IMA Journal of Numerical Analysis (2009) **29**, 350–375
doi:10.1093/imanum/drn014
Advance Access publication on April 4, 2008

Block-diagonal preconditioning for spectral stochastic finite-element systems

CATHERINE E. POWELL†
School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, UK

AND

HOWARD C. ELMAN‡
Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA

[Received on 25 May 2007; revised on 9 January 2008]

Deterministic models of fluid flow and the transport of chemicals in flows in heterogeneous porous media incorporate partial differential equations (PDEs) whose material parameters are assumed to be known exactly. To tackle more realistic stochastic flow problems, it is fitting to represent the permeability coefficients as random fields with prescribed statistics. Traditionally, large numbers of deterministic problems are solved in a Monte Carlo framework and the solutions are averaged to obtain statistical properties of the solution variables. Alternatively, so-called stochastic finite-element methods (SFEMs) discretize the probabilistic dimension of the PDE directly leading to a single structured linear system. The latter approach is becoming extremely popular but its computational cost is still perceived to be problematic as this system is orders of magnitude larger than for the corresponding deterministic problem. A simple block-diagonal preconditioning strategy incorporating only the mean component of the random field coefficient and based on incomplete factorizations has been employed in the literature and observed to be robust, for problems of moderate variance, but without theoretical analysis. We solve the stochastic Darcy flow problem in primal formulation via the spectral SFEM and focus on its efficient iterative solution. To achieve optimal computational complexity, we base our block-diagonal preconditioner on algebraic multigrid. In addition, we provide new theoretical eigenvalue bounds for the preconditioned system matrix. By highlighting the dependence of these bounds on all the SFEM parameters, we illustrate, in particular, why enriching the stochastic approximation space leads to indefinite system matrices when unbounded random variables are employed.

Keywords: finite elements; stochastic finite elements; fast solvers; preconditioning; multigrid.

1. Introduction

Fluid flow and the transport of chemicals in flows in heterogeneous porous media are modelled mathematically using partial differential equations (PDEs). In deterministic modelling, inputs such as material properties, boundary conditions and source terms are assumed to be known explicitly. Such assumptions lead to tractable computations. However, simulations based on such over-simplifications cannot be used in practice to quantify the probability of an unfavourable event such as, say, a chemical being transported

†Corresponding author. Email: c.powell@manchester.ac.uk
‡Email:elman@cs.umd.edu

at a lethal level of concentration in groundwater. If the input variables for the system being studied are subject to uncertainty, then it is fitting to represent them as random fields. Solutions to the resulting stochastic PDEs are then necessarily also random fields. Strategic decision making cannot be made without some form of uncertainty quantification. Typically, a few moments of the solution variables are required, or the probability distribution of a particular quantity of interest.

We focus on the case of uncertainty in material properties. The simplest and most commonly employed way of dealing with this is via the Monte Carlo Method (MCM). Large numbers of realizations of the random system inputs are generated and each resulting deterministic problem is solved using the available numerical methods and solvers. Results are post-processed to determine the desired statistical properties of the solution variables. Care must be taken, however, to ensure that enough realizations are generated so that the probability space is sampled appropriately. The exact number of trials required depends on the problem at hand but hundreds of thousands of experiments are not untypical for realistic flow problems with large variance. Easy access to parallel computers makes this feasible in the 21st century. Quasi MCMs and variance reduction techniques can be used to reduce the overall number of trials, making this technology even more competitive. However, minimizing the computational cost of solving each deterministic problem is still a crucial and non-trivial step.

An alternative approach, pioneered in Ghanem & Spanos (2003), couples a Karhunen–Loève (KL) expansion of the random field coefficients in the stochastic PDE with a traditional finite-element discretization on the spatial domain. The stochastic dimension of the problem is discretized directly. The advantage of this so-called stochastic finite-element method (SFEM) is that a single linear system needs to be solved. However, this is orders of magnitude larger than the subproblems solved in the MCM. The components of the discrete solution are coefficients of a probabilistic expansion of the solution variables which can easily be post-processed to recover the mean, variance and probability distribution of quantities of interest. SFEMs are becoming increasingly popular but their computational cost is perceived to be high. The linear systems in question are, however, highly structured and researchers have been slow to take up the challenge of solving them efficiently. Initial attempts were made in Ghanem & Kruger (1996) and Pellissetti & Ghanem (2000). More recently, a fast and efficient linear algebra for alternative SFEMs (see Deb *et al.*, 2001; Babuška & Chatzipantelidis, 2002; Babuška *et al.*, 2004) has been proposed by linear algebra specialists (see Eiermann *et al.*, 2007; Elman *et al.*, 2005a; Elman & Furnival, 2007) and fast solvers and parallel computer architectures have been exploited in Keese (2003, 2004) and Keese & Matthies (2002, 2003).

We focus on the numerical solution of the steady-state diffusion problem which, in its deterministic formulation, is written as

$$\begin{aligned} -\nabla \cdot K \nabla u &= f, \quad \text{in } D \subset \mathbb{R}^d, \\ u &= g, \quad \text{on } \partial D_D \neq \emptyset, \\ K \nabla u \cdot \vec{n} &= 0, \quad \text{on } \partial D_N = \partial D \backslash \partial D_D. \end{aligned} \tag{1.1}$$

The boundary-value problem (1.1) is the primal formulation of the standard second-order elliptic problem and provides a simplified model for single-phase flow in a saturated porous medium (e.g. see Russell & Wheeler, 1983; Ewing & Wheeler, 1983). In that physical setting, u and $\vec{q} = K \nabla u$ are the residual pressure and velocity field, respectively. K is a prescribed scalar function or a $d \times d$ symmetric and uniformly positive-definite tensor, representing permeability. Since the permeability coefficients of a heterogeneous porous medium can never, in reality, be known at every point in space, we consider, here, the case where $K = K(\vec{x}, \omega)$ is a random field. We assume only statistical properties of K.

To make these notions precise, let $(\Omega, \mathscr{B}, P)$ denote a probability space where Ω, $\mathscr{B}$, and P are the set of random events, the minimal σ-algebra of the subsets of Ω and an appropriate probability measure, respectively. Then $K(\vec{x}, \omega) : D \times \Omega \to \mathbb{R}$. For a fixed spatial location $\vec{x} \in D$, $K(\cdot, \omega)$ is a random variable, while for a fixed realization $\omega \in \Omega$, $K(\vec{x}, \cdot)$ is a spatial function in $\vec{x}$. The stochastic problem then reads: find a random field $u(\vec{x}, \omega) : D \times \Omega \to \mathbb{R}$ such that P-almost surely

$$\begin{aligned} -\nabla \cdot K(\vec{x}, \omega)\nabla u(\vec{x}, \omega) &= f(\vec{x}), \quad \vec{x} \in D, \\ u(\vec{x}, \omega) &= g(\vec{x}), \quad \vec{x} \in \partial D_D, \\ K(\vec{x}, \omega)\nabla u(\vec{x}, \omega) \cdot \vec{n} &= 0, \quad \vec{x} \in \partial D_N = \partial D \backslash \partial D_D, \end{aligned} \tag{1.2}$$

where $f(\vec{x})$ and $g(\vec{x})$ are suitable deterministic functions. The source term f can also be treated as a random field in a straightforward manner (see Elman *et al.*, 2005a; Deb *et al.*, 2001) but we shall not consider that case.

1.1 *Overview*

The focus of this work is the design of fast solvers for (1.2). In Section 2, we summarize the classical spectral SFEM discretization from Ghanem & Spanos (2003) and discuss some modelling issues that affect the spectral properties of the resulting linear systems and ultimately the solver performance. We highlight the structure and algebraic properties of the resulting linear system and implement the block-diagonal preconditioning scheme advocated in Ghanem & Kruger (1996). For the subproblems, however, we replace the traditional incomplete factorization schemes used in Ghanem & Kruger (1996) and Pellissetti & Ghanem (2000) with a black-box algebraic multigrid (AMG) solver. We also compare the computational effort required with that of traditional MCMs. Our main contributions, namely the derivation of key properties of the finite-element matrices and eigenvalue bounds for the preconditioned system matrices, are presented in Section 3. The bounds are shown to be tight for test problems commonly used in the literature. Numerical results are presented in Section 4.

2. Spectral SFEMs for the steady-state diffusion problem

SFEMs can be divided into two categories: non-spectral and spectral methods. The former (see Deb *et al.*, 2001; Elman *et al.*, 2005a) achieves a prescribed accuracy via polynomial approximation of a fixed degree on an increasingly fine partition of a probability range space. The latter requires no such formal partition and error is reduced by increasing the degree, p, of polynomial approximation. The methods can be further subcategorized according to the choice of stochastic basis functions (orthogonal, as in Ghanem (1998), doubly-orthogonal as in Babuška *et al.* (2004), etc.) We focus on the classical spectral method outlined in Ghanem & Spanos (2003) and employ a standard orthogonal polynomial chaos basis. The main advantage is that the dimension of this space grows more slowly than for other choices (see Babuška *et al.* (2004) or Deb *et al.* (2001) for alternatives). However, the stochastic terms are fully coupled and this is much more challenging for solvers. For notational convenience, we illustrate the derivation of the spectral SFEM equations for the case of homogeneous Dirichlet boundary conditions only. This derivation is completely standard and full details can be found in Ghanem & Spanos (2003), Deb *et al.* (2001), Babuška & Chatzipantelidis (2002) and Babuška *et al.* (2004).

If the coefficient $K(\vec{x}, \omega)$ is bounded and strictly positive, i.e.

$$0 < k_1 \leqslant K(\vec{x}, \omega) \leqslant k_2 < +\infty, \quad \text{a.e. in } D \times \Omega, \tag{2.1}$$

then (1.2) can be cast in weak form in the usual way, and existing theory (i.e. the classical Lax–Milgram lemma) can be used to establish existence and uniqueness of a solution. In the present stochastic setting, the idea is to seek a weak solution in a Hilbert space $H = H_0^1(D) \otimes L^2(\Omega)$, consisting of tensor products of deterministic functions defined on the spatial domain and stochastic functions defined on the probability space.

In order to set up variational problems, some notation is first required. Let X be a real random variable belonging to $(\Omega, \mathscr{B}, P)$ and assume that there exists a density function $\rho : \mathbb{R}^d \to \mathbb{R}$ such that the expected value can be expressed via the integral

$$\langle X \rangle = \int_{\mathbb{R}} x\rho(x)\mathrm{d}x.$$

If $\langle X \rangle < \infty$, then $X \in L^1(\Omega)$. The space $L^2(D) \otimes L^2(\Omega) = \{v(\vec{x}, \omega) : D \times \Omega \to \mathbb{R} \,|\, \|v\| < \infty\}$ consists of random functions with finite second moment where the norm $\|\cdot\|$ is defined via

$$\|v(\vec{x}, \omega)\|^2 = \left\langle \int_D v^2(\vec{x}, \omega)\mathrm{d}\vec{x} \right\rangle. \tag{2.2}$$

Next, we define $V = \{v(\vec{x}, \omega) : D \times \Omega \to \mathbb{R} \,|\, \|v\|_V < \infty,\ v|_{\partial D \times \Omega} = 0\}$, where the 'stochastic energy' norm $\|\cdot\|_V$ is defined via

$$\|v(\vec{x}, \omega)\|_V^2 = \left\langle \int_D K(\vec{x}, \omega)|\nabla v(\vec{x}, \omega)|^2 \,\mathrm{d}\vec{x} \right\rangle. \tag{2.3}$$

If condition (2.1) holds, then the norm is well defined and it can be shown that there exists a unique $u = u(\vec{x}, \omega) \in V$ satisfying the continuous variational problem

$$\left\langle \int_D K(\vec{x}, \omega)\nabla u(\vec{x}, \omega) \cdot \nabla v(\vec{x}, \omega)\mathrm{d}\vec{x} \right\rangle = \left\langle \int_D f(\vec{x})v(\vec{x}, \omega)\mathrm{d}\vec{x} \right\rangle \quad \forall\, v(\vec{x}, \omega) \in V. \tag{2.4}$$

