East Asian Journal on Applied Mathematics
doi: 10.4208/eajam

Vol. **xx**, No. x, pp. 1-23
xxx 2018

Efficient spectral stochastic finite element methods for Helmholtz equations with random inputs

Guanjie Wang and Qifeng Liao*

School of Information Science and Technology, ShanghaiTech University, Shanghai, China

Abstract. The implementation of spectral Galerkin stochastic finite element approximation methods for Helmholtz equations with random inputs is addressed in this work. The corresponding linear systems are formulated, of which the coefficient matrices have a Kronecker product structure. The sparsity of the matrices is analyzed and a mean-based preconditioner is developed. Computational results suggest that the mean-based preconditioner is efficient when the underlying stochastic Helmholtz problem is not too close to a resonant frequency.

AMS subject classifications: 65C30, 65F08, 65N30, 35J05

Key words: Helmholtz equations, PDEs with random data, generalized Polynomial Chaos, stochastic finite elements, iterative solvers.

1. Introduction

During the last few decades there has been a rapid development in efficient uncertainty quantification approaches for solving partial differential equations (PDEs) with random inputs. These random inputs typically arise from lack of knowledge or measurement of realistic model parameters, for example, permeability coefficients in diffusion problems [29, 55], viscosity parameters in incompressible flow problems [12, 42, 47, 49], and shape parameters in acoustic scattering problems [58]. In particular, stochastic Helmholtz equations currently gain a lot of interest, and this paper is devoted to investigating efficient solution strategies for them.

The Helmholtz equation is the fundamental governing PDEs for modeling ocean acoustic, optic and electromagnetic problems [28,33,37,51]. When modeling acoustics wave problems, uncertainties typically come from refractive indices (or wave number parameters), source functions, and shapes of scattering surfaces. Elman et al. [14] first consider the Helmholtz equations with random forcing functions and boundary conditions, and develop efficient multigrid solvers for the corresponding stochastic finite element approximation. In the work by Xiu and Shen [58], generalized polynomial chaos (gPC) approximations [55, 56] (and see [30]

*Corresponding author. *Email addresses:* wanggj@shanghaitech.edu.cn (G. Wang), liaoqf@shanghaitech.edu.cn (Q. Liao)

for polynomial chaos) based on stochastic collocation methods [3, 54] are developed for problems with uncertain scattering surface shapes. Tang and Zhou investigate the stochastic collocation method for scalar hyperbolic equations with a random wave speed and demonstrate that the rate of convergence depends on the regularity of the solutions [60]. After that, the studies [39, 40] consider stochastic wave numbers and impedance parameters, and develop multifidelity approaches for the corresponding stochastic optimization problems. The recent work [21, 22] considers stochastic refractive indices, and develops a Monte Carlo interior penalty discontinuous method based on a multimodal representation. Feng *et al.* develop an efficient stochastic Galerkin method for Maxwell's equations with random input [23].

In this work, we investigate spectral Galerkin stochastic finite element methods [2, 30, 52] for the stochastic Helmholtz equations, where uncertainties in the refractive indices are considered. Specifically, we discretize the stochastic parameter space using gPC methods [55, 56] and discretize the physical space using finite element methods [6, 15], which leads to linear systems in Kronecker formulation [10, 41, 43]. We note that efficient iterative solvers for stochastic Galerkin linear systems in general are currently in rapid development, e.g., mean-based preconditioning methods [41, 43], hierarchical preconditioners [47, 48], and preconditioned low-rank projection methods [36], to name a few. However, to the authors' knowledge, the performance of these preconditioned iterative methods has not been studied for the stochastic Helmholtz problems. In this paper we analyze the sparsity of the stochastic Galerkin linear systems associated with stochastic Helmholtz problems, and investigate the corresponding mean-based preconditioning scheme.

An outline of the paper is as follows. In the next section, we first present our problem setting and spectral Galerkin stochastic finite element approximation. After that, the sparsity of the underling linear systems is analyzed and a detailed formulation of the linear systems associated with uniform random inputs is presented. In Section 3, iterative methods and mean-based preconditioning are discussed. Numerical results are discussed in Section 4. Section 5 concludes the paper.

