Gamma_n} \prod_{j=1}^{k} |y - y_j|$ (see [10]). Here we take the Leja sequence of points with g_{TD} from (5.2) and the isotropic linear growth function $\boldsymbol{m}$ from (5.6).

5.2. Cost of solving the SCFEM systems. To construct the fully discrete approximation with the SCFEM, we must solve M_L distinct decoupled finite element systems, each dependent on a realization of the parameters $\mathbf{y}_k \in \mathcal{G}_L^{\boldsymbol{m},g}$ for $k = 1, \ldots, M_L$. Similar to the SGFEM, we can apply the PCG method to the solution of each system. Let $N_{\mathrm{iter}}^{(k)}$ be the number of iterations required by the CG method to solve the finite element system corresponding to $\mathbf{y}_k$ and $N_{\mathrm{iter}}^{\mathrm{p}(k)}$ be the corresponding number of iterations when a preconditioner is used. We are interested in choosing a suitable preconditioning strategy to decrease the total number of iterations $N_{\mathrm{iter}}^{\mathrm{pSC}} = \sum_{k=1}^{M_L} N_{\mathrm{iter}}^{\mathrm{p}(k)}$ required to obtain the fully discrete approximation $u_{h,L}$. We present a preconditioning strategy of choosing

$$\mathbf{P}_0 := \mathbf{A}(\mathbf{y}_1), \tag{5.7}$$

with $\mathbf{A}(\mathbf{y})$ from (3.2) the finite element stiffness matrix corresponding to the sample point $\mathbf{y}_1 \in \mathcal{G}_L^{\boldsymbol{m},g}$, as the preconditioner for all of the individual finite element solutions. We refer to this choice of preconditioner as the level-zero preconditioner since it corresponds to the SC approximation at level $L = 0$.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Since we apply CG to the solution of each individual finite element system, the work in floating point operations (FLOPs) required to obtain a fully discrete approximation with the SCFEM without a preconditioner is given by

$$W^{\mathrm{SC}} \approx \mathcal{O}\left(J_h * \sum_{k=1}^{M_L} N_{\mathrm{iter}}^{(k)} \right). \tag{5.8}$$

On the other hand, the "level-zero" preconditioner induces an an additional matrix-vector product requiring $\mathcal{O}(J_h)$ FLOPs per iteration, hence the work of solving (2.5) with PCG is given by

$$W^{\mathrm{pSC}} \approx 2 * \mathcal{O}\left(J_h * \sum_{k=1}^{M_L} N_{\mathrm{iter}}^{\mathrm{p}(k)} \right). \tag{5.9}$$

Here, the reduction in work due to preconditioning will be seen in the number of iterations saved in each individual count $N_{\mathrm{iter}}^{\mathrm{p}(k)}$ contributing to the sum.

5.3. Comparing the explicit cost bounds of the SGFEM and SCFEM. Given a particular "sparse" index set Λ_p, we can find increasing functions $\boldsymbol{m} : \mathbb{N}_+^N \to \mathbb{N}_+^N$ and $g : \mathbb{N}_+^N \to \mathbb{N}_+$, and $L \in \mathbb{N}$ such that $\Lambda_p = \Lambda_L^{\boldsymbol{m},g}$ from (5.1). In this setting, we can either use Galerkin projection or construct an interpolant to obtain an approximation to u in $\mathcal{P}_{\Lambda_p}(\Gamma)$. Let u_{Λ_p} denote the Galerkin projection of u onto the space $\mathcal{P}_{\Lambda_p}(\Gamma)$. Then in the $L^2_\varrho(\Gamma)$ norm, we have the estimate

$$\|u - u_{\Lambda_p}\|_{L^2_\varrho(\Gamma)} \leq C_a \min_{v \in \mathcal{P}_{\Lambda_p}(\Gamma)} \|u - v\|_{L^2_\varrho(\Gamma)}$$

where $C_a > 0$ depends on the coefficient $a(x, \mathbf{y})$ and the bounds from assumption (A1). This estimate expresses optimality in the $L^2_\varrho(\Gamma)$ error of the Galerkin projection since C_a does not grow with Λ_p, and suggests that the Galerkin method is the best choice for approximating u in the space $\mathcal{P}_{\Lambda_p}(\Gamma)$. We can also define an interpolation operator $\mathcal{I}_L^{\boldsymbol{m},g} : C^0(\Gamma) \to \mathcal{P}_{\Lambda_p}(\Gamma)$, and then we have the estimate

$$\begin{aligned} \|u - \mathcal{I}_L^{\boldsymbol{m},g}[u]\|_{L^\infty(\Gamma)} &\leq (C_{\Lambda_L} + 1) \min_{v \in \mathcal{P}_{\Lambda_p}(\Gamma)} \|u - v\|_{L^\infty(\Gamma)} \\ &= (C_{\Lambda_L} + 1) \|u - u_{\Lambda_p}\|_{L^\infty(\Gamma)} \end{aligned} \tag{5.10}$$

where $C_{\Lambda_L} = \|\mathcal{I}_L^{\boldsymbol{m},g}\|_{L^\infty(\Gamma)}$ is the Lebesgue constant of $\mathcal{I}_L^{\boldsymbol{m},g}$. A good interpolant will be one for which C_{Λ_L} grows moderately with $\#\Lambda_L^{\boldsymbol{m},g}$. For example, it is known (see [11,16]) that for a one-dimensional Lagrange interpolation operator using a Clenshaw-Curtis rule, the Lebesgue constant is bounded by $\frac{2}{\pi}\log(m-1) + 1$, where m is the number of points. For the SC method, we define SDOF to be the total number of points needed to construct the approximation. From (5.10), if we only consider the number of SDOF needed to represent the solution, we expect the error for the

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Galerkin approximation to be much lower than the error in the interpolant. Indeed, this is reflected in our numerical results in Figures 6 and 8, and has been observed in previous comparisons [3, 13].