To convert the stochastic problem (2.4) into a deterministic one, we require a finite set of random variables $\{\xi_1(\omega), \ldots, \xi_M(\omega)\}$ that represent appropriately and sufficiently the stochastic variability of $K(\vec{x}, \omega)$. One possibility is to approximate $K(\vec{x}, \omega)$ by a truncated KL expansion, a linear combination of a finite set of uncorrelated random variables. We discuss this in Section 2.1. After formally replacing $K(\vec{x}, \omega)$ by $K_M(\vec{x}, \vec{\xi})$, it can be shown that the corresponding solution also has finite stochastic dimension and the variational problem (2.4) can be restated as follows: find $u = u(\vec{x}, \vec{\xi}) \in W$ satisfying

$$\int_\Gamma \rho(\vec{\xi}) \int_D K_M(\vec{x}, \vec{\xi})\nabla u(\vec{x}, \vec{\xi}) \cdot \nabla w(\vec{x}, \vec{\xi})\mathrm{d}\vec{x}\,\mathrm{d}\vec{\xi} = \int_\Gamma \rho(\vec{\xi}) \int_D f(\vec{x})w(\vec{x}, \vec{\xi})\mathrm{d}\vec{x}\,\mathrm{d}\vec{\xi} \tag{2.5}$$

$\forall\, w(\vec{x}, \vec{\xi}) \in W$. Here, $\rho(\vec{\xi})$ denotes the joint probability density function of the random variables and $\Gamma = \Gamma_1 \times \cdots \times \Gamma_M$ is the joint image of the random vector $\vec{\xi}$. A key point is that if the random variables are mutually independent, then the density function is separable, i.e. $\rho(\vec{\xi}) = \rho_1(\xi_1) \cdot \rho_2(\xi_2) \cdots \rho_M(\xi_M)$ and the integrals in (2.5) simplify greatly. Many practitioners use Gaussian random variables because uncorrelated Gaussian random variables are independent (e.g. see Ghanem & Spanos, 2003, or Elman & Furnival, 2007). However, (2.1) is not satisfied in this case, so the resulting problem (2.4) is not well posed. Other authors (see Deb *et al.*, 2001; Elman *et al.*, 2005a; Babuška & Chatzipantelidis, 2002) work with random variables with bounded images in order to satisfy (2.1) and introduce independence as an extra modelling assumption. An approach that leads to well-posed problems without an explicit assumption of independent variables is given in Babuška *et al.* (2007).

Formally, the space W differs from V since the definition of the norm induced by the inner product in (2.5) is defined in terms of the density $\rho(\vec{\xi})$. Hence, to make (2.5) understood, we define

$$W = H_0^1(D) \otimes L^2(\Gamma) = \{w(\vec{x}, \vec{\xi}) \in L^2(D \times \Gamma) \mid \|w(\vec{x}, \vec{\xi})\|_W < \infty \text{ and } w|_{\partial D \times \Gamma} = 0\} \tag{2.6}$$

and the energy norm

$$\|w(\vec{x}, \vec{\xi})\|_W^2 = \int_\Gamma \rho(\vec{\xi}) \int_D K_M(\vec{x}, \vec{\xi}) |\nabla w(\vec{x}, \vec{\xi})|^2 \, \mathrm{d}\vec{x} \, \mathrm{d}\vec{\xi}.$$

Representing the stochastic behaviour of $K(\vec{x}, \omega)$ by a finite set of random variables (a form of model order reduction) can be viewed as the first step in the discretization process. To obtain a fully discrete version of (2.5), we now need a finite-dimensional subspace $W_h \subset W = H_0^1(D) \otimes L^2(\Gamma)$. The key idea of the SFEM is to discretize the deterministic space $H_0^1(D)$ and the stochastic space $L^2(\Gamma)$ separately. Hence, given bases

$$X_h = \text{span}\{\phi_i(\vec{x})\}_{i=1}^{N_x} \subset H_0^1(D), \quad S = \text{span}\{\psi_j(\vec{\xi})\}_{j=1}^{N_\xi} \subset L^2(\Gamma), \tag{2.7}$$

which may be chosen independently of one another, we define

$$W^h = X_h \otimes S = \{v(\vec{x}, \vec{\xi}) \in L^2(D \times \Gamma) \mid v(\vec{x}, \vec{\xi}) \in \text{span}\{\phi(\vec{x})\psi(\vec{\xi}), \phi \in X_h, \psi \in S\}\}. \tag{2.8}$$

We choose the basis for X_h by defining the functions $\phi_i(\vec{x})$ to be the standard hat functions associated with piecewise linear (or bilinear) approximation associated with a partition T_h of the spatial domain D into triangles (or rectangles). Different classes of SFEMs are distinguished by their choices for S. In Elman *et al.* (2005a), Deb *et al.* (2001) and Babuška *et al.* (2004), tensor products of piecewise polynomials on the subdomains Γ_i are employed. In this approach, the polynomial degree is fixed and approximation is improved by refining the partition of Γ. The classical, so-called spectral SFEM (see Ghanem & Spanos, 2003; Elman & Furnival, 2007; Le Maitre *et al.*, 2003; Sudret & Der Kiureghian, 2000) employs global polynomials of total degree p in M random variables ξ_i on Γ. In this approach, there is no partition of Γ and approximation is improved by increasing the polynomial degree. We shall adopt the latter method.

When the underlying random variables are Gaussian, the spectral approach uses a basis of multidimensional Hermite polynomials of total degree p, termed the 'polynomial chaos' (see Wiener, 1938). The use of Hermite polynomials ensures that the corresponding basis functions are orthogonal with respect to the Gaussian probability measure. This leads to sparse linear systems, a crucial property that must be exploited for fast solution schemes. If alternative distributions are used to model the input random field, then appropriate stochastic basis functions should be used to ensure orthogonality with respect to the probability measure they induce (see Xiu & Karniadakis, 2003). For example, if uniform random variables with zero mean and unit variance (having support on the bounded interval $[-\sqrt{3}, \sqrt{3}]$) are selected, Legendre polynomials are the correct choice. Convergence and approximation properties of the resulting SFEM, when random variables with bounded images are employed, are discussed in Babuška *et al.* (2004).

To illustrate the construction of S, consider the case of Gaussian random variables with $M = 2$ and $p = 3$. Then S is the set of two-dimensional Hermite polynomials (the product of a univariate Hermite polynomial in ξ_1 and a univariate polynomial in ξ_2) of degree less than or equal to three. Each basis function is associated with a multi-index $\alpha = (\alpha_1, \alpha_2)$, where the components represent the degrees of polynomials in ξ_1 and ξ_2. Since the total degree of the polynomial is three, we have the possibilities

$\alpha = (0,0), (1,0), (2,0), (3,0), (0,1), (1,1), (2,1), (0,2), (1,2)$ and $(0,3)$. Given that the univariate Hermite polynomials of degrees $0, 1, 2, 3$ are $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$ and $H_3(x) = x^3 - 3x$, we obtain

$$\begin{aligned} S &= \operatorname{span}\{\psi_j(\vec{\xi})\}_{j=1}^{10} \\ &= \{1, \xi_1, \xi_1^2 - 1, \xi_1^3 - 3\xi_1, \xi_2, \xi_1\xi_2, (\xi_1^2 - 1)\xi_2, \xi_2^2 - 1, (\xi_2^2 - 1)\xi_1, \xi_2^3 - 3\xi_2\}. \end{aligned}$$

Note that the dimension of this space is

$$N_\xi = 1 + \sum_{s=1}^{p} \frac{1}{s!} \prod_{r=0}^{s-1} (M + r) = \frac{(M+p)!}{M!p!}.$$

In the sequel, we provide results for Gaussian random variables and Hermite polynomials, which are popular with practitioners (see Elman & Furnival, 2007; Ghanem & Kruger, 1996; Ghanem & Spanos, 2003; Keese, 2004; Pellissetti & Ghanem, 2000) as well as bounded, independent uniform random variables with Legendre polynomials (see Deb *et al.*, 2001; Elman *et al.*, 2005a). Other distributions can be accommodated in the same framework provided that the correct choice of orthogonal polynomial is made.

2.1 *KL expansion*

A random field $K(\vec{x}, \omega)$ with continuous covariance function

$$C(\vec{x}, \vec{y}) = \langle (K(\vec{x}, \omega) - \langle K(\vec{x}) \rangle)(K(\vec{y}, \omega) - \langle K(\vec{y}) \rangle) \rangle = \sigma^2 \varrho(\vec{x}, \vec{y}), \quad \vec{x}, \vec{y} \in D,$$

admits a proper orthogonal decomposition (see Loéve, 1960) or KL expansion

$$K(\vec{x}, \omega) = \mu + \sigma \sum_{i=1}^{\infty} \sqrt{\lambda_i} c_i(\vec{x}) \xi_i, \tag{2.9}$$

where $\mu = \langle K(\vec{x}) \rangle$, the random variables $\{\xi_1, \xi_2, \ldots\}$ are 'uncorrelated' and $\{\lambda_i, c_i(\vec{x})\}$ are the set of eigenvalues and eigenfunctions of $\varrho(\vec{x}, \vec{y})$. Here, $C(\cdot, \cdot)$ is non-negative definite, the eigenvalues are real and we label them in descending order $\lambda_1 > \lambda_2 > \ldots$. Now we can employ the truncated expansion

$$K(\vec{x}, \vec{\xi}) \approx K_M(\vec{x}, \vec{\xi}) = \mu + \sigma \sum_{i=1}^{M} \sqrt{\lambda_i} c_i(\vec{x}) \xi_i, \tag{2.10}$$

for computational purposes in (2.5). This choice is motivated by the fact that quadratic mean square convergence of $K_M(\vec{x}, \vec{\xi})$ to $K(\vec{x}, \omega)$ is guaranteed as $M \to \infty$. The truncation criterion, and hence the choice of M, is usually based on the speed of decay of the eigenvalues since $|D|\text{Var(K)} = \sum \lambda_i$. Hence, in applications where the eigenvalues decay slowly, due to small correlation lengths, M might be very large. However, care must be taken to ensure that for the chosen M, (2.5) is well posed. For the conventional analysis, we require that the truncated coefficient be strictly positive and bounded, and thus satisfy

$$0 < k_1 \leqslant K_M(\vec{x}, \omega) \leqslant k_2 < \infty \quad \text{a.e. in } D \times \Gamma.$$

This is not the same as (2.1). In Babuška & Chatzipantelidis (2002), it is shown that we require $K_M(\vec{x}, \omega)$ to converge to $K(\vec{x}, \omega)$ uniformly as $M \to \infty$. To achieve this, we require that $\xi_1, \ldots, \xi_M$ have uniformly bounded images and that the eigenvalues of the covariance function decay sufficiently fast (see Frauenfelder *et al.*, 2005). Gaussian random variables have unbounded images, but are widely used by practitioners, as we previously mentioned, because uncorrelated Gaussian random variables are independent and this simplifies (2.5). At a discrete level, it often seems that Gaussian random variables are adequate since, for a fixed variance, it is always possible to choose the parameters M and p so that the system matrix to be defined in (3.4) is positive definite. This is misleading since it is not clear what solution is being approximated. Indeed, we will show that it is always possible to choose values of M and p that lead to an indefinite or singular system matrix. Using random variables with bounded images is not a simple fix, however. Uncorrelated random variables with bounded images are not necessarily independent. Independent random variables are assumed for the standard stochastic finite-element technology described here. If it is not fitting, however, to assume independence of the random variables, alternative techniques such as those considered in Babuška *et al.* (2007) and Eiermann *et al.* (2007) should be considered.