2. The stochastic Helmholtz equation and its discretization

Let $D \subset \mathbb{R}^d$ $(d = 2, 3)$ denote a physical domain which is bounded, connected and with a polygonal boundary ∂D, and $\boldsymbol{x} \in \mathbb{R}^d$ denote a physical variable. Let $\boldsymbol{\xi}$ be a vector which collects a finite number of real-valued random variables. The dimension of $\boldsymbol{\xi}$ is denoted by N, i.e., we write $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^T$. The image of $\boldsymbol{\xi}$ is denoted by Γ and the probability density function of $\boldsymbol{\xi}$ is denoted by $\pi(\boldsymbol{\xi})$. In this paper, we consider the following stochastic Helmholtz problem: find the unknown function $u(\boldsymbol{x}, \boldsymbol{\xi})$ satisfying

$$-\nabla^2 u(\boldsymbol{x}, \boldsymbol{\xi}) - \kappa^2(\boldsymbol{x}, \boldsymbol{\xi}) u(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}) \qquad \forall (\boldsymbol{x}, \boldsymbol{\xi}) \in D \times \Gamma, \tag{2.1}$$

$$u(\boldsymbol{x}, \boldsymbol{\xi}) = 0 \qquad \forall (\boldsymbol{x}, \boldsymbol{\xi}) \in \partial D_D \times \Gamma, \tag{2.2}$$

$$\frac{\partial u}{\partial n} - \mathbf{i}\kappa(\boldsymbol{x}, \boldsymbol{\xi}) u = 0 \qquad \forall (\boldsymbol{x}, \boldsymbol{\xi}) \in \partial D_R \times \Gamma, \tag{2.3}$$

where κ is the refractive index and takes values in $\mathbb{R}$, $\mathbf{i} = \sqrt{-1}$, $\partial u/\partial n$ is the outward normal derivative of u on the boundaries, and the Dirichlet boundaries ∂D_D and the radiation (Sommerfeld) boundaries ∂D_R satisfy $\partial D_D \cup \partial D_R = \partial D$ and $\partial D_D \cap \partial D_R = \emptyset$. The refractive index in (2.1) is assumed to have the following forms:

$$\kappa(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{m=0}^{N} \kappa_m(\boldsymbol{x})\xi_m, \tag{2.4}$$

where $\{\kappa_m(\boldsymbol{x})\}_{m=0}^{N}$ are real-valued deterministic functions, and we set $\xi_0 = 1$ for convenience.

To ensure the well-posedness of our problem, we assume that there exists a constant $\epsilon > 0$, such that $\kappa(\boldsymbol{x}, \boldsymbol{\xi}) > \epsilon$ for all $(\boldsymbol{x}, \boldsymbol{\xi}) \in D \times \Gamma$, and eigenvalues associated of deterministic versions of (2.1) have modulus greater than ϵ. That is, for each realization of $\boldsymbol{\xi}$, considering the following deterministic Helmholtz eigenvalue problem (see [27, 34, 35])

$$-\nabla^2 u(\boldsymbol{x}, \boldsymbol{\xi}) - \kappa^2(\boldsymbol{x}, \boldsymbol{\xi})u(\boldsymbol{x}, \boldsymbol{\xi}) = \lambda(\boldsymbol{\xi})u(\boldsymbol{x}, \boldsymbol{\xi}) \tag{2.5}$$

with boundary conditions (2.2)–(2.3), we collect all its eigenvalues (i.e., all values of $\lambda(\boldsymbol{\xi})$ in (2.5)) into a set denoted by $\Lambda_{\boldsymbol{\xi}}$, and assume that $|\lambda| > \epsilon$ for all $\lambda \in \cup_{\boldsymbol{\xi}\in\Gamma}\Lambda_{\boldsymbol{\xi}}$.