However, if we are willing to change the space Λ_p, e.g., adding more interpolation points to gain a more stable interpolant by changing $\boldsymbol{m}$ or changing which points are included in the set $\Lambda_L^{\boldsymbol{m},g}$ by changing g, it might be possible to obtain an approximation with lower complexity to reach a given tolerance, despite having to solve more systems. Therefore, to properly compare the work involved in constructing u_{Λ_p} and $\mathcal{I}_L^{\boldsymbol{m},g}[u]$, we consider the computational complexity of both methods, not in terms of SDOF, but in terms of floating point operations (FLOPs). For a chosen Λ_p, this reduces to studying the complexity of the system resulting from Galerkin projections and the stability properties of the interpolant $\mathcal{I}_L^{\boldsymbol{m},g}$.

Let $\tilde{u}_{h,L}$ denote the numerical solution to the fully discrete approximation $u_{h,L}$ obtained with the SCFEM from (5.4) found by the PCG, and observe that we have a similar splitting to (4.7) for the error in the approximation

$$\|u-\tilde{u}_{h,L}\|_{\mathcal{H}_\varrho^2} \le \underbrace{\|u-u_h\|_{\mathcal{H}_\varrho^2}}_{\text{SC(I)}} + \underbrace{\|u_h-u_{h,L}\|_{\mathcal{H}_\varrho^2}}_{\text{SC(II)}} + \underbrace{\|u_{h,L}-\tilde{u}_{h,L}\|_{\mathcal{H}_\varrho^2}}_{\text{SC(III)}} \tag{5.11}$$

Note that unlike in the case of the SGFEM, the SCFEM does not require a further projection of the coefficient $a(x,\mathbf{y})$, so that we do not need to consider the error $\|u-u^r\|_{\mathcal{H}_\varrho^2}$ from (4.7). In addition, we do not need to worry about well-posedness of the truncation as discussed in Remark 3.3. Similar to the complexity analysis for the SGFEM, we must choose $h \le h_{\max}$ and $L \ge L_{\min}$ so that the errors $\|u-u_h\|_{\mathcal{H}_\varrho^2}$ from the finite element discretization and $\|u_h-u_{h,L}\|_{\mathcal{H}_\varrho^2}$ from the SC interpolation are both bounded by $\varepsilon/3$. From this, a minimum tolerance $\tau_{\min}$ for the PCG solver can be derived and the maximum number of PCG iterations, with a zero initial guess, can be estimated [16]. In what follows, we present a result, whose proof can be found in [16, Theorem 4.7] that bounds the number of PCG iterations in the context of the work estimate (5.9). Using this estimate we can compare the cost in FLOPs for the SCFEM with the SGFEM results from Theorems 4.8 and 4.9 in the previous section.

Theorem 5.1. *Let $u \in L_\varrho^2(\Gamma; H_0^1(D)\cap H^{s+1}(D))$ be the solution to* (2.1)*. Then for $\varepsilon > 0$ arbitrary, the work of finding $\tilde{u}_{h,L}$, the approximation to the fully discrete SC solution $u_{h,L}$ from* (5.4) *found by PCG, denoted by W^{pSC}, can be bounded by*

$$\begin{aligned} W^{\mathrm{pSC}} \le{}& 2C_7\left(\frac{3C_{\mathrm{FEM}}}{\varepsilon}\right)^{\frac{d}{s}} C_8\left[\log\left(\frac{3C_{\mathrm{SC}}}{\varepsilon}\right)\right]^N \left[C_9+\frac{1}{\log 2}\log\log\left(\frac{3C_{\mathrm{SC}}}{\varepsilon}\right)\right]^{N-1} \\ &\times \frac{1}{\log\left(\frac{\sqrt{\bar{\kappa}}+1}{\sqrt{\bar{\kappa}}-1}\right)}\left\{\log\left(\frac{C_{10}}{\varepsilon}\right)+C_{11}+2N\log\log\left[\frac{1}{rN}\log\left(\frac{3C_{\mathrm{SC}}}{\varepsilon}\right)\right]\right\}. \end{aligned} \tag{5.12}$$

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

Here C_{FEM} from Lemma 4.6, C_7 from Theorem 4.8, and C_8, C_9, C_{10}, C_{11}, C_{SC}, and r from [16, Theorem 4.7] are positive constants independent of ε. Moreover, we define $\bar{\kappa} = \sup_{\mathbf{y}\in\Gamma} \kappa(\mathbf{y})$ where $\kappa(\mathbf{y})$ is the condition number of the preconditioned system $\mathbf{P}_0^{-1}\mathbf{A}(\mathbf{y})$ and $\mathbf{P}_0$ is given in (5.7)*, noting that $\bar{\kappa}$ is also independent of ε.*

Theorem 5.1 follows from the fact that $N_{\text{iter}}^{\text{pSC}} = \sum_{k=1}^{M_L} N_{\text{iter}}^{\text{p}(k)} \leq N_{\text{zero}}$, where N_{zero} is the number of iterations needed by the SCFEM with a PCG and a zero vector initial guess. Substituting the bound on N_{zero} shown in [16, Theorem 4.7] into the work estimate (5.9) and using $J_{h_{\max}} = C_7(3C_{\text{FEM}}/\varepsilon)^{d/s}$ as in Theorem 4.8, puts the result in terms of FLOPs. Given Theorem 5.1 we see that the work of obtaining the fully discrete approximation with the SCFEM with the PCG method is asymptotically bounded by:

$$\underbrace{\mathcal{O}\left(\frac{1}{\varepsilon}\right)^{\frac{d}{s}}}_{\text{(SC.1)}} \underbrace{\left[\log\left(\frac{1}{\varepsilon}\right)\right]^{N} \left[\log\log\left(\frac{1}{\varepsilon}\right)\right]^{N-1}}_{\text{(SC.2)}} \underbrace{\left(\frac{\log\left(\frac{1}{\varepsilon}\right)}{\log\left(\frac{\sqrt{\bar{\kappa}}+1}{\sqrt{\bar{\kappa}}-1}\right)}\right)}_{\text{(SC.3)}} \tag{5.13}$$

where (SC.1), (SC.2), and (SC.3) correspond to the work required by the finite element, SC interpolant, and PCG methods, respectively. Since the costs associated with the finite element discretization are the same for both methods, we focus only on the costs associated with the SG projection and the SC interpolation, coupled with the costs of the PCG method. In particular, as the work required by the SG approximation from (SG.2) of (4.17) has different bounds depending on whether the coefficient is a fixed order polynomial or is given by a total degree orthogonal expansion having order depending on ε, we now provide a comparison in both of these cases.