2.1.1 *Truncated KL expansion.* To illustrate the positivity issue, consider the following example. The covariance function employed in Ghanem & Spanos (2003), Elman & Furnival (2007), and Deb *et al.* (2001) and in the MATLAB-based code described in Sudret & Der Kiureghian (2000) is

$$C(\vec{x}, \vec{y}) = \sigma^2 \exp\left(-\frac{|x_1 - y_1|}{c_1} - \frac{|x_2 - y_2|}{c_2}\right), \tag{2.11}$$

where c_1 and c_2 are correlation lengths and $D = [-a, a] \times [-a, a]$. The attraction of working with (2.11) is that analytical expressions for the eigenfunctions and eigenvalues exist. To see this, note that the kernel is separable and so the eigenfunctions and eigenvalues can be expressed as the products of those of two corresponding 1D problems. That is, $c_i(\vec{x}) = c_k^1(x_1)c_j^2(x_2)$ and $\lambda_i = \lambda_k^1\lambda_j^2$, where the eigenpairs $\{c_k^1(x_1), \lambda_k^1\}_{k=1}^{\infty}$ and $\{c_j^2(x_2), \lambda_j^2\}_{j=1}^{\infty}$ are solutions to

$$\begin{aligned}
\int_{-a}^{a} \exp(-b_1|x_1 - y_1|)c_k^1(y_1)\mathrm{d}y_1 &= \lambda_k^1 c_k^1(x_1), \\
\int_{-a}^{a} \exp(-b_2|x_2 - y_2|)c_j^2(y_2)\mathrm{d}y_2 &= \lambda_j^2 c_j^2(x_2),
\end{aligned} \tag{2.12}$$

with $b_i = c_i^{-1}$, $i = 1, 2$. Solutions are given in Ghanem & Spanos (2003). As i increases, the eigenfunctions become more oscillatory. The more random variables we use to represent $K(\vec{x}, \vec{\xi})$, the more scales of fluctuation we incorporate. In Fig. 1, we plot a sample of the eigenfunctions for the case $c_1 = 1 = c_2$. In Fig. 2, we plot three realizations of the corresponding truncated coefficient (2.10) with standard deviation $\sigma = 0.5$. Observe that in one of these realizations ($M = 50$), the truncated KL expansion is not strictly positive. This fits with theoretical arguments given in Babuška & Chatzipantelidis (2002). The truncated coefficient is not strictly positive a.e in $D \times \Omega$. However, if the variance is 'not large', we can still choose M and p so that the discrete SFEM system has a positive-definite system matrix. We investigate this issue further in Sections 3 and 4.

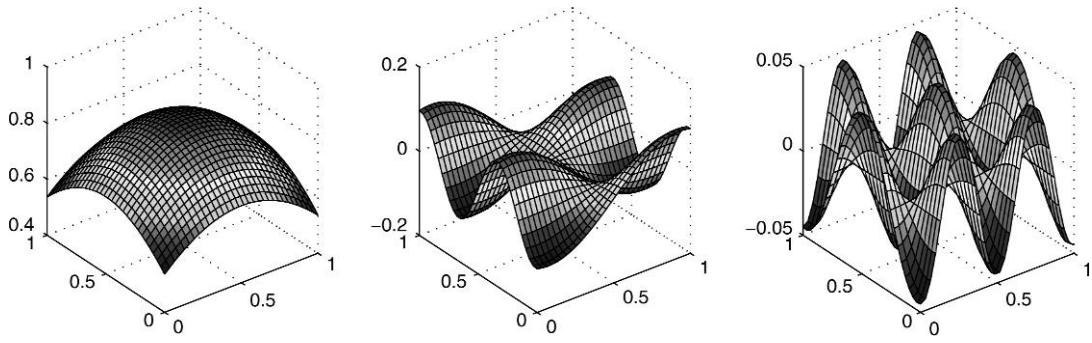

FIG. 1. First, 20th and 50th eigenfunctions of the covariance kernel in (2.11) with $c_1 = c_2 = 1$.

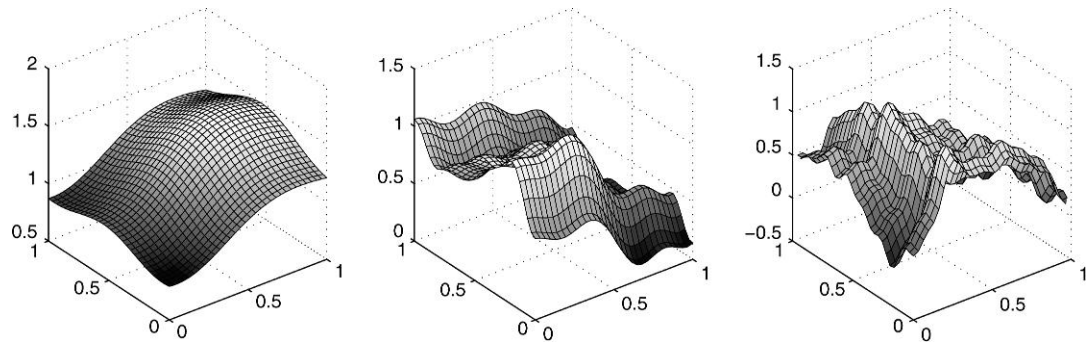

FIG. 2. Realizations of $K_M(\vec{x}, \omega)$ with $M = 5, 20, 50$, $c_1 = 1 = c_2$, $\mu = 1$, $\sigma = 0.5$ and $\xi_i \sim N(0, 1)$, $i = 1 : M$.

3. Linear algebra aspects of spectral SFEM formulation

Given $K_M(\vec{x}, \vec{\xi})$ and bases for X_h and S, we now seek a finite-dimensional solution $u_{hp}(\vec{x}, \vec{\xi}) \in W^h = X_h \otimes S$ satisfying

$$\int_\Gamma \rho(\vec{\xi}) \int_D K_M(\vec{x}, \vec{\xi}) \nabla u_{hp}(\vec{x}, \vec{\xi}) \cdot \nabla w(\vec{x}, \vec{\xi}) \mathrm{d}\vec{x}\, \mathrm{d}\vec{\xi} = \int_\Gamma \rho(\vec{\xi}) \int_D f(\vec{x}) w(\vec{x}, \vec{\xi}) \mathrm{d}\vec{x}\, \mathrm{d}\vec{\xi} \tag{3.1}$$

$\forall\, w(\vec{x}, \vec{\xi}) \in W^h$. Expanding the solution and the test functions in the chosen bases in (3.1), we see that

$$u_{hp}(\vec{x}, \vec{\xi}) = \sum_{s=1}^{N_\xi} \sum_{r=1}^{N_x} u_{r,s} \phi_r(\vec{x}) \psi_s(\vec{\xi}) = \sum_{s=1}^{N_\xi} u_s \psi_s(\vec{\xi}), \tag{3.2}$$

leads to a linear system $A\underline{u} = \underline{f}$ of dimension $N_x N_\xi \times N_x N_\xi$ with block-structure

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,N_\xi} \\ A_{2,1} & A_{2,2} & \dots & A_{2,N_\xi} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N_\xi,1} & A_{N_\xi,2} & \dots & A_{N_\xi,N_\xi} \end{pmatrix}, \quad \underline{u} = \begin{pmatrix} \underline{u}_1 \\ \underline{u}_2 \\ \vdots \\ \underline{u}_{N_\xi} \end{pmatrix}, \quad \underline{f} = \begin{pmatrix} \underline{f}_1 \\ \underline{f}_2 \\ \vdots \\ \underline{f}_{N_\xi} \end{pmatrix}. \tag{3.3}$$

The blocks of A are linear combinations of $M+1$ weighted stiffness matrices of dimension N_x, each with a sparsity pattern equivalent to that of the corresponding deterministic problem. That is,

$$A_{r,s} = \langle \psi_r(\vec{\xi})\psi_s(\vec{\xi})\rangle K_0 + \sum_{k=1}^{M} \langle \xi_k \psi_r(\vec{\xi})\psi_s(\vec{\xi})\rangle K_k,$$

$$K_0(i,j) = \int_D \mu \nabla\phi_i(\vec{x})\nabla\phi_j(\vec{x})\mathrm{d}\vec{x}, \quad K_k(i,j) = \sigma\sqrt{\lambda_k}\int_D c_k(\vec{x})\nabla\phi_i(\vec{x})\nabla\phi_j(\vec{x})\mathrm{d}\vec{x}, \tag{3.4}$$

where $k = 1:M$ and $\mu = \langle K(\vec{x})\rangle$. K_0 contains the mean information of the permeability coefficient, while the other K_k blocks represent fluctuations. In tensor product notation, we have

$$A = G_0 \otimes K_0 + \sum_{k=1}^{M} G_k \otimes K_k, \quad f = \underline{g}_0 \otimes \underline{f}_0, \tag{3.5}$$

where the stochastic matrices G_k are defined via

$$G_0(r,s) = \langle \psi_r, \psi_s\rangle, \quad G_k(r,s) = \langle \xi_k \psi_r \psi_s\rangle, \quad k = 1:M, \tag{3.6}$$

and the vectors $\underline{g}_0$ and $\underline{f}_0$ are given by $\underline{g}_0(i) = \langle \psi_i\rangle$, $\underline{f}_0(i) = \int_D f(\vec{x})\phi(\vec{x})\mathrm{d}\vec{x}$. Since the stochastic basis functions are orthogonal with respect to the probability measure of the distribution of the chosen random variables, G_0 is diagonal. If doubly orthogonal polynomials are used (see Babuška *et al.*, 2004), then each G_k is diagonal, so that A is block-diagonal. This can be handled very easily by solving N_ξ decoupled systems of dimension N_x. We do not consider that case here.

The block-structure of A obtained from the spectral SFEM is illustrated in Fig. 3. Many of the coefficients in the summation in (3.4) are zero, due to the orthogonality properties of the stochastic basis functions (see Section 3.1), and the matrix is highly sparse in a block sense. In particular, K_0 occurs only on the main diagonal blocks. It should also be noted that A is never fully assembled. As pointed out in Ghanem & Kruger (1996), we store only $M+1$ matrices of dimension $N_x \times N_x$ and the entries of each G_k in (3.6). If the discrete problem is well posed, then A is symmetric and positive definite but is ill conditioned with respect to the discretization parameters. We can solve the system

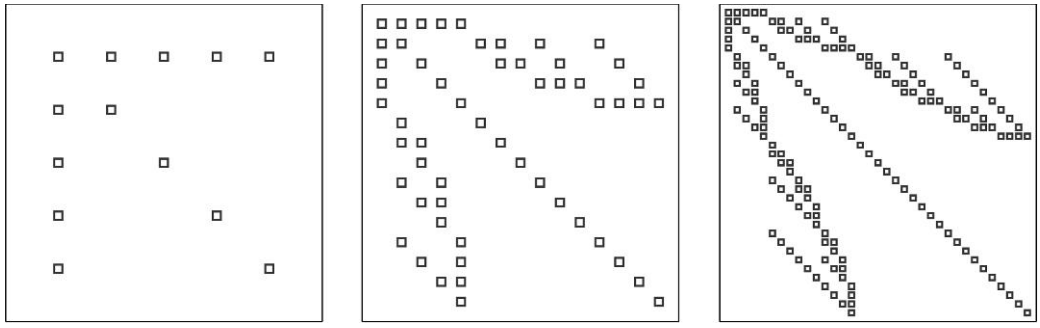

FIG. 3. Matrix block-structure (each block has dimension $N_x \times N_x$), $M = 4$ with $p = 1, 2, 3$ (left to right).

iteratively, using the conjugate gradient (CG) method (performing matrix–vector products intelligently) but a preconditioner is required. We discuss this in Section 3.1.

3.1 *Matrix properties*

We examine first the properties of the stochastic G matrices in (3.6). Each stochastic basis function $\psi_i(\vec{\xi})$ is the product of M univariate orthogonal polynomials. That is, $\psi_i(\vec{\xi}) = \psi_{i_1}(\xi_1)\psi_{i_2}(\xi_2)\cdots\psi_{i_M}(\xi_M)$, where the index i into the stochastic basis is identified with a multi-index $i = (i_1, \ldots, i_M)$, $\sum i_s \leqslant p$, where p is the total polynomial degree, and M is the number of random variables retained in (2.10). If Gaussian random variables are used, each $\psi_{i_s}(\xi_s)$ is a univariate Hermite polynomial of degree i_s. If uniform random variables are more appropriate, we use Legendre polynomials. The ordering of these multi-indices is not important for the calculations but some simple eigenvalue bounds for these matrices are obvious if a specific ordering is used.