2.1. Variational formulation

To introduce the variational form of (2.1)–(2.3), some notations are required. We first define the space of complex-valued functions that are square integrable,

$$L^2(D) := \left\{ v : D \to \mathbb{C} \;\middle|\; \int_D v\bar{v}\,\mathrm{d}\boldsymbol{x} < \infty \right\}, \tag{2.6}$$

and denote the (function) L^2 norm by

$$\|v\|_2 := \left(\int_D v\bar{v}\mathrm{d}\boldsymbol{x} \right)^{1/2}. \tag{2.7}$$

We next define the space

$$H_0^1(D) := \left\{ v \in H^1(D) \,|\, v = 0 \text{ on } \partial D_D \right\},$$

where $H^1(D)$ is the complex-valued Sobolev space

$$H^1(D) := \left\{ v \in L^2(D),\, \partial v/\partial x_i \in L^2(D), i = 1, \ldots, d \right\}.$$

As usual, for a given function $g(\boldsymbol{\xi}) : \Gamma \to \mathbb{C}$, its expectation (mean value) is defined as

$$\mathbb{E}\big[g(\boldsymbol{\xi})\big] := \int_\Gamma \pi(\boldsymbol{\xi})g(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi},$$

where $\pi(\xi)$ is the probability density function of ξ.

The solution and test function space can then be defined as

$$\begin{aligned} W &:= H_0^1(D) \otimes L_\pi^2(\Gamma) \\ &:= \left\{ w(\boldsymbol{x},\xi) : D \times \Gamma \to \mathbb{C} \mid \|w(\boldsymbol{x},\xi)\|_W < \infty \text{ and } w|_{\partial D_D \times \Gamma} = 0 \right\}, \end{aligned}$$

where $L_\pi^2(\Gamma) := \{ g : \Gamma \to \mathbb{C} \mid \mathbb{E}[g\bar{g}] < \infty \}$ and the norm $\|\cdot\|_W$ is defined by $\|w(\boldsymbol{x},\xi)\|_W^2 := \int_\Gamma \pi(\xi) \int_D |\nabla w|^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\xi$.

Following [22, 43, 53], the variational form of (2.1)–(2.3) can be written as: find $u \in W$, such that

$$\mathbb{E}\left[\int_D \nabla u \cdot \nabla \bar{w} - \int_D \kappa^2 u \bar{w} - \mathbf{i} \int_{\partial D_R} \kappa u \bar{w} \right] = \mathbb{E}\left[\int_D f \bar{w} \right], \ \forall w \in W. \tag{2.8}$$

2.2. Discretization

A discrete version of (2.8) is obtained by introducing a finite-dimensional subspace W^h to approximate W. Specifically, we first denote finite-dimensional subspaces of the corresponding stochastic and physical spaces by

$$S = \operatorname{span}\left\{ \Phi_j(\xi) \right\}_{j=1}^{N_\xi} \subset L_\pi^2(\Gamma), \ V^h = \operatorname{span}\left\{ v_s(\boldsymbol{x}) \right\}_{s=1}^{N_x} \subset H_0^1(D), \tag{2.9}$$

where $\Phi_j(\xi)$ and $v_s(\boldsymbol{x})$ refer to basis functions. We next define a finite-dimensional subspace of the overall solution (and test) function space W by

$$W^h := V^h \otimes S := \operatorname{span}\left\{ v(\boldsymbol{x})\Phi(\xi) \,\middle|\, v \in V^h, \Phi \in S \right\}.$$

There are many choices for the bases in (2.9) corresponding to different discretization methods, e.g., piecewise linear functions [2, 6, 9, 15] and global orthogonal polynomials [5, 30, 45, 53]. The global orthogonal polynomial approximations for the stochastic space consist of three main kinds: the polynomial chaos methods [29, 30], the generalized polynomial chaos methods [56], and the dynamically bi-orthogonal methods [7, 8, 38, 61]. In this paper, we focus on generalized polynomial chaos methods to discrete the stochastic parameter space and finite element methods for the physical space. We review the generalized polynomial chaos methods introduced by [55, 57] in the following for completeness.