Comparison in the affine and non-affine polynomial cases, e.g., Examples 2.1 and 2.2. The terms (SG.2) from (4.17) and (SC.2) from (5.13) are asymptotic estimates of the number of coupled and decoupled finite element systems that must be solved by the SGFEM and SCFEM to construct the stochastic approximation, respectively. For the coefficients from Examples 2.1 and 2.2, (SG.2) from (4.17) for the SGFEM is $\mathcal{O}([\log(1/\varepsilon)]^N)$. Hence, our analysis shows that in these cases, the number of coupled finite element systems in the SGFEM matrix is smaller than the number of decoupled finite element systems required by the SCFEM by a factor of $(\log\log(1/\varepsilon))^{N-1}$. This difference is enough to suggest that, if the condition numbers of both the preconditioned coupled system from the SGFEM and the preconditioned individual systems from the SCFEM are of the same order, then the SGFEM will outperform the SCFEM in terms of minimum work required to obtain a fully discrete approximation. In fact, whenever $a(x, \mathbf{y})$ is a general non-affine, polynomial coefficient, e.g., Example 2.2, having fixed order $\bar{r} < \infty$, the complexity of matrix-vector products involving $\mathbf{K}$ from (3.8) is approximately $\mathcal{O}(J_h M_p M_{\bar{r}} \min\{2^{\bar{r}}, M_{\lceil \bar{r}/2 \rceil}\})$ from Remark 4.3. Hence, our analysis shows that when $\mathcal{O}(M_{\bar{r}} \min\{2^{\bar{r}}, M_{\lceil \bar{r}/2 \rceil}\}) < \mathcal{O}((\log\log(1/\varepsilon))^{N-1})$, which,

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

in limit as $\varepsilon \to 0$, is always the case, and the condition numbers of both systems are of the same order, the SGFEM will outperform the SCFEM. However, in practical applications, it may require unrealistically small tolerance ε to see this when $\overline{r}$ is large.

Comparison in the non-affine, transcendental case, e.g., Example 2.3. In the case that $a(x, \mathbf{y})$ is a non-affine, transcendental functon of the random parameters, e.g., Example 2.3, the estimate for (SG.2) is $\mathcal{O}([\log(1/\varepsilon)]^{3N})$. In this case, our analysis shows that the number of coupled finite element matrices present in the SG system $\mathbf{K}_r$ from (3.14) dominates the number of decoupled finite element systems needed by the SCFEM by a factor of $[\log(1/\varepsilon)]^{2N}$. From this, even if the condition numbers of both the preconditioned coupled SG system and the preconditioned decoupled SCFEM systems are of the same order, it becomes difficult to see how the SGFEM can compete with the SCFEM.

Note that, as in the case of the SGFEM, the term (SC.3) has a dependence on the condition numbers of the preconditioned finite element systems through the bound $\bar{\kappa} = \sup_{\mathbf{y}\in\Gamma} \kappa(\mathbf{y})$. For the unpreconditioned systems $\mathbf{A}(\mathbf{y})$, the condition numbers can be bounded by $\kappa(\mathbf{A}(\mathbf{y})) \leq (C_\kappa/h)^2$ for every $\mathbf{y} \in \Gamma$, following from assumption (A1) and the quasi-uniformity of the mesh $\mathcal{T}_h$. However, for the preconditioned systems $\mathbf{P}_0^{-1}\mathbf{A}(\mathbf{y})$, the condition numbers are bounded independent of h and $\mathbf{y} \in \Gamma$, so that $\bar{\kappa}$ is independent of mesh size h, level L, and stochastic dimension N. For the SGFEM, the condition number $\tilde{\boldsymbol{\kappa}}_r$ of the preconditioned system $\mathbf{P}^{-1}\mathbf{K}_r$ can be bounded by

$$\tilde{\boldsymbol{\kappa}}_r \leq \frac{1 + \overline{\tau}_r}{1 - \underline{\tau}_r},$$

where $\underline{\tau}_r$ and $\overline{\tau}_r$ are defined in (4.19), hence depend on ε when $r \geq r_{\min}$ from Lemma 4.7. If $\underline{\tau}_r$ and $\overline{\tau}_r \to \infty$ as $\varepsilon \to 0$, the SG system will eventually become indefinite, and the PCG method will fail to converge. In this case another strategy must be used to solve (3.14), such as GMRES, though the cost must be modified in (3.20) to account for the number of operations required to obtain the numerical solution. On the other hand, if $\underline{\tau}_r \to 1$ as $r \to \infty$, then the condition number will become increasingly large and a significant number of iterations will be needed to reduce the error of the numerical solution with PCG. Other preconditioners may be used to reduce the dependence on r, but then their associated costs must be accounted for in the work estimate (3.20) as well. For fixed $\mathbf{y} \in \Gamma$, since $a(x, \mathbf{y})$ satisfies assumption (A1), the system $\mathbf{A}(\mathbf{y})$ is positive definite, hence this is not an issue for the SCFEM.