Using orthogonality of the polynomials and independence of the random variables yields

$$G_0(i,j) = \int_\Gamma \psi_i(\vec{\xi})\psi_j(\vec{\xi})\rho(\vec{\xi})\mathrm{d}\vec{\xi} = \prod_{s=1}^{M}\int_{\Gamma_s} \psi_{i_s}(\xi_s)\psi_{j_s}(\xi_s)\rho_s(\xi_s)\mathrm{d}\xi_s = \prod_{s=1}^{M}\left\langle \psi_{i_s}^2(\xi_s)\right\rangle \delta_{i_s,j_s}.$$

G_0 is a diagonal matrix and is the identity matrix if the stochastic basis functions are normalized. For example, if Hermite polynomials in Gaussian random variables on the interval $(-\infty, \infty)$ are employed, we obtain

$$G_0(i,j) = \prod_{s=1}^{M} i_s!\delta_{i_s,j_s} = \left(\prod_{s=1}^{M} i_s!\right)\delta_{i,j} = \begin{cases} \prod_{s=1}^{M} i_s!, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

Using, in addition, the well-known three-term recurrence for the Hermite polynomials,

$$\psi_{k+1}(x) = x\psi_k(x) - k\psi_{k-1}(x) \tag{3.8}$$

for $k = 1 : M$, we obtain

$$\begin{aligned} G_k(i,j) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \xi_k\psi_i(\vec{\xi})\psi_j(\vec{\xi})\rho(\vec{\xi})\mathrm{d}\vec{\xi} \\ &= \left(\prod_{s=1,s\neq k}^{M}\left\langle \psi_{i_s}(\xi_s)\psi_{j_s}(\xi_s)\right\rangle\right)\left\langle \xi_k\psi_{i_k}(\xi_k)\psi_{j_k}(\xi_k)\right\rangle \\ &= \begin{cases} \left(\prod_{s=1,s\neq k}^{M} i_s!\delta_{i_s,j_s}\right)(i_k+1)!, & \text{if } i_k = j_k - 1, \\ \left(\prod_{s=1,s\neq k}^{M} i_s!\delta_{i_s,j_s}\right)i_k!, & \text{if } i_k = j_k + 1, \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \left(\prod_{s=1,s\neq k}^{M} i_s!\right)(i_k+1)!, & \text{if } i_k = j_k - 1 \text{ and } i_s = j_s,\ s = \{1:M\}\setminus\{k\}, \\ \left(\prod_{s=1}^{M} i_s!\right), & \text{if } i_k = j_k + 1 \text{ and } i_s = j_s,\ s = \{1:M\}\setminus\{k\}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Due to (3.8), G_k has at most two nonzero entries per row. $G_k(i, j)$ is nonzero only when the multi-indices corresponding to i and j agree in all components except the kth one, where the entries differ by one. This is true when the basis is built from any set of univariate orthogonal polynomials.

In Section 3.2, it will be necessary to have a handle on the eigenvalues of $G_0^{-1}G_k$ or, equivalently, the symmetrically preconditioned matrices

$$\hat{G}_k = G_0^{-\frac{1}{2}} G_k G_0^{-\frac{1}{2}}, \quad k = 1 : M,$$

for a fixed value of p. The next result relies on the well-known fact that roots of orthogonal polynomials are eigenvalues of certain tridiagonal matrices (see Golub & Welsch (1969) or a standard numerical analysis text such as Stoer & Bulirsch (1980)).

LEMMA 3.1 If Hermite polynomials of total degree p in M Gaussian random variables are used for the stochastic basis, the eigenvalues of $\hat{G}_k = G_0^{-1/2} G_k G_0^{-1/2}$, for each $k = 1 : M$, lie in the interval $[-H_{p+1}^{\max}, H_{p+1}^{\max}]$, where $H_{p+1}^{\max}$ is the maximum positive root of the univariate Hermite polynomial of degree $p + 1$.

Proof. Using the definitions of G_0 and G_k, observe that $\hat{G}_k$ has at most two nonzeroes per row:

$$\hat{G}_k(i, j) = \begin{cases} \dfrac{\left(\prod_{s=1,\, s\neq k}^{M} i_s!\right)(i_k+1)!}{\sqrt{\left(\prod_{s=1}^{M} i_s!\right)}\sqrt{\left(\prod_{s=1}^{M} j_s!\right)}}, & \text{if } i_k = j_k - 1 \text{ and } i_s = j_s,\ s = \{1 : M\} \setminus \{k\}, \\ \dfrac{\left(\prod_{s=1}^{M} i_s!\right)}{\sqrt{\left(\prod_{s=1}^{M} i_s!\right)}\sqrt{\left(\prod_{s=1}^{M} j_s!\right)}}, & \text{if } i_k = j_k + 1 \text{ and } i_s = j_s,\ s = \{1 : M\} \setminus \{k\}, \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \sqrt{i_k + 1}, & \text{if } i_k = j_k - 1 \text{ and } i_s = j_s,\ s = \{1 : M\} \setminus \{k\}, \\ \sqrt{i_k}, & \text{if } i_k = j_k + 1 \text{ and } i_s = j_s,\ s = \{1 : M\} \setminus \{k\}, \\ 0, & \text{otherwise.} \end{cases}$$

Let M and p be fixed but arbitrary and consider, first, the matrix $\hat{G}_1$. It is possible to choose an ordering of the stochastic basis functions that causes $\hat{G}_1$ to be block tridiagonal. Recall that the sum of the multi-index components does not exceed p. First, list multi-indices with first component ranging from 0 to p with entries in the second to Mth components summing to zero: $(0, 0, \ldots, 0), (1, 0, \ldots, 0), \ldots, (p, 0, \ldots, 0)$. This accounts for $p + 1$ basis functions. Given the definition of $\hat{G}_1$, the leading $(p + 1) \times (p + 1)$ block, namely T_{p+1}, is then necessarily tridiagonal

$$T_{p+1} = \begin{pmatrix} 0 & 1 & & & & \\ 1 & 0 & \sqrt{2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \sqrt{p-1} & 0 & \sqrt{p} & \\ & & & \sqrt{p} & 0 & \end{pmatrix}. \tag{3.9}$$

Next, we list multi-indices with first components ranging from 0 to $p-1$ and with entries in the second to Mth components that add up to one, but grouped to have the same entries in those components:

$$\begin{array}{cccc}
(0,0,\ldots,0,1) & (0,0,\ldots,1,0) & \cdots\cdots & (0,1,\ldots,0,0) \\
(1,0,\ldots,0,1) & (1,0,\ldots,1,0) & \cdots\cdots & (1,1,\ldots,0,0) \\
\vdots & \vdots & \cdots\cdots & \vdots \\
(p-1,0,\ldots,0,1) & (p-1,0,\ldots,1,0) & \cdots\cdots & (p-1,1,\ldots,0,0)
\end{array}$$

This accounts for $(M-1)\times p$ basis functions. $\hat{G}_1$ then has $M-1$ copies of a tridiagonal matrix T_p defined analogously to T_{p+1}. We continue to order the multi-indices in this way until, finally, we list multi-indices that are 0 in the first component and in the second to Mth components are the same and have entries that add up to p. Then, $\hat{G}_1$ is a symmetric block tridiagonal matrix with multiple copies of the symmetric tridiagonal matrices $T_{p+1}, T_p, \ldots, T_1 = 0$ as the diagonal blocks. The number of copies of T_{p+1} is one and the number of copies of T_j, $j = 1:p$, that appear is

$$\frac{1}{(p-j+1)!}\prod_{r=0}^{p-j}(M-1+r).$$

The eigenvalues of $\hat{G}_1$ are the eigenvalues of the $\{T_j\}$. The eigenvalues of each tridiagonal block are just roots of a characteristic polynomial $p_j(\lambda)$ that satisfies the recursion (3.8). That is,

$$p_{j+1}(\lambda) = (\lambda - 0)p_j(\lambda) - (-\sqrt{j})^2 p_{j-1}(\lambda).$$

Hence, $p_j(\lambda)$ is the Hermite polynomial of degree j (see Golub & Welsch, 1969, or Stoer & Bulirsch, 1980, Chapter 3). Since the roots of lower-degree Hermite polynomials are bounded by the extremal eigenvalues of higher-degree polynomials, the maximum eigenvalue of $\hat{G}_1$ is the maximum root of the $(p+1)$th-degree polynomial H_{p+1} or, equivalently, the maximum eigenvalue of T_{p+1}. The minimum eigenvalue is identical to the maximum eigenvalue but with a sign change.

Now, if the basis functions have not been chosen to give $\hat{G}_1$ explicitly as a block tridiagonal matrix, there exists a permutation matrix P_1 (corresponding to a reordering of the stochastic basis functions) such that $\tilde{G}_1 = P_1\hat{G}_1P_1^{\mathrm{T}}$ is the block tridiagonal matrix described above. The eigenvalues of $\tilde{G}_1$ are the same as those of $\hat{G}_1$. The same argument applies for the other matrices $\hat{G}_k$, $k = 2:M$. There exists a permutation matrix P_k so that $\tilde{G}_k = P_k\hat{G}_kP_k^{\mathrm{T}}$ is block tridiagonal and whose extremal eigenvalues are given by those of T_{p+1}. □

REMARK 3.2 The above result refers specifically to Gaussian random variables. However, it can be easily extended to other types of random variables. The stochastic basis is always constructed from a set of orthogonal univariate polynomials that satisfy a three-term recurrence. Hence, $\hat{G}_k$ is always a permutation of a symmetric, block tridiagonal matrix. The characteristic polynomial for the eigenvalues of each block always inherits the same three-term recurrence as the original set of orthogonal polynomials. Further discussion and generalization of this point, as well as a discussion of other properties of the matrices $\hat{G}_k$, can be found in Ernst & Ullmann (2008).

We specify the result in the case of uniform random variables as follows.

LEMMA 3.3 If Legendre polynomials of total degree p in M uniform random variables with support on the bounded symmetric interval $[-\gamma, \gamma]$ are used for the stochastic basis, the eigenvalues of $\hat{G}_k$, for each $k = 1 : M$, lie in the interval $[-L_{p+1}^{\max}, L_{p+1}^{\max}]$, where $L_{p+1}^{\max}$ is the maximum positive root of the univariate Legendre polynomial of degree $p + 1$.

Proof. Follow the proof of Lemma 3.1, replacing the definitions of G_0 and G_k and the three-term recurrence (3.8) by those appropriate to the Legendre polynomials on the interval $[-\gamma, \gamma]$. □

For Hermite polynomials, $H_{p+1}^{\max}$ is bounded by $\sqrt{p-1} + \sqrt{p}$. This is observed by applying Gershgorin's theorem to T_{p+1} in (3.9). In contrast, for Legendre polynomials, $L_{p+1}^{\max}$ is bounded by γ independently of p. If the uniform random variables have mean zero and unit variance, then $\gamma = \sqrt{3}$.

We now examine the eigenvalues of the matrices K_k coming from the spatial discretization.