As introduced in [55], a gPC approximation of the solution $u(\boldsymbol{x},\xi)$ is written as

$$u(\boldsymbol{x},\xi) \approx u^p(\boldsymbol{x},\xi) := \sum_{j=1}^{N_\xi} u_j(\boldsymbol{x}) \Phi_j(\xi), \tag{2.10}$$

where $S = \{\Phi_j(\xi)\}_{j=1}^{N_\xi}$ is an orthogonal basis with respect to the inner product

$$\mathbb{E}\left[\Phi_j(\xi)\Phi_k(\xi) \right] = \int_\Gamma \pi(\xi) \Phi_j(\xi) \Phi_k(\xi) \, \mathrm{d}\xi. \tag{2.11}$$

For an one-dimensional random input with probability density function $\pi(\xi)$, the basis functions in (2.10) are $\Phi_j(\xi) = \phi_{j-1}(\xi),\ j = 1,\dots,N_\xi$, where $\{\phi_j\}_{j=0}^{N_\xi-1}$ is a sequence of orthogonal polynomials with respect to the inner product

$$\mathbb{E}\big[\phi_j(\xi)\phi_k(\xi)\big] = \int_\Gamma \pi(\xi)\phi_j(\xi)\phi_k(\xi)\,\mathrm{d}\xi\,.$$

For more details about the orthogonal polynomials, see [1,45].

For multi-dimensional random inputs ($N > 1$), supposing $\xi_1,\dots,\xi_N$ are independent random variables, each stochastic basis function $\Phi_j(\boldsymbol{\xi})$ for $j \in \{1,\dots,N_\xi\}$ is a product of N univariate orthogonal polynomials. More precisely, $\Phi_j(\boldsymbol{\xi}) = \phi_{j_1}^{(1)}(\xi_1)\cdots\phi_{j_N}^{(N)}(\xi_N)$, where $\{\phi_k^{(i)}(\xi_i)\}_{k=0}^{\infty}$ is the univariate orthogonal basis corresponding to ξ_i's probability density function $\pi_i(\xi_i)$ for $i = 1,\dots,N$. Each single-index $j \in \{1,\dots,N_\xi\}$ here can be represented by a multi-index $\boldsymbol{j} = (j_1,\dots,j_N)$, and $|\boldsymbol{j}| = j_1+\cdots+j_N$ specifies the total degree of $\Phi_j(\boldsymbol{\xi})$. To exactly define each $\Phi_j(\boldsymbol{\xi})$, a bijection from single index to multi-index are introduced, i.e.,

$$\mathscr{M}_b : j \longleftrightarrow (j_1,\dots,j_N). \tag{2.12}$$

Once the bijection $\mathscr{M}_b$ is specified, $\Phi_j(\boldsymbol{\xi})$ is then determined. A popular choice is arranging the multi-index $\boldsymbol{j}$ in graded lexicographic order [53]. That is, if $|\boldsymbol{i}| > |\boldsymbol{j}|$, or $|\boldsymbol{i}| = |\boldsymbol{j}|$ and the first nonzero entry in the difference $\boldsymbol{i}-\boldsymbol{j}$ is positive, we set $\mathscr{M}_b^{-1}(\boldsymbol{i}) > \mathscr{M}_b^{-1}(\boldsymbol{j})$.

Following [59], for a given integer (the gPC order) $p > 0$, the gPC approximation (2.10) can be rewritten in the following multi-index form

$$u^p(\boldsymbol{x},\boldsymbol{\xi}) = \sum_{j=1}^{N_\xi} u_j(\boldsymbol{x})\Phi_j(\boldsymbol{\xi}) = \sum_{|\boldsymbol{j}|=0}^{p} u_{\boldsymbol{j}}(\boldsymbol{x})\Phi_{\boldsymbol{j}}(\boldsymbol{\xi}),\ |\boldsymbol{j}| = j_1+\cdots+j_N, \tag{2.13}$$

where $N_\xi = \binom{N+p}{p}$ (see [55]).

For the spatial point of view, each $u_j(\boldsymbol{x})$ can be approximated by

$$u_j(\boldsymbol{x}) \approx \sum_{s=1}^{N_x} u_{js}v_s(\boldsymbol{x}),\ v_s(\boldsymbol{x}) \in V^h, \tag{2.14}$$

where $\{v_s(\boldsymbol{x})\}_{s=1}^{N_x}$ refers to a finite element basis of V^h. Combining (2.10) and (2.14), the overall approximation of the solution $u(\boldsymbol{x},\boldsymbol{\xi})$ is written as

$$u^{ph}(\boldsymbol{x},\boldsymbol{\xi}) := \sum_{j=1}^{N_\xi}\sum_{s=1}^{N_x} u_{js}v_s(\boldsymbol{x})\Phi_j(\boldsymbol{\xi})\,. \tag{2.15}$$