6. Numerical examples. In this section, we provide illustrative numerical examples comparing the complexity of the SGFEM in the three cases of Examples 2.1, 2.2, and 2.3. We then compare these results with SCFEM and the results of the theoretical complexity comparison of the previous section. We solve the model problem (2.1), on the unit square $D = [0, 1]^2$. For a general coefficient $a(x, \mathbf{y})$ we do not know the exact solution to (2.1). Hence we check the convergence against a "highly

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

enriched" approximation, which we consider close enough to the exact one. To construct this "exact" solution $u_{\text{ex}}(x, \mathbf{y})$, we make use of the isotropic SCFEM based on Clenshaw-Curtis abscissas using the level L_{ex}. We approximate the computational error for the SGFEM with orders $p = 0, 1, 2, \ldots, p_{\max}$ and for the SCFEM with levels $L = 0, 1, 2, \ldots, L_{\max}$ as

$$\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{\ell^\infty} \approx \|\mathbb{E}[u_{\text{ex}} - \tilde{u}_{h,p}]\|_{\ell^\infty} \qquad \text{and} \qquad \|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{\ell^\infty} \approx \|\mathbb{E}[u_{\text{ex}} - \tilde{u}_{h,L}]\|_{\ell^\infty}, \quad (6.1)$$

where $\tilde{u}_{h,p}$ and $\tilde{u}_{h,L}$ are the fully discrete approximations (3.5) and (5.4), respectively, found by the PCG method, described in §3 and §5. In §6.3, we measure $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{\ell^\infty} \approx \|\mathbb{E}[u_{\text{ex}} - \tilde{u}^r_{h,p}]\|_{\ell^\infty}$ where $\tilde{u}^r_{h,p}$ denotes the solution of (3.14) with the projected coefficient $a^r(x, \mathbf{y})$.

As stated in §3.4 and §5.2, we use PCG with the mean-based preconditioner for SGFEM and the level-zero preconditioner for the SCFEM. Hence, we believe this puts both methods at a similar starting point for comparison, if not providing a slight advantage for the SGFEM. With these choices, the complexity results are presented in terms of the work estimates (3.20) and (5.9), respectively. The amount of work to reach a given error in PCG is also dependent on the tolerance used by the solver. If the tolerance is too small, we may see that the PCG method "over-resolves" the solution. To ensure that we do not over-resolve either solution, we set the tolerance of the solvers to be $\|\mathbb{E}[\varepsilon_{\text{SG}}]\|_{\ell^\infty}/10$ and $\|\mathbb{E}[\varepsilon_{\text{SC}}]\|_{\ell^\infty}/10$ respectively, where these quantities are first estimated for each order p and level L using a tolerance of 1.0×10^{-12}. In practice, we find that this does not affect the convergence results much.

In all three examples, we use the SG approximation constructed in terms of the orthonormal Legendre polynomials $\{\Psi_{\boldsymbol{p}}\}_{\boldsymbol{p}\in\Lambda_p}$ for given index sets Λ_p. In the presentation of the results that follow, we use the following abbreviations. For the SGFEM, we use: "SG-TD" to denote the approximation in the total degree subspace $\mathcal{P}_{\Lambda_p^{\text{TD}}}(\Gamma)$ with Λ_p^{TD} given in (3.4), and "SG-SM" to denote the approximation in the sparse Smolyak subspace $\mathcal{P}_{\Lambda_p^{\text{SM}}}(\Gamma)$ with Λ_p^{SM} given in (3.4). For the SCFEM, we use: "SC-GL" and "SC-LJ" to denote the Smolyak approximation constructed on Gauss-Legendre abscissas and the Leja approximation constructed on Clenshaw-Curtis abscissas, both defined in terms of g_{TD} and $\boldsymbol{m}$ given in (5.2) and (5.6), respectively, and "SC-CC" to denote the Smolyak approximation constructed on Clenshaw-Curtis abscissas with g_{SM} and $\boldsymbol{m}$ given in (5.2) and (5.5).

6.1. Piecewise affine coefficients. One common example in engineering and the physical sciences is that of isotropic thermal diffusion problem with a stochastic conductivity coefficient. Consider a partitioning of $D = [0, 1]^2$ into 8 circular inclusions arrayed about 1 square inclusion as in Figure 5. We present the following example

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

from [3], where the coefficient was given by

$$a(x, \mathbf{y}) = b_0(x) + \sum_{n=1}^{8} y_n \chi_n(x), \tag{6.2}$$

with $b_0 = 1$, and $y_n \sim \mathcal{U}(-0.99, -0.2)$. Here, χ_n are indicator functions corresponding to the 8 circular inclusions of radius $r = 1.3$. In this example, we also set the forcing term to be

$$f(x) = 100\chi_F(x), \tag{6.3}$$

where $F = [0.4, 0.6]^2$, is the square inclusion centered in D with side length 0.2. Figure 5 shows the expected value of the solution to this problem. To solve (2.1) with the coefficient (6.2) and forcing function (6.3), we use a piecewise linear finite element basis in the deterministic space over a nonuniform mesh $\mathcal{T}_h$. Here, the nodes of $\mathcal{T}_h$ are adapted to the geometry of our problem, that is, we fix the nodes that lie on the boundaries of the inclusions in our domain. From this fixed boundary data, we then use the `distmesh` MATLAB program [28] to generate a non-degenerate triangulation that adequately resolves the details of our subdomain geometry. We further specify the subsets of the total set of nodes that belong to each geometric inclusion, and to the boundaries of the inclusion, so that the interface conditions for the coefficient may be correctly applied. The final mesh consists of 10,604 elements, 5,377 total nodes, and 5,229 unknowns.

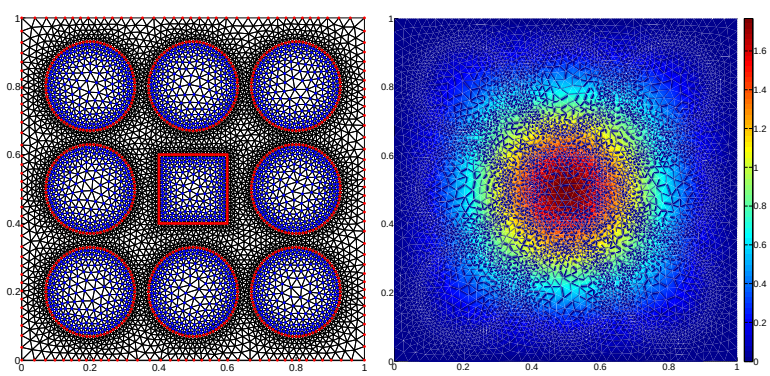

Figure 5: Left: a triangulation of the domain D with circular and square inclusions. Red nodes highlight the boundary of an inclusion or the domain D, blue nodes highlight nodes on the interior of an inclusion. **Right:** the expected value of the solution of (2.1) with stochastic conductivity coefficient (6.2).