LEMMA 3.4 Let K_0 and K_k be the stiffness matrices defined in (3.4). If $c_k(\vec{x}) \geqslant 0$, where $\{\lambda_k, c_k(\vec{x})\}$ is the kth eigenpair of $\varrho(\vec{x}, \vec{y})$, then

$$0 \leqslant \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\min} \leqslant \frac{\underline{x}^{\mathrm{T}}K_k\underline{x}}{\underline{x}^{\mathrm{T}}K_0\underline{x}} \leqslant \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\max} \quad \forall\, \underline{x} \in \mathbb{R}^{N_x},$$

where $c_k^{\min} = \inf_{\vec{x}\in D} c_k(\vec{x})$ and $c_k^{\max} = \sup_{\vec{x}\in D} c_k(\vec{x}) = \|c_k(\vec{x})\|_\infty$. Alternatively, if $c_k(\vec{x})$ is not uniformly positive, then

$$-\frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty \leqslant \frac{\underline{x}^{\mathrm{T}}K_k\underline{x}}{\underline{x}^{\mathrm{T}}K_0\underline{x}} \leqslant \frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty \quad \forall\, \underline{x} \in \mathbb{R}^{N_x}.$$

Proof. Given any $\underline{x} \in \mathbb{R}^{N_x}$, define a function $v \in X_h$ via $v = \sum x_i\phi_i(\vec{x})$. If $c_k(\vec{x}) \geqslant 0$, then

$$\underline{x}^{\mathrm{T}}K_k\underline{x} = \int_D \sigma\sqrt{\lambda_k}c_k(\vec{x})\nabla v \cdot \nabla v \,\mathrm{d}D \leqslant \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\max}\int_D \mu\nabla v \cdot \nabla v \,\mathrm{d}D = \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\max}\underline{x}^{\mathrm{T}}K_0\underline{x},$$

$$\underline{x}^{\mathrm{T}}K_k\underline{x} = \int_D \sigma\sqrt{\lambda_k}c_k(\vec{x})\nabla v \cdot \nabla v \,\mathrm{d}D \geqslant \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\min}\int_D \mu\nabla v \cdot \nabla v \,\mathrm{d}D = \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\min}\underline{x}^{\mathrm{T}}K_0\underline{x}.$$

If μ is positive, dividing through by the quantity $\underline{x}^{\mathrm{T}}K_0\underline{x}$ gives the first result. Now, if $c_k(\vec{x})$ also takes on negative values, we have

$$|\underline{x}^{\mathrm{T}}K_k\underline{x}| = \sigma\sqrt{\lambda_k}\left|\int_D c_k(\vec{x})\nabla v \cdot \nabla v \,\mathrm{d}D\right| \leqslant \frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty\underline{x}^{\mathrm{T}}K_0\underline{x}.$$

Arguing as in the first case gives the second result. □

3.2 *Preconditioning*

When the global matrix A is symmetric and positive definite, we can use the CG method as a solver. However, the system is ill conditioned and a preconditioner is required. In Pellissetti & Ghanem (2000) and Ghanem & Kruger (1996), it is noted that if the variance of $K(\vec{x}, \omega)$ is small, then the preconditioner P composed of the diagonal blocks of A, i.e.

$$P = G_0 \otimes K_0, \tag{3.10}$$

is heuristically the simplest and most appropriate choice. Working under the assumption that the variance is 'sufficiently small' is not suitable for some applications but the user is, in fact, limited to this if employing Hermite polynomials in Gaussian random variables. In this section, we obtain a theoretical handle on these observations. First, we explain why preconditioning is required.

LEMMA 3.5 If G_0 is defined using Hermite polynomials in Gaussian random variables and piecewise linear (or bilinear) approximation is used for the spatial discretization, on quasi-uniform meshes, the eigenvalues of $(G_0 \otimes K_0)$ lie in the interval $[\mu\alpha_1 h^2, \mu\alpha_2 p!]$, where μ is the mean value of $K(\vec{x}, \omega)$, p is the degree of stochastic polynomials, h is the characteristic spatial mesh-size and α_1 and α_2 are constants independent of h, M and p.

Proof. If $\underline{v}_\xi$ is an eigenvector of G_0 with corresponding eigenvalue λ_ξ and $\underline{v}_x$ is an eigenvector of K_0 with corresponding eigenvalue λ_x, then $(G_0 \otimes K_0)(\underline{v}_\xi \otimes \underline{v}_x) = \lambda_\xi \lambda_x (\underline{v}_\xi \otimes \underline{v}_x)$. Using (3.7), we deduce that $1 \leqslant \lambda_\xi \leqslant p!$. A bound for the eigenvalues of K_0 can be obtained in the usual way, e.g. see Elman *et al.* (2005b, pp. 57–59), to give

$$\mu\alpha_1 h^2 \leqslant \frac{\underline{v}_x^{\mathrm{T}} K_0 \underline{v}_x}{\underline{v}_x^{\mathrm{T}} \underline{v}_x} \leqslant \mu\alpha_2 \quad \forall\, \underline{v}_x \in \mathbb{R}^{N_x}.$$

The result immediately follows. □

REMARK 3.6 Note that if the polynomial chaos basis functions are normalized with $\langle \psi_i, \psi_j \rangle = \delta_{ij}$, then with any choice of random variables, the stochastic mass matrix is the identity matrix and the above eigenvalue bound is simply $[\mu\alpha_1 h^2, \mu\alpha_2]$ and is independent of p. It is always worthwhile normalizing the basis functions for this reason. We shall assume that this is the case in the sequel.

Now, we can expect the eigenvalues of the global unpreconditioned system matrix (3.5) to be a perturbation of the eigenvalues of $G_0 \otimes K_0$.

LEMMA 3.7 If the matrices G_k in (3.6) are defined using either normalized Hermite polynomials in Gaussian random variables or normalized Legendre polynomials in uniform random variables on a bounded symmetric interval $[-\gamma, \gamma]$, and piecewise linear (or bilinear) approximation is used for the spatial discretization, on quasi-uniform meshes, then the eigenvalues of the global stiffness matrix A in (3.5) are bounded and lie in the interval $[\mu\alpha_1 h^2 - \delta, \mu\alpha_2 + \delta]$, where

$$\delta = \alpha_2 \sigma C_{p+1}^{\max} \sum_{k=1}^{M} \sqrt{\lambda_k} \|c_k(\vec{x})\|_\infty,$$

$C_{p+1}^{\max}$ is the maximal root of an orthogonal polynomial of degree $p+1$, h is the spatial discretization parameter and α_1 and α_2 are constants independent of h, M and p.

Proof. First note that the maximum and minimum eigenvalues $\nu_{\max}$ and $\nu_{\min}$ of

$$\left((G_0 \otimes K_0) + \sum_{k=1}^{M} (G_k \otimes K_k) \right) \underline{v} = \nu \underline{v}$$

can be bounded in terms of the maximum and minimum eigenvalues of the matrices in the sum. Using normalized stochastic basis functions, the matrices $\hat{G}_k$ in Lemmas 3.1 and 3.3 are the same as the matrices G_k in (3.6). Hence, the eigenvalues of G_k belong to the symmetric interval $[-C_{p+1}^{\max}, C_{p+1}^{\max}]$,

where $C_{p+1}^{\max}$ is equal to $H_{p+1}^{\max}$ or $L_{p+1}^{\max}$. Using a similar argument to that presented in Lemma 3.4, the eigenvalues of K_k, $k = 1 : M$, lie in the bounded interval

$$\begin{aligned} &[\sigma\sqrt{\lambda_k}c_k^{\min}\alpha_1 h^2,\ \sigma\sqrt{\lambda_k}c_k^{\max}\alpha_2], && \text{if } c_k(\vec{x}) \geqslant 0,\\ &[-\sigma\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty\alpha_2,\ \sigma\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty\alpha_2], && \text{otherwise.}\end{aligned}$$

Denoting the minimum and maximum eigenvalues of $(G_k \otimes K_k)$ by $\gamma_{\min}^k$ and $\gamma_{\max}^k$, respectively, and applying the result of Lemma 3.5, we have

$$\nu_{\min} \geqslant \mu\alpha_1 h^2 + \sum_{k=1}^{M} \gamma_{\min}^k, \quad \nu_{\max} \leqslant \mu\alpha_2 + \sum_{k=1}^{M} \gamma_{\max}^k.$$

Now, noting that the eigenvalues of the Kronecker product of two matrices are the products of the eigenvalues of the individual matrices, we have, for any k,

$$\nu_{\min} \geqslant \mu\alpha_1 h^2 - \sigma\alpha_2 C_{p+1}^{\max} \sum_{k=1}^{M} \sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty, \quad \nu_{\max} \leqslant \mu\alpha_2 + \sigma\alpha_2 C_{p+1}^{\max} \sum_{k=1}^{M} \sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty.$$

□

Using the preceding arguments, we can now establish a result that determines the efficiency of the chosen preconditioner.

THEOREM 3.8 The eigenvalues $\{\nu_i\}$ of the generalized eigenvalue problem, $A\underline{x} = \nu P\underline{x}$, where the matrices G_k are defined using either normalized Hermite polynomials in Gaussian random variables or normalized Legendre polynomials in uniform random variables on a bounded symmetric interval $[-\gamma, \gamma]$, lie in the interval $[1 - \tau, 1 + \tau]$, where

$$\tau = \frac{\sigma}{\mu} C_{p+1}^{\max} \sum_{k=1}^{M} \sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty, \tag{3.11}$$

σ and μ are the standard deviation and mean of $K(\vec{x}, \omega)$, $\{\lambda_k, c_k(\vec{x})\}$ are the eigenpairs of $\rho(\vec{x}, \vec{y})$ and $C_{p+1}^{\max}$ is a constant (possibly) depending on p.

Proof. First note that the eigenvalues that we are seeking satisfy $\nu = \theta + 1$, where

$$\sum_{k=1}^{M} (G_0 \otimes K_0)^{-1}(G_k \otimes K_k)\underline{v} = \theta\underline{v}.$$

Hence, using standard properties of the matrix Kronecker product, and assuming normalized stochastic basis functions, we have

$$\sum_{k=1}^{M} (G_k \otimes K_0^{-1}K_k)\underline{v} = \theta\underline{v}.$$

Now, let $\hat{K}_k = K_0^{-1}K_k$. Applying Lemmas 3.4, 3.1 and 3.3, the eigenvalues of G_k belong to the symmetric interval $[-C_{p+1}^{\max}, C_{p+1}^{\max}]$, where $C_{p+1}^{\max}$ is equal to $H_{p+1}^{\max}$ or $L_{p+1}^{\max}$ depending, on the choice of random

variables, and the eigenvalues of $\hat{K}_k$ belong to the interval

$$\left[\frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\min}, \frac{\sigma}{\mu}\sqrt{\lambda_k}c_k^{\max}\right] \quad \text{or} \quad \left[-\frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty, \frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty\right],$$

depending on the positivity of $c_k(\vec{x})$. Proceeding as in Lemma 3.7, and denoting the minimum and maximum eigenvalues of $G_k \otimes \hat{K}_k$ by $\gamma_{\min}^k$ and $\gamma_{\max}^k$, we have, in both cases,

$$\theta_{\min} \geqslant \sum_{k=1}^{M} \gamma_{\min}^k \geqslant -\sum_{k=1}^{M} C_{p+1}^{\max}\frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty, \quad \theta_{\max} \leqslant \sum_{k=1}^{M} \gamma_{\max}^k \leqslant \sum_{k=1}^{M} C_{p+1}^{\max}\frac{\sigma}{\mu}\sqrt{\lambda_k}\|c_k(\vec{x})\|_\infty.$$

The eigenvalues that we need are the values $\nu_i = 1 + \theta_i$, $i = 1 : N_x N_\xi$. □

REMARK 3.9 As $\sigma\mu^{-1} \to 0$, the bound collapses to a single cluster at one. This is intuitively correct, since the off-diagonal blocks, which are not represented in the preconditioner, become insignificant. For increasing $\sigma\mu^{-1}$, the upper and lower bounds move away from one. The lower bound may be negative. Note that when $\sigma\mu^{-1}$ is too large, the condition (2.1) is violated for even low values of M and the unpreconditioned matrix A is not positive definite.

REMARK 3.10 The bound depends on the value $C_{p+1}^{\max}$. As we have seen, if Gaussian random variables are used, this constant grows like $\sqrt{p-1} + \sqrt{p}$. Hence, the preconditioner P does not improve the conditioning of A with respect to p. If uniform random variables are employed, $C_{p+1}^{\max} = L_{p+1}^{\max} \leqslant \gamma$ and so there is no ill-conditioning in the preconditioned or unpreconditioned systems with respect to p.