The unknown coefficients u_{js} for $j = 1,\dots,N_\xi$ and $s = 1,\dots,N_x$ in (2.15), can be obtained through solving the following linear system (with size $N_xN_\xi \times N_xN_\xi$)

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}\,, \tag{2.16}$$

where

$$\boldsymbol{A} = \boldsymbol{G}_{00} \otimes \boldsymbol{K} - \sum_{l=0}^{N}\sum_{m=0}^{N} \boldsymbol{G}_{lm} \otimes \boldsymbol{M}_{lm} - \sum_{l=0}^{N} \mathbf{i}\boldsymbol{G}_{l0} \otimes \boldsymbol{L}_{l}; \tag{2.17}$$

$$\boldsymbol{b} = \boldsymbol{h} \otimes \boldsymbol{f}. \tag{2.18}$$

In (2.17)–(2.18), $\otimes$ denotes Kronecker tensor product and

$$\boldsymbol{h}(j) = \mathbb{E}\big[\Phi_j(\xi)\big], \qquad \boldsymbol{f}(s) = \int_D f v_s \,\mathrm{d}\boldsymbol{x}, \tag{2.19}$$

$$\boldsymbol{M}_{lm}(s,t) = \int_D \kappa_l \kappa_m v_s v_t \,\mathrm{d}\boldsymbol{x}, \qquad \boldsymbol{L}_l(s,t) = \int_{\partial D_R} \kappa_l v_s v_t \,\mathrm{d}s, \tag{2.20}$$

$$\boldsymbol{G}_{lm}(j,k) = \mathbb{E}\big[\xi_l \xi_m \Phi_j(\xi)\Phi_k(\xi)\big], \qquad \boldsymbol{K}(s,t) = \int_D \nabla v_s \cdot \nabla v_t \,\mathrm{d}\boldsymbol{x}, \tag{2.21}$$

where $l,m = 0,1,\ldots,N$; $j,k = 1,\ldots,N_\xi$ and $s,t = 1,\ldots,N_x$.

Based on the above discussion and following [43], (2.16) can be rewritten as the following block form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \ldots & \boldsymbol{A}_{1N_\xi} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \ldots & \boldsymbol{A}_{2N_\xi} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{N_\xi 1} & \boldsymbol{A}_{N_\xi 2} & \ldots & \boldsymbol{A}_{N_\xi N_\xi} \end{bmatrix}, \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_{N_\xi} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_{N_\xi} \end{bmatrix}, \tag{2.22}$$

where each $\boldsymbol{A}_{jk}$ for $j,k = 1,\ldots,N_\xi$ is a $N_x \times N_x$ matrix.

Once the approximation $u^{ph}(\boldsymbol{x},\xi)$ (see (2.15)) is obtained through solving (2.16), the mean function of the solution can be approximated by

$$\mathbb{E}\big[u(\boldsymbol{x},\xi)\big] \approx \mathbb{E}\big[u^{ph}(\boldsymbol{x},\xi)\big], \tag{2.23}$$

and the variance function of the solution can be approximated by

$$\mathrm{Var}(u(\boldsymbol{x},\xi)) \approx \mathrm{Var}\big[u^{ph}(\boldsymbol{x},\xi)\big] := \mathbb{E}\Big[\big|u^{ph}(\boldsymbol{x},\xi) - \mathbb{E}\big[u^{ph}(\boldsymbol{x},\xi)\big]\big|^2\Big]. \tag{2.24}$$

2.3. Sparsity of the coefficient matrix

As presented in (2.17) and (2.22), the coefficient matrix $\boldsymbol{A}$ can be written as a block matrix, while each block has the same sparsity pattern as the corresponding deterministic problem [43]. We note that the general sparsity and structural properties of the coefficient matrix are studied

in [19]. In the setting of the stochastic Helmholtz problem considered in this paper, for even weight functions, the number of nonzero entries of $\boldsymbol{G}_{lm}$ for $l = 1,\dots,N$ and $m = 0$ are given in [19]. We in the following investigate more general cases—the sparsity of $\boldsymbol{G}_{lm}$ for $l,m = 0