The coefficient (6.2) is an example of a coefficient $a(x, \mathbf{y})$ having affine dependence on the parameters, e.g., Example 2.1. Figure 6 displays the convergence of the stochastic Galerkin and collocation methods against the total number of SDOF.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

For the SGFEM we take the SDOF to be the cardinality of the set Λ_p used in constructing the fully discrete approximation $u_{h,p}$ from (3.5) by solving (3.6), and for the SC method we take the SDOF to be the number of points $\#\mathcal{G}_L^{\boldsymbol{m},g}$ corresponding to an index set $\Lambda_L^{\boldsymbol{m},g}$ used in constructing the fully discrete approximation $u_{h,L}$ from (5.4). From the discussion of §5.3, we expect to see that the approximation obtained with the SGFEM requires fewer SDOF than the SCFEM to achieve the same error, and this is indeed the observed result. For example, both the SG-TD and SC-LJ approximations require the same number of SDOF, but the error of the SC-LJ approximation is much higher. This, of course, is a consequence of the estimate (5.10), where the errors of the SC approximations are bounded above by their respective Lebesgue constants against the best-approximation error in the space $\mathcal{P}_{\Lambda_L^{\boldsymbol{m},g}}(\Gamma)$.

Figure 6 also displays the convergence of both methods in terms of error versus the total computational cost of solving the system with the work estimates of (3.20) and (5.9), respectively. Here, we compute the error in $\|\mathbb{E}[\varepsilon_{\mathrm{SG}}]\|_{\ell^\infty}$ and $\|\mathbb{E}[\varepsilon_{\mathrm{SC}}]\|_{\ell^\infty}$ as given in (6.1). Our analysis shows the work corresponding to the SG discretization for SG-TD is asymptotically bounded by $\mathcal{O}([\log(1/\varepsilon)]^N)$ while the analysis from [16] shows that the work corresponding to the SC discretization for SC-CC is asymptotically bounded by $\mathcal{O}([\log(1/\varepsilon)]^N[\log\log(1/\varepsilon)]^{N-1})$. This closely matches the results of the numerical experiments in Figure 6, where it can be seen that for polynomial order $p \geq 2$, the SG-TD approximation yields the best results with the least computational cost for this problem.

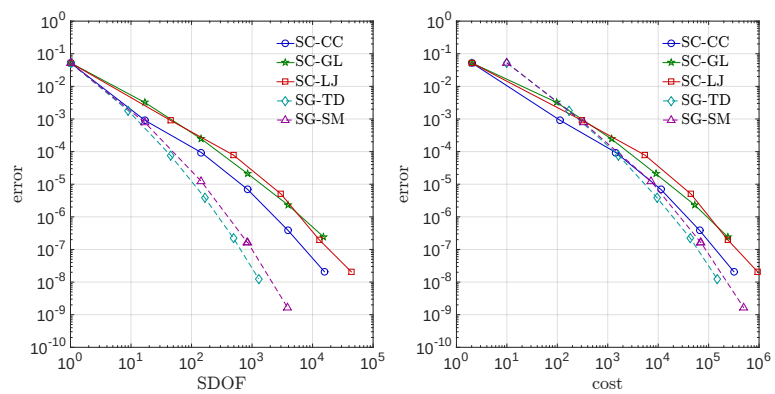

Figure 6: Left: Error versus SDOF in solving problem (2.1) with coefficient (6.2) and forcing (6.3). **Right:** Error versus computational cost with the work estimates given in (3.20) and (5.9) based on total number of matrix-vector products used by the CG method.

6.2. Polynomial coefficients. The next example we present is that of a polynomial function of the random paramters $\mathbf{y}$, e.g. the coefficient from Example 2.2. We

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

consider the following function

$$a(x,\mathbf{y}) = 5 + \sum_{|\boldsymbol{r}|\le \overline{r}} e^{-1.5|\boldsymbol{r}|}\varsigma_{\boldsymbol{r}}(x)\mathbf{y}^{\boldsymbol{r}}, \tag{6.4}$$

$$\varsigma_{\boldsymbol{r}}(x) = \begin{cases} \sin(|\boldsymbol{r}|\pi x_1)\cos(|\boldsymbol{r}|\pi x_2) & \text{if } |\boldsymbol{r}| \text{ is even,} \\ \cos(|\boldsymbol{r}|\pi x_1)\sin(|\boldsymbol{r}|\pi x_2) & \text{if } |\boldsymbol{r}| \text{ is odd,} \end{cases}$$

with $y_n \sim \mathcal{U}(-1,1)$ for all $n = 1,\dots,N$ and forcing term $f(x) = 1\ \forall x \in \overline{D}$. For the results that follow we fix $N = 4$ and study the convergence of the SGFEM and SCFEM in the cases $\overline{r} = 1,3,7$ in (6.4). As in §6.1, we set the finite element space for the spatial discretization to be the span of piecewise linear polynomials, but here we use a uniform triangulation of D with $4,934$ elements and $2,340$ spatial unknowns.

Figure 7 displays the convergence of the SGFEM and SCFEM in terms of error versus the total computational cost of solving the system with the work estimates of (3.20) and (5.9). Here, we compute the error in $\|\mathbb{E}[\varepsilon_{\mathrm{SG}}]\|_{\ell^\infty}$ and $\|\mathbb{E}[\varepsilon_{\mathrm{SC}}]\|_{\ell^\infty}$ as given in (6.1). As we increase the order $\overline{r}$ in (6.4), we see that the work for the SGFEM increases, corresponding to the decreasing sparsity of the matrix $\mathbf{K}$ from (3.8). Here, the work of matrix-vector multiplications with $\mathbf{K}$ are of the order $\mathcal{O}(J_h M_p M_{\overline{r}} \min\{2^{\overline{r}}, M_{\lceil \overline{r}/2\rceil}\})$, where $M_{\overline{r}} \min\{2^{\overline{r}}, M_{\lceil \overline{r}/2\rceil}\}$ is a large constant that grows rapidly with $\overline{r}$. As a result, we see that for $\overline{r} = 1$, the SGFEM outperforms the other methods for $p \ge 4$. However, for $\overline{r} = 3,7$, the extra work of the matrix-vector multiplications of the coupled SG system dominates the overall convergence.