REMARK 3.11 The bounds are pessimistic in M, due to the fact that we have bounded the maximum eigenvalue of a sum of matrices by the sum of the maximum eigenvalues of the individual matrices (and similarly with the minimum eigenvalue). The bound is sharp when $M = 1$ with any p, μ and σ (since then there is no sum) and is tighter in M when the eigenvalues decay rapidly or when σ is very small.

For the case $p = 1$, a tighter bound (with respect to M) can be established. We illustrate this below for Gaussian random variables.

THEOREM 3.12 When $p = 1$, for any M, the eigenvalues $\{\nu_i\}$ in $A\underline{x} = \nu P\underline{x}$, where A and P are defined in (3.5) and (3.10) and the matrices G_k are defined as in (3.6) using normalized Hermite polynomials in Gaussian random variables, lie in the interval $[1-\tau, 1+\tau]$, where

$$\tau = \frac{\sigma}{\mu}\left(\sum_{k=1}^{M} \lambda_k \|c_k(\vec{x})\|_\infty^2\right)^{\frac{1}{2}}, \tag{3.12}$$

σ and μ are the standard deviation and mean of $K(\vec{x}, \omega)$ and $\{\lambda_k, c_k(\vec{x})\}$ are the eigenpairs of $\rho(\vec{x}, \vec{y})$.

Proof. When $p = 1$, each $\hat{G}_k$ is a permutation of an $(M+1) \times (M+1)$ block tridiagonal matrix G_* with leading block

$$T_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and all remaining rows and columns filled with zeros. Hence, following the proof of Theorem 3.8,

$$\sum_{k=1}^{M} \hat{G}_k \otimes \hat{K}_k = \begin{pmatrix} 0 & \hat{K}_M & \dots & \hat{K}_1 \\ \hat{K}_M & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{K}_1 & 0 & \dots & 0 \end{pmatrix}$$

(or some block permutation thereof). It is then a trivial task to show that the eigenvalues of this sum are either 0 or $\pm\sqrt{\lambda\big(\sum_{k=1}^{M} \hat{K}_k^2\big)}$ or, in other words, the eigenvalues of the matrix

$$\hat{G}_* \otimes \left(\sum_{k=1}^{M} \hat{K}_k^2\right)^{\frac{1}{2}}$$

(since the eigenvalues of G_* are -1, 0, 1). Using the result of Lemma 3.4, noting that the eigenvalues of $\hat{K}_k^2$ are non-negative, and denoting the maximum eigenvalue of $\hat{K}_k$ by $\gamma_{\max}^k$, we have

$$\theta_{\max} \leqslant \left(\sum_{k=1}^{M} (\gamma_{\max}^k)^2\right)^{\frac{1}{2}} \leqslant \frac{\sigma}{\mu}\left(\sum_{k=1}^{M} \lambda_k \|c_k(\vec{x})\|_\infty^2\right)^{\frac{1}{2}},$$

$$\theta_{\min} \geqslant -\left(\sum_{k=1}^{M} (\gamma_{\max}^k)^2\right)^{\frac{1}{2}} \geqslant -\frac{\sigma}{\mu}\left(\sum_{k=1}^{M} \lambda_k \|c_k(\vec{x})\|_\infty^2\right)^{\frac{1}{2}}.$$

The eigenvalues we need are the values $\nu_i = 1 + \theta_i$, $i = 1 : N_x N_\xi$. □

REMARK 3.13 The above bound is tighter with respect to M as the sequence $\lambda_1, \lambda_2, \dots$ decays more rapidly than the sequence $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots$. Unfortunately, for other values of p there is no nice representation for the sum of matrices in the form $G_* \otimes X$ for some matrix X that is easy to handle.

We now explore the accuracy of the bounds. In each example below, we list the computed extremal eigenvalues of $P^{-1}A$ and the bounds on those eigenvalues calculated using Theorems 3.8 and 3.12. For the stochastic basis, we employ Hermite polynomials in Gaussian random variables.

EXAMPLE 3.14 We consider first the case where the covariance function is (2.11) with $\sigma = 0.1$, $\mu = 1$, $c_1 = 1 = c_2$ and $h = \frac{1}{8}$. Computed eigenvalues and their estimated bounds are listed in Table 1.

EXAMPLE 3.15 Next, we consider the same example but with a very small standard deviation $\sigma = 0.01$. Computed eigenvalues and their estimated bounds are listed in Table 2.

EXAMPLE 3.16 Observe what happens when we use a large standard deviation $\sigma = 0.3$ and increase p, the stochastic polynomial degree. In this example, in Table 3, we also list the extremal eigenvalues of A. Thus, it can be seen that when Hermite polynomials (with infinite support) are employed, for fixed values of h, M and σ, we can always find a value of p that causes the system matrix A and the preconditioned system matrix $P^{-1}A$ to be indefinite. The eigenvalue bounds in Theorems 3.8 and 3.12 predict this. Figure 4 summarizes this for the case $M = 1$.

TABLE 1 *Example 3.14: extremal eigenvalues of $P^{-1}A$ and bounds on extremal eigenvalues of $P^{-1}A$*

M	p	$\nu_{\min}(P^{-1}A)$	$\nu_{\max}(P^{-1}A)$	Bounds	$H_{p+1}^{\max}$
1	1	0.9155	1.0845	[0.9151, 1.0849]	1
	2	0.8537	1.1463	[0.8529, 1.1471]	1.7321
	3	0.8028	1.1972	[0.8017, 1.1983]	2.3344
	4	0.7586	1.2414	[0.7573, 1.2427]	2.8570
2	1	0.9125	1.0875	[0.9037, 1.0963]	1
	2	0.8485	1.1515	[0.7743, 1.2257]	1.7321
	3	0.7959	1.2041	[0.6958, 1.3042]	2.3344
	4	0.7502	1.2498	[0.6277, 1.3723]	2.8570
3	1	0.9107	1.0893	[0.8935, 1.1065]	1
	2	0.8453	1.1547	[0.6957, 1.3043]	1.7321
	3	0.7915	1.2085	[0.5899, 1.4101]	2.3344
	4	0.7449	1.2551	[0.4981, 1.5019]	2.8570

TABLE 2 *Example 3.15: extremal eigenvalues of $P^{-1}A$ and bounds on extremal eigenvalues of $P^{-1}A$*

M	p	$\nu_{\min}(P^{-1}A)$	$\nu_{\max}(P^{-1}A)$	Bounds	$H_{p+1}^{\max}$
1	1	0.9916	1.0170	[0.9915, 1.0085]	1
	2	0.9854	1.0146	[0.9853, 1.0147]	1.7321
	3	0.9803	1.0197	[0.9802, 1.0198]	2.3344
	4	0.9759	1.0241	[0.9757, 1.0243]	2.8570
2	1	0.9913	1.0176	[0.9904, 1.0096]	1
	2	0.9849	1.0151	[0.9774, 1.0226]	1.7321
	3	0.9796	1.0204	[0.9696, 1.0304]	2.3344
	4	0.9750	1.0250	[0.9628, 1.0372]	2.8570
3	1	0.9911	1.0089	[0.9893, 1.0107]	1
	2	0.9845	1.0155	[0.9696, 1.0304]	1.7321
	3	0.9792	1.0208	[0.9590, 1.0410]	2.3344
	4	0.9745	1.0255	[0.9498, 1.0502]	2.8570

EXAMPLE 3.17 Finally, consider the case where the covariance function is (2.11) with $\sigma = 0.1$, $\mu = 1$, $c_1 = 10 = c_2$ and $h = \frac{1}{8}$. Here, the eigenvalues of the covariance functions decay more quickly than in the first two examples. Eigenvalues of the preconditioned system are listed in Table 4.

In all cases, the extremal eigenvalues of $P^{-1}A$ exhibit the behaviour anticipated by the bounds in Theorems 3.8 and 3.12. They are symmetric about one, increase very slightly with p and retract to one for small variance. For small values of σ, the dependence on p is not evident. These results together with Theorem 3.8 tell us that when Gaussian random variables are used, the preconditioned system is positive definite only when the variance and the polynomial degree are not too large. Now we turn to the question of implementation and focus on cases where A is positive definite.

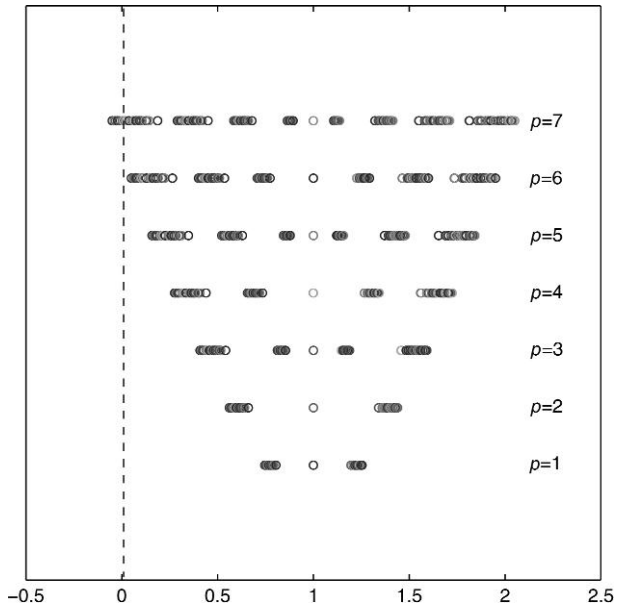

FIG. 4. $h = \frac{1}{8}$. Example 3.16: eigenvalues of $P^{-1}A$, $\sigma = 0.3$, $M = 1$, varying p.

TABLE 3 *Example 3.16: extremal eigenvalues of A and $P^{-1}A$ and bounds on extremal eigenvalues of $P^{-1}A$*

M	p	$\nu_{\min}(A)$	$\nu_{\max}(A)$	$\nu_{\min}(P^{-1}A)$	$\nu_{\max}(P^{-1}A)$	Bounds	$H_{p+1}^{\max}$
1	4	0.1080	6.4326	0.2758	1.7242	[0.2720, 1.7280]	2.8570
	5	0.0756	6.8636	0.1574	1.8426	[0.1529, 1.8471]	3.3243
	6	0.0427	7.2569	0.0493	1.9507	[0.0443, 1.9557]	3.7504
	7	**−0.1545**	**7.6206**	**−0.0506**	**2.0506**	**[−0.0561, 2.0561]**	4.1445
	8	**−0.4640**	**7.9605**	**−0.1439**	**2.1439**	**[−0.1500, 2.1500]**	4.5127
2	4	0.1052	6.5085	0.2505	1.7495	[−0.1169, 2.1169]	2.8570
	5	0.0717	6.9464	0.1279	1.8721	[−0.2996, 2.2996]	3.3243
	6	0.0333	7.3450	0.0161	1.9839	[−0.4662, 2.4662]	3.7504
	7	**−0.2725**	**7.7130**	**−0.0873**	**2.0873**	**[−0.6202, 2.6202]**	4.1445
	8	**−0.5972**	**8.0563**	**−0.1838**	**2.1838**	**[−0.7642, 2.7642]**	4.5127

4. Numerical results

In this section, we present iteration counts and timings for two test problems using Gaussian random variables. We implement block-diagonal preconditioning with CG. The theoretical results above tell us that we can expect the iteration count to be independent of the spatial discretization parameter, h, and almost independent of p (polynomial degree) and M (KL terms). It is required, however, in each CG iteration to approximate the quantity $P^{-1}\underline{r}$, where $\underline{r}$ is a residual error vector. Applying the preconditioner therefore requires N_ξ approximate solutions of subsidiary systems with coefficient matrix K_0. The number of subproblems can be very large for increasing M and p. (See Table 5 for details.) Fortunately, approximately inverting each of the diagonal blocks of the preconditioner is equivalent to solving