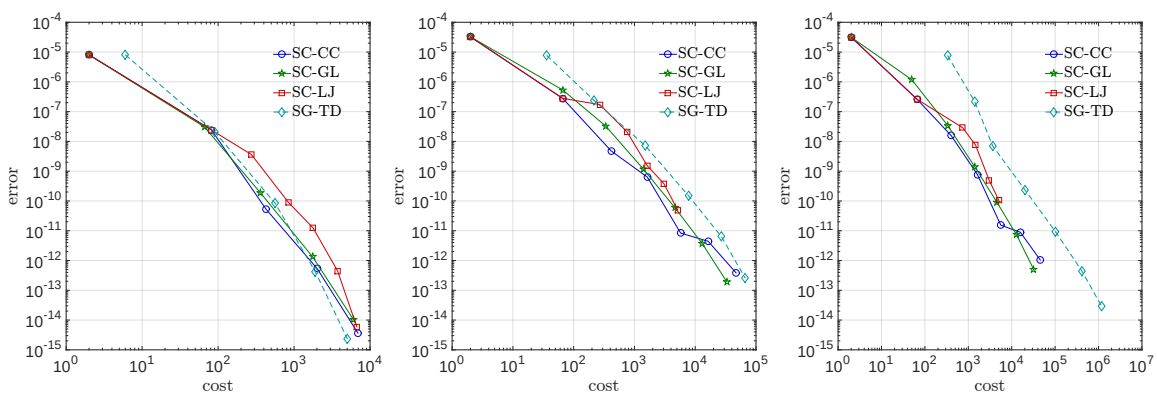

Figure 7: Error versus cost for solving problem (2.1) with coefficient (6.4) having $\overline{r} = 1$ (left), $\overline{r} = 3$ (middle), and $\overline{r} = 7$ (right), with forcing $f(x) = 1$. The cost, given in (3.20) and (5.9), is based on total number of matrix-vector products used by the PCG method.

6.3. Transcendental coefficients. The next example we present is that of a random coefficient defined in terms of the truncated Karhunen-Loève expansion of the function $\log(a(x,\mathbf{y}) - a_{\min})$, for $a_{\min} > 0$. This example represents a commonly

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

used transcendental function of the physical and random parameters, e.g., Example 2.3, and is often presented in the context of enforcing the positivity of $a(x, \mathbf{y})$ required in assumption (A1). Coefficients of this type are commonly found in groundwater flow models. For these models, the permeability can exhibit large variance within each layer of sediment, and as a result are better represented on a logarithmic scale. We recall the problem of solving (2.1) with a coefficient having one-dimensional (layered) spatial dependence and a deterministic load $f(x_1, x_2, \omega) = 2\cos(x_1)\sin(x_2)$ as studied in [25, 26], where $a(x, \mathbf{y})$ was given by

$$\log(a(x,\omega) - 0.5) = 1 + y_1(\omega)\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} + \sum_{n=2}^{N} \zeta_n \varphi_n(x) y_n(\omega), \tag{6.5}$$

$$\zeta_n = (\sqrt{\pi}L)^{1/2} \exp\left(-\left(\left\lfloor \tfrac{n}{2} \right\rfloor \pi L\right)^2 / 8\right), \text{ for } n > 1,$$

$$\varphi_n(x) = \begin{cases} \sin\left(\left\lfloor \frac{n}{2} \right\rfloor \pi x_1 / L_p\right), & \text{if } n \text{ is even,} \\ \cos\left(\left\lfloor \frac{n}{2} \right\rfloor \pi x_1 / L_p\right), & \text{if } n \text{ is odd.} \end{cases}$$

Here, $\{y_n(\omega)\}_{n=1}^{\infty}$ are independent random variables uniformly distributed in $[-\sqrt{3}, \sqrt{3}]$ with zero mean and unit variance. For $x_1 \in [0, b]$, let L_c be a desired physical correlation length for the random field $a(x, \mathbf{y})$, chosen so that the random variables $a(x_1, \omega)$ and $a(x_1', \omega)$ become essentially uncorrelated for $|x_1 - x_1'|\mathbf{g}L_c$. Also, let $L_p = \max\{b, 2L_c\}$ and $L = L_c/L_p$. Expression (6.5) represents a possible truncation of a one-dimensional random field with stationary covariance,

$$\mathrm{cov}[\log(a - 0.5)](x_1, x_2) = \exp\left(\frac{-(x_1 - x_2)^2}{L_c^2}\right).$$

Direct integration with the coefficient $a(x, \mathbf{y})$ from (6.5) yields a fully-block dense linear system $\mathbf{K}$ from (3.8) that is computationally infeasible to solve [14, 23, 34, 35]. The purpose of this example is to highlight the difficulties of obtaining a fully discrete approximation with the SGFEM in this case.

As in the previous example in §6.2, we set the finite element space for the spatial discretization to be the span of piecewise linear polynomials and use a uniform triangulation of D with $4,934$ elements and $2,340$ spatial unknowns. For the results that follow, we fix the truncation length $N = 9$ and correlation length $L_c = 1/64$ in (6.5). To maintain sparsity of the SG system, we use the strategy of projecting the coefficient $a(x, \mathbf{y})$ from (6.5) onto the space $\mathcal{P}_{\Lambda_r}(\Gamma)$, as in (3.10), where $\Lambda_r = \Lambda_r^{\mathrm{TD}}$ for the SG-TD approximation, and $\Lambda_r = \Lambda_r^{\mathrm{SM}}$ for the SG-SM approximation, obtaining the matrix $\mathbf{K}_r$ from (3.14). We then increase r while p is fixed until the error in the solution stagnates, in practice finding that, for this problem, $r = p$ is sufficient to guarantee the error of the projection does not exceed that of the solution, while maintaining sparsity of the linear system.

Figure 8 compares the error versus SDOFs. There we see that for order $p \geq 3$, the

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

SG-TD approximation provides the best approximation with respect to SDOFs. As discussed in §5.3, this is to be expected since the computational complexity of solving the coupled and decoupled systems is not taken into account. Figure 8 also displays the convergence in error versus the total computational cost of solving the system with the work estimates of (3.20) and (5.9). Here however, the results show that the SGFEM requires significantly more work to obtain the same error than the SCFEM. In Remark 4.4, we observed that for the TD-SG approximation, when $r = p$ the cost of solving (3.14) with the PCG method is of the order $\mathcal{O}(J_h M^3_{\lceil p/2 \rceil} N^{\mathrm{SG}}_{\mathrm{iter}})$, growing much more rapidly than the cost in the affine and polynomial coefficient cases, e.g., Examples 2.1 and 2.2, as we increase the order p of the total degree approximation with the SG method. Table 1 shows the amount of work required to achieve an error on the order of 1.0×10^{-k} for some values of k and the savings in computational effort of solving the system with the SC method instead of the SGFEM.

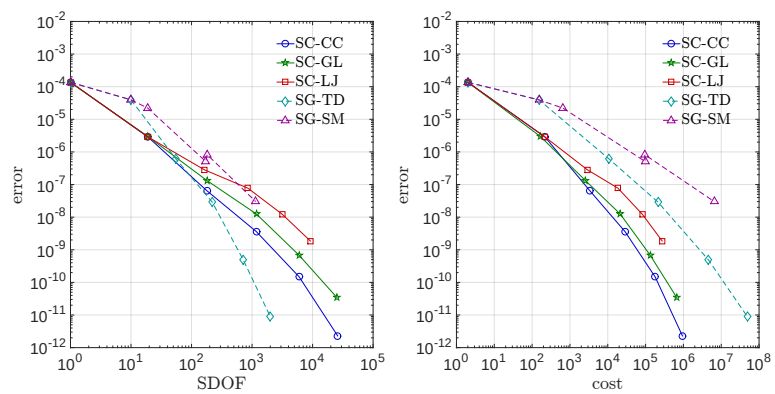

Figure 8: Left: Error versus SDOFs in solving problem (2.1) with coefficient (6.5) and forcing $f(x_1, x_2, \omega) = \cos(x_1)\sin(x_2)$. **Right:** Error versus cost with the work estimates given in (3.20) and (5.9).

7. Conclusions. In this work, we presented explicit cost bounds for applying the SGFEM to the solution of an elliptic PDE having both affine and non-affine random coefficients. To this end, we have conducted a rigorous counting argument for the sparsity of the linear system that results from the SG discretization with a global orthogonal basis defined on an isotropic total degree index set. Our analysis shows that when the coefficient is an affine or non-affine function of the random variables having fixed polynomial order, the computational cost of solving the coupled SG system grows linearly with the dimension of the polynomial subspace. In these cases, the results only differ by a constant depending on the polynomial order of the random coefficient and the dimension of the parameter domain.

On the other hand, when the coefficient is a non-affine, transcendental function of the random variables requiring an additional orthogonal expansion, our analysis shows that the computational complexity, no longer grows linearly with the polynomial

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

SC-CC Level	SC-CC Error	Mat-vec cost of SC-CC	SG-TD Order	SG-TD Error	Mat-vec cost of SG-TD
0	1.3626×10^{-4}	2	0	1.3626×10^{-4}	4
1	2.8884×10^{-6}	218	1	3.9444×10^{-5}	152
2	6.3652×10^{-8}	3,398	2	6.1427×10^{-7}	10,710
3	3.6021×10^{-9}	28,638	3	2.8851×10^{-8}	213,010
4	1.4794×10^{-10}	178,894	4	4.9210×10^{-10}	4,579,575
5	2.2869×10^{-12}	944,220	5	8.9123×10^{-12}	49,089,051

Table 1: Comparison of cost in matrix-vector products for solving problem (2.1) with coefficient (6.5) and forcing $f(x_1, x_2, \omega) = \cos(x_1)\sin(x_2)$ using the SC-CC and SG-TD approximations, with the strategy of picking the CG tolerance to be $\|\mathbb{E}[u_{ex} - \tilde{u}^r_{h,p}]\|_{\ell^\infty}/10$ for the SGFEM and $\|\mathbb{E}[u_{ex} - \tilde{u}_{h,L}]\|_{\ell^\infty}/10$ for the SC method. Cost in matrix-vector products for SG-TD method is given by (3.20) and for SC-CC is given by (5.9) normalized by the cost of a finite element matrix vector product.

subspace dimension. For such coefficients, we are able to provide bounds on the complexity that depend on the truncation order of the coefficient. These estimates imply that a truncation of the expansion should be used, when possible, though attention must be paid to the well-posedness of the resulting PDE.

The analysis conducted herein motivates the study of the total computational complexity of obtaining fully discrete approximations with such methods. We have seen that, despite the fact that the SG method yields an approximation that is optimal in the L^2 sense for a given polynomial subspace, the associated computational costs of obtaining SG approximations are not optimal for all problems. Moreover, we have observed, both through theoretical comparisons and numerical examples, that changing the underlying polynomial subspace and method used for obtaining the fully discrete approximation can often yield a solution that requires far less work to obtain, but has the same error.

Acknowledgements. The first author would like to acknowledge Dr. Miroslav Stoyanov for his insightful comments and assistance in producing the stochastic collocation results with the TASMANIAN package [1].

This material is based upon work supported in part by the U.S. Air Force of Scientific Research under grant number 1854-V521-12; by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under contract numbers ERKJ259, and ERKJE45 and by the Laboratory Directed Research and Development program at the Oak Ridge National Laboratory, which is operated by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

REFERENCES

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

[1] *Toolkit for Adaptive Stochastic Modeling and Non-Intrusive ApproximatioN*. `http://tasmanian.ornl.gov/`.

[2] K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, vol. 39 of Texts in Applied Mathematics, Springer, New York, NY, 3rd ed., 2005.