TABLE 4 *Example 3.17: extremal eigenvalues of $P^{-1}A$ and bounds on extremal eigenvalues of $P^{-1}A$*

M	p	$\nu_{\min}(P^{-1}A)$	$\nu_{\max}(P^{-1}A)$	Bounds	$H_{p+1}^{\max}$
1	2	0.8298	1.1702	[0.8297, 1.1703]	1.7321
	3	0.7706	1.2294	[0.7704, 1.2296]	2.3344
	4	0.7192	1.2808	[0.7190, 1.2810]	2.8570
2	2	0.8291	1.1709	[0.7961, 1.2039]	1.7321
	3	0.7697	1.2303	[0.7252, 1.2748]	2.3344
	4	0.7182	1.2818	[0.6637, 1.3363]	2.8570
3	2	0.8286	1.1714	[0.7626, 1.2374]	1.7321
	3	0.7689	1.2311	[0.6800, 1.3200]	2.3344
	4	0.7172	1.2828	[0.6084, 1.3916]	2.8570

TABLE 5 *Values of N_ξ (dimension of stochastic basis) for varying M and p*

p	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$	$M=15$	$M=20$	$M=30$
1	3	5	7	9	11	16	21	31
2	6	15	28	45	66	136	231	992
3	10	35	84	165	286	816	1,771	32,736

a standard diffusion problem. Exact solves are too costly for highly refined spatial meshes. However, we can benefit from our experience of solving deterministic problems by replacing the exact solves for K_0 with either an incomplete factorization preconditioner (see Pellissetti & Ghanem, 2000, and Ghanem & Kruger, 1996) or a multigrid V-cycle. In fact, any fast solver for a Poisson problem is a potential candidate. Moreover, the N_ξ approximate solves required at each CG iteration are independent of one another and can be performed in parallel. Crucially, set-up of the approximation to or factorization of K_0 needs to be performed only once.

Below, we implement the preconditioner using both incomplete Cholesky factorization and one V-cycle of AMG with symmetric Gauss–Seidel (SGS) smoothing to approximately invert K_0. The latter method has the key advantage that the computational cost grows linearly in the problem size. Our particular AMG code (see Silvester & Powell, 2007) is implemented in MATLAB and based on the traditional Ruge-Stüben algorithm (see Ruge & Stüben, 1985). No parameters are tuned. We apply the method as a black box in each experiment. Using geometric multigrid to solve these systems is discussed in Elman & Furnival (2007) and Le Maitre *et al.* (2003). All iterations are terminated when the relative residual error, measured in the Euclidean norm, is reduced to 10^{-10}. All computations are performed in serial using MATLAB 7.3 on a laptop PC with 512MB of RAM.

4.1 *Homogeneous Dirichlet boundary condition*

First, we reproduce an experiment performed in Deb *et al.* (2001). The chosen covariance function is (2.11) with $c_1 = 1 = c_2$, standard deviation $\sigma = 0.1$ and mean $\mu = \langle K(\vec{x})\rangle = 1$. We solve (1.2) on

$D = [-0.5, 0.5]\times[-0.5, 0.5]$ with homogenous Dirichlet boundary condition and $f = 2(0.5-x^2-y^2)$. Post-processing the coefficient blocks of the solution in the spectral expansion (3.2) to recover the mean and variance of the solution is trivial. Solutions obtained on a 32×32 uniform spatial grid are plotted in Fig. 5. The maximum values of the mean and variance obtained with $h^{-1} = 16$, $p = 4$ and $M = 6$ are 0.063113 and 2.3600×10^{-05}, respectively. Using the SFEM, a single system of dimension $15^2 \times 210$ is solved. By way of comparison, in Table 6 we record the maximum values of the estimated mean and variance of the pressure solution obtained using a traditional MCM, with N realizations of $K(\vec{x}, \omega)$. The random field inputs were generated using the circulant embedding method described in Dietrich & Newsam (1997), with the same grid used for the spatial discretization. Note that the value $\sigma\mu^{-1}$ is sufficiently small in this example that no negative values of the sampled diffusion coefficients are encountered.

In Table 7, we record iteration counts and timings for preconditioned CG applied to the SFEM systems, with varying h, M and p. Now we can compare implementations based on incomplete Cholesky factorization and on our suggested AMG solver. Note that the performance of the former is sensitive to the choice of drop tolerance parameter and we have not sought to optimize this. The black-box AMG version of the preconditioning scheme proved to be optimal with respect to the spatial discretization without tuning any parameters. Indeed, the matrix V corresponding to a single V-cycle of the AMG algorithm is a spectrally equivalent approximation to K_0 (see Table 8). The maximum eigenvalue of $V^{-1}K_0$ is one, independently of h. The efficiency of this approximation is completely unaffected by the choice of p, M and standard deviation σ.

The efficiency of both implementations of the block-diagonal preconditioner deteriorates with increasing $\sigma\mu^{-1}$. For fixed μ, as σ increases, the off-diagonal blocks of A become more significant and they are not represented in the preconditioner. Iteration counts, for exact preconditioning, for fixed

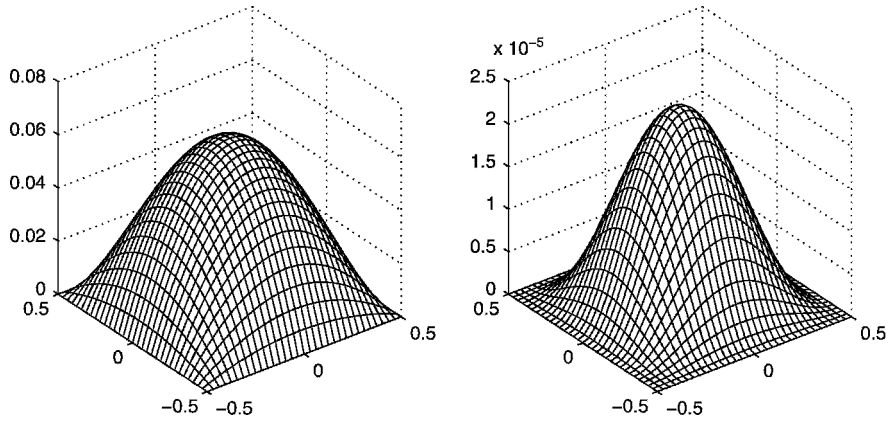

FIG. 5. Mean (left) and variance (right) of pressure on a 32×32 mesh for the case $M = 4$ with $p = 2$.

TABLE 6 *Maximum values of sample mean and standard deviation after N realizations*

	$N = 100$	$N = 1,000$	$N = 10,000$	$N = 40,000$
Max(sample mean)	0.063608	0.063299	0.063127	0.063134
Max(sample variance)	2.1611×10^{-05}	2.4065×10^{-05}	2.2584×10^{-05}	2.3160×10^{-05}

TABLE 7 *Preconditioned CG iterations and timings in seconds (set-up + total iteration times)*

Preconditioner	h	$p = 2$	$p = 3$	$p = 4$
$M = 4$				
None	$\frac{1}{4}$	19	30	55
	$\frac{1}{8}$	43	73	139
	$\frac{1}{16}$	92	161	314
	$\frac{1}{32}$	188	335	666
Block-diagonal	$\frac{1}{16}$	10 (0.00 + 0.33)	11 (0.00 + 1.05)	12 (0.00 + 2.88)
(cholinc, 1×10^{-3})	$\frac{1}{32}$	11 (0.02 + 1.23)	13 (0.01 + 4.24)	15 (0.01 + 9.42)
	$\frac{1}{64}$	21 (0.10 + 9.68)	22 (0.10 + 30.33)	24 (0.09 + 67.74)
	$\frac{1}{128}$	38 (0.61 + 91.44)	42 (0.61 + 272.17)	45 (0.61 + 613.92)
Block-diagonal	$\frac{1}{16}$	10 (0.06 + 0.53)	12 (0.06 + 1.76)	13 (0.13 + 4.50)
(AMG)	$\frac{1}{32}$	11 (0.20 + 1.60)	12 (0.20 + 5.71)	13 (0.29 + 13.41)
	$\frac{1}{64}$	11 (0.88 + 6.38)	12 (0.99 + 20.50)	13 (0.96 + 47.15)
	$\frac{1}{128}$	12 (6.72 + 36.40)	13 (6.84 + 104.64)	14 (5.18 + 233.44)
$M = 6$				
None	$\frac{1}{4}$	19	29	55
	$\frac{1}{8}$	44	76	146
	$\frac{1}{16}$	93	169	332
	$\frac{1}{32}$	190	350	702
Block-diagonal	$\frac{1}{16}$	10 (0.07 + 0.90)	11 (0.00 + 4.11)	12 (0.00 + 17.21)
(cholinc, 1×10^{-3})	$\frac{1}{32}$	13 (0.01 + 3.98)	13 (0.00 + 14.51)	15 (0.01 + 44.00)
	$\frac{1}{64}$	21 (0.09 + 24.17)	23 (0.10 + 91.21)	24 (0.60 + 242.06)
	$\frac{1}{128}$	38 (0.61 + 240.10)	42 (0.68 + 876.49)	46 (0.60 + 2,610.48)
Block-diagonal	$\frac{1}{16}$	11 (0.06 + 1.39)	12 (0.06 + 6.06)	13 (0.06 + 24.02)
(AMG)	$\frac{1}{32}$	11 (0.20 + 4.37)	12 (0.20 + 16.66)	13 (0.20 + 51.93)
	$\frac{1}{64}$	11 (0.88 + 15.67)	13 (0.98 + 58.32)	14 (6.75 + 180.48)
	$\frac{1}{128}$	12 (6.78 + 89.66)	13 (6.75 + 310.61)	14 (6.75 + 886.46)

TABLE 8 $M = 4$. *Minimum eigenvalue of $V^{-1}K_0$ where V is one V-cycle of AMG (with SGS smoothing)*

h	$p = 2$	$p = 3$	$p = 4$
$\frac{1}{8}$	0.9882	0.9882	0.9882
$\frac{1}{16}$	0.9707	0.9707	0.9707
$\frac{1}{32}$	0.9525	0.9525	0.9252

TABLE 9 *CG iteration counts with exact block-diagonal preconditioning,* $h = \frac{1}{16}$, $M = 4$

$\frac{\sigma}{\mu}$	$p = 2$	$p = 3$	$p = 4$
0.1	8	10	11
0.2	11	14	17
0.3	14	21	30
0.4	18	35	532

TABLE 10 *Dimension of global stiffness matrix*

	h	$p = 2$	$p = 3$	$p = 4$
$M = 4$	$\frac{1}{16}$	4,335	10,115	20,230
	$\frac{1}{32}$	16,335	38,115	76,230
	$\frac{1}{64}$	63,375	147,875	295,750
	$\frac{1}{128}$	249,615	582,435	1,116,870
$M = 6$	$\frac{1}{16}$	8,092	24,276	60,690
	$\frac{1}{32}$	30,492	91,476	228,690
	$\frac{1}{64}$	121,968	365,904	887,250
	$\frac{1}{128}$	465,948	1,397,844	3,494,610

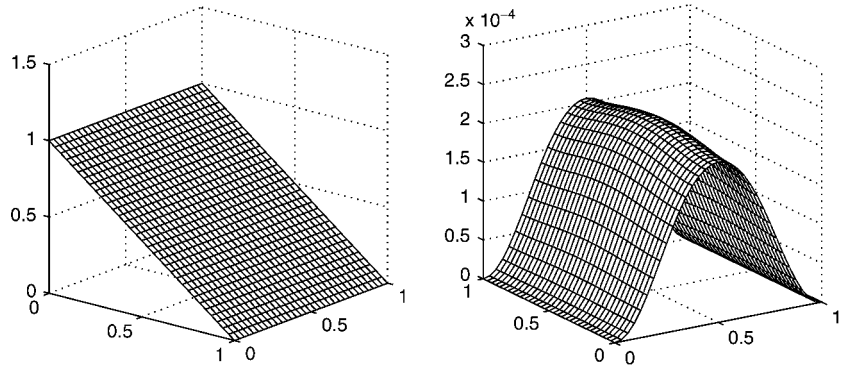

FIG. 6. Mean (left) and variance (right) of pressure on a 32 × 32 mesh for the case $M = 4$ with $p = 2$.