[3] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone, *Stochastic Spectral Galerkin and Collocation Methods for PDEs with Random Coefficients: A Numerical Comparison*, in Spectral and High Order Methods for Partial Differential Equations, J. S. Hesthaven and E. M. Rønquist, eds., vol. 76 of Lecture Notes in Computational Science and Engineering, Springer Berlin Heidelberg, 2011, pp. 43–62.

[4] A. Chkifa, A. Cohen, and C. Schwab, *Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs*, Journal de Mathématiques Pures et Appliquées, 103 (2014), pp. 400–428.

[5] C. W. Clenshaw and A. R. Curtis, *A method for numerical integration on an automatic computer*, Numerische Mathematik, 2 (1960), pp. 197–205.

[6] A. Cohen, A. Chkifa, R. DeVore, and C. Schwab, *Sparse adaptive Talor approximation algorithms for parametric and stochastic elliptic PDEs*, ESAIM: Mathematical Modelling and Numerical Analysis, (2012), pp. 1–27.

[7] A. Cohen, R. DeVore, and C. Schwab, *Convergence Rates of Best N-term Galerkin Approximations for a Class of Elliptic sPDEs*, Foundations of Computational Mathematics, 10 (2010), pp. 615–646.

[8] A. Cohen, R. Devore, and C. Schwab, *Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDEs*, Analysis and Applications, 09 (2011), pp. 11–47.

[9] P. J. Davis, *Interpolation and Approximation*, Dover, 1975.

[10] S. De Marchi, *On Leja sequences: Some results and applications*, Applied Mathematics and Computation, 152 (2004), pp. 621–647.

[11] V. K. Dzjadyk and V. V. Ivanov, *On asymptotics and estimates for the uniform norms of the Lagrange interpolation polynomials corresponding to the Chebyshev nodal points*, Analysis Mathematica, 9 (1983), pp. 85–97.

[12] M. Eiermann, O. G. Ernst, and E. Ullmann, *Computational aspects of the stochastic finite element method*, in Computing and Visualization in Science, vol. 10, 2007, pp. 3–15.

[13] H. C. Elman, C. W. Miller, E. T. Phipps, and R. S. Tuminaro, *Assessment of collocation and Galerkin approaches to linear diffusion equations with random data*, International Journal for Uncertainty Quantification, 1 (2011), pp. 19–33.

[14] O. G. Ernst and E. Ullmann, *Stochastic Galerkin Matrices*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 1848–1872.

[15] G. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Series in Operations Research and Financial Engineering, Springer, 1996.

[16] D. Galindo, P. Jantsch, C. G. Webster, and G. Zhang, *Accelerating stochastic collocation methods for PDEs with random input data*, Tech. Rep. ORNL/TM-2015/219, Oak Ridge National Laboratory, 2015.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

[17] M. Gunzburger and C. G. Webster, *Uncertainty quantification for partial differential equations with stochastic coefficients*, The Mathematical Intelligencer, (2014). To appear.

[18] M. Gunzburger, C. G. Webster, and G. Zhang, *An adaptive wavelet stochastic collocation method for irregular solutions of partial differential equations with random input data*, in Sparse Grids and Applications - Munich 2012, vol. 97 of Lecture Notes in Computational Science and Engineering, Springer International Publishing, 2014, pp. 137–170.

[19] M. D. Gunzburger, C. G. Webster, and G. Zhang, *Stochastic finite element methods for partial differential equations with random input data*, Acta Numerica, 23 (2014), pp. 521–650.

[20] C. Johnson, *Numerical solution of partial differential equations by the finite element method*, Courier Dover Publications, 2012.

[21] O. P. Le Maître and O. M. Knio, *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*, Scientific Computation, Springer, 2010.

[22] M. Loève, *Probability Theory*, no. v. 2 in Graduate Texts in Mathematics, Springer, 1978.

[23] H. G. Matthies and A. Keese, *Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations*, Computer Methods in Applied Mechanics and Engineering, 194 (2005), pp. 1295–1331.

[24] F. Nobile and R. Tempone, *Analysis and implementation issues for the numerical approximation of parabolic equations with random coefficients*, Online, 80 (2009), pp. 979–1006.

[25] F. Nobile, R. Tempone, and C. G. Webster, *A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 2309–2345.

[26] ——, *An Anisotropic Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 2411–2442.

[27] M. F. Pellissetti and R. G. Ghanem, *Iterative Solution of Systems of Linear Equations Arising in the Context of Stochastic Finite Elements*, Advances in Engineering Software, 31 (2000), pp. 607–616.

[28] P.-O. Persson and G. Strang, *A Simple Mesh Generator in MATLAB*, SIAM Review, 46 (2004), pp. 329–345.

[29] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA Journal of Numerical Analysis, 29 (2009), pp. 350–375.

[30] S. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, Dokl. Akad. Nauk SSSR, (1963), pp. 4:240–243.

[31] G. Szegö, *Orthogonal polynomials*, vol. XXIII, Amer. Math. Soc., 4 ed., 1975.

[32] R. A. Todor and C. Schwab, *Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients*, IMA Journal of Numerical Analysis, 27 (2007), pp. 232–261.

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.

[33] H. Tran, C. G. Webster, and G. Zhang, *Analysis of quasi-optimal polynomial approximations for parameterized PDEs with deterministic and stochastic coefficients*, Tech. Rep. ORNL/TM-2014/468, Oak Ridge National Laboratory, 2014. Submitted.

[34] E. Ullmann, *A Kronecker Product Preconditioner for Stochastic Galerkin Finite Element Discretizations*, SIAM Journal on Scientific Computing, 32 (2010), pp. 923–946.

[35] E. Ullmann, H. C. Elman, and O. G. Ernst, *Efficient Iterative Solvers for Stochastic Galerkin Discretizations of Log-Transformed Random Diffusion Problems*, SIAM Journal on Scientific Computing, 34 (2012), pp. A659–A682.

[36] N. Wiener, *The Homogeneous Chaos*, American Journal of Mathematics, 60 (1938), pp. pp. 897–936.

[37] D. Xiu and G. E. Karniadakis, *The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Copyright © by ORNL. Unauthorized reproduction of this article is prohibited.