M and h and varying σ are listed in Table 9. Choosing σ to be too large compared to μ causes A to become indefinite and in that case, CG breaks down. This is observed when $p = 4$ and $\sigma\mu^{-1} = 0.4$.

Dimensions of the global systems for the problems considered are summarized in Table 10. Observe then that using our multigrid method, we can solve more than 3.5 million equations on a laptop PC in under 15 min. Furthermore, it should be noted that multigrid algorithms have lower memory requirements than incomplete factorization methods, even with optimized parameters.

4.2 *Mixed boundary conditions*

Next we consider steady flow from left to right on the domain $D = [0,1] \times [0,1]$ with $f = 0$, $\partial D_D = \{0,1\} \times [0,1]$ and $\partial D_N = \partial D \backslash \partial D_D$. We set $\vec{q} \cdot \vec{n} = 0$ at the two horizontal walls so that flow is tangent to those boundaries. The Dirichlet data are $u = 1$ on $\{0\} \times [0,1]$ and $u = 0$ on $\{1\} \times [0,1]$. Again, we employ the covariance function (2.11) with $c_1 = 1 = c_2$, $\sigma = 0.1$ and $\mu = 1$. The mean and variance of the primal variable, obtained on a 32×32 uniform grid using four terms in the KL expansion of $K(\vec{x}, \omega)$ and quadratic Hermite polynomial chaos functions for the stochastic discretization, are plotted in Fig. 6. Preconditioned CG iteration counts and timings are recorded in Table 11. Again we observe that convergence is insensitive to M and h and slightly dependent on p (since we have used

TABLE 11 *Preconditioned CG iterations and timings in seconds (set-up + total iteration times)*

Preconditioner	h	$p = 2$	$p = 33$	$p = 4$
$M = 4$				
None	$\frac{1}{4}$	39	61	118
	$\frac{1}{8}$	73	129	247
	$\frac{1}{16}$	139	246	498
	$\frac{1}{32}$	266	485	984
Block-diagonal	$\frac{1}{16}$	10 (0.00 + 0.45)	11 (0.00 + 1.11)	11 (0.00 + 2.75)
(cholinc, 1×10^{-3})	$\frac{1}{32}$	13 (0.02 + 1.55)	14 (0.18 + 4.71)	15 (0.02 + 10.97)
	$\frac{1}{64}$	23 (0.10 + 11.63)	24 (0.12 + 32.78)	26 (0.11 + 76.11)
	$\frac{1}{128}$	42 (0.66 + 106.59)	45 (0.66 + 284.94)	49 (0.66 + 650.37)
Block-diagonal	$\frac{1}{16}$	10 (0.07 + 0.75)	11 (0.07 + 1.89)	12 (0.07 + 4.70)
(AMG)	$\frac{1}{32}$	10 (0.27 + 1.95)	11 (0.21 + 5.63)	12 (0.21 + 12.25)
	$\frac{1}{64}$	10 (0.91 + 6.36)	11 (1.00 + 18.63)	12 (0.91 + 40.64)
	$\frac{1}{128}$	10 (5.25 + 32.62)	11 (6.89 + 84.65)	12 (6.84 + 193.24)
$M = 6$				
None	$\frac{1}{4}$	40	68	128
	$\frac{1}{8}$	75	138	270
	$\frac{1}{16}$	142	264	533
	$\frac{1}{32}$	273	511	1,029
Block-diagonal	$\frac{1}{16}$	10 (0.00 + 0.88)	11 (0.00 + 4.42)	11 (0.00 + 16.51)
(cholinc, 1×10^{-3})	$\frac{1}{32}$	13 (0.02 + 3.44)	14 (0.18 + 16.69)	15 (0.02 + 53.79)
	$\frac{1}{64}$	22 (0.11 + 26.93)	24 (0.11 + 103.18)	26 (0.11 + 307.86)
	$\frac{1}{128}$	42 (0.66 + 260.22)	45 (0.66 + 877.54)	48 (0.66 + 2,566.47)
Block-diagonal	$\frac{1}{16}$	10 (0.07 + 1.54)	11 (0.07 + 6.40)	12 (0.07 + 23.21)
(AMG)	$\frac{1}{32}$	10 (0.22 + 4.53)	11 (0.22 + 16.91)	12 (0.22 + 52.48)
	$\frac{1}{64}$	10 (1.00 + 15.34)	11 (0.92 + 54.60)	12 (1.02 + 166.29)
	$\frac{1}{128}$	10 (6.90 + 73.30)	11 (7.05 + 270.38)	12 (6.74 + 767.19)

Gaussian random variables). The efficiency of the preconditioning deteriorates for increasing standard deviation, σ.

5. Conclusions

The focus of this work was the design of a fast and robust solver for the model elliptic stochastic boundary-value problem (1.2). Our goals were to provide a theoretical basis for a simple, popular preconditioning scheme employed by other authors and to suggest a practical, efficient implementation based on multigrid. We described the classical spectral SFEM discretization and outlined the structure of the resulting symmetric linear systems. We analysed the exact block-diagonal preconditioner proposed in Ghanem & Kruger (1996), based on the mean component of the system matrix, and established an eigenvalue bound for the preconditioned system in the case that either Gaussian random variables or uniform random variables are employed to represent the diffusion coefficient. Those eigenvalues are independent of h but depend on σ and additionally on p if unbounded random variables are used. In that case, the bounds predict that the system matrix will become indefinite when the stochastic approximation space is enriched. This corresponds to the fact that the underlying variational problem is not well posed. The bound is slightly pessimistic in M, the number of terms retained in the truncated KL expansion of $K(\vec{x}, \omega)$, but the dependence on all other SFEM parameters is sharp. We tested the robustness of the preconditioner with approximate solves for the mean stiffness matrix computed via incomplete Cholesky factorization and using a V-cycle of black-box AMG. The black-box AMG scheme was robust with respect to the spatial discretization parameter without tuning any parameters. It also has lower memory requirements than factorization methods for fine spatial meshes.

Funding

The Nuffield Foundation (NAL/0076/G); the British Council (1279); the US Department of Energy (DE-FG02-04ER25619).

References

Babuška, I. & Chatzipantelidis, P. (2002) On solving elliptic stochastic partial differential equations. *Comput. Methods Appl. Mech. Eng.*, **191**, 4093–4122.

Babuška, I., Nobile, F. & Tempone, R. (2007) A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer Anal.*, **45**, 1005–1034.

Babuška, I., Tempone, R. & Zouraris, G. E. (2004) Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer Anal.*, **42**, 800–825.

Deb, M. K., Babuška, I. & Oden, J. T. (2001) Solution of stochastic partial differential equations using Galerkin finite element techniques. *Comput. Methods Appl. Mech. Eng.*, **90**, 6359–6372.

Dietrich, C. R. & Newsam, G. N. (1997) Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, **18**, 1088–1107.

Eiermann, M., Ernst, O. G. & Ullmann, E. (2007) Computational aspects of the stochastic finite element method. *Comput. Visual Sci.*, **10**, 3–15.

Elman, H., Ernst, O. G., O'Leary, D. P. & Stewart, M. (2005a) Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering. *Comput. Methods Appl. Mech. Eng.*, **18**, 1037–1055.

Elman, H. & Furnival, D. (2007) Solving the stochastic steady-state diffusion problem using multigrid. *IMA J. Numer. Anal.*, **27**, 675–688.

ELMAN, H., SILVESTER, D. & WATHEN, A. (2005b) *Finite Elements and Fast Iterative Solvers*. Oxford: Oxford University Press.

ERNST, O. G. & ULLMANN, E. (2008) On stochastic Galerkin matrices (in preparation).

EWING, R. E. & WHEELER, M. F. (1983) Computational aspects of mixed finite element methods. *Numerical Methods for Scientific Computing* (R. Stepleman ed.). Amsterdam: North-Holland publishing company, pp. 163–172.

FRAUENFELDER, P., SCHWAB, C. & TODOR, R. A. (2005) Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Eng.*, **194**, 205–228.

GHANEM, R. G. (1998) Probabilistic characterization of transport in heterogeneous media. *Comput. Methods Appl. Mech. Eng.*, **158**, 199–220.

GHANEM, R. G. & KRUGER, R. M. (1996) Numerical solution of spectral stochastic finite element systems. *Comput. Methods Appl. Mech. Eng.*, **129**, 289–303.

GHANEM, R. G. & SPANOS, P. D. (2003) *Stochastic Finite Elements: A Spectral Approach*. New York: Dover Publications.

GOLUB, G. H. & WELSCH, J. H. (1969) Calculation of Gauss quadrature rules. *Math. Comput.*, **23**, 221–230.

KEESE, A. (2003) A review of recent developments in the numerical solution of stochastic partial differential equations (stochastic finite elements). *Technical Report 2003-6*. Braunschweig, Germany: Institute of Scientific Computing, Technical University Braunschweig.

KEESE, A. (2004) Numerical solution of systems with stochastic uncertainties: a general purpose framework for stochastic finite elements. *Ph.D. Thesis*, Fachbereich Mathematik und Informatik, TU Braunschweig, Braunschweig, Germany.

KEESE, A. & MATTHIES, H. G. (2002) Efficient solvers for nonlinear stochastic problems. *Proceedings of the Fifth World Congress on Computational Mechanics, Vienna.* http://wccm.tuwien.ac.at/publications/Papers/fp81007.pdf.

KEESE, A. & MATTHIES, H. G. (2003) Pavallel computation of stochastic groundwater flow. *Technical Report 2003–09*. Braunschweig, Germany: Institute of Scientific Computing, Technical University of Braunschweig.

LE MAITRE, O. P., KNIO, O. M., DEBUSSCHERE, B. J., NAJM, H. N. & GHANEM, R. G. (2003) A multigrid solver for two-dimensional stochastic diffusion equations. *Comput. Methods Appl. Mech. Eng.*, **192**, 4723–4744.

LOÈVE, M. (1960) *Probability Theory*. New York: Van Nostrand.

PELLISSETTI, M. F. & GHANEM, R. G. (2000) Iterative solution of systems of linear equations arising in the context of stochastic finite elements. *Adv. Eng. Softw.*, **313**, 607–616.

RUGE, J. W. & STÜBEN, K. (1985) Efficient solution of finite difference and finite element equations by algebraic multigrid (AMG). *Multigrid Methods for Integral and Differential Equations* (D. J. Paddon & H. Holstein eds). The Institute of Mathematics and its Applications Conference Series. New Series 3. Oxford: Clarendon Press, pp. 169–212.

RUSSELL, T. F. & WHEELER, M. F. (1983) Finite element and finite difference methods for continuous flows in porous media. *The Mathematics of Reservoir Simulation* (R. E. Ewing ed.). Philadelphia, PA: SIAM, pp. 35–106.

SILVESTER, D. J. & POWELL, C. E. (2007) PIFISS Potential (Incompressible) Flow & Iterative Solution Software guide. *MIMS technical report 2007.14*. University of Manchester. http://eprints.ma.man.ac.uk/700/.

STOER, J. & BULIRSCH, R. (1980) *Introduction to Numerical Analysis*. New York: Springer.

SUDRET, B. & DER KIUREGHIAN, A. (2000) Stochastic finite element methods and reliability, a state-of-the-art-report. *Report No. UCB/SEMM-2000/08*. Berkeley, CA: Department of Civil and Environmental Engineering, University of California.

WIENER, N. (1938) The homogeneous chaos. *Amer. J. Math.*, **60**, 897–936.

XIU, D. & KARNIADAKIS, G. E. (2003) Modeling uncertainty in steady state diffusion problems via generalised polynomial chaos. *Comput. Methods Appl. Mech. Eng.*, **191**, 4927–4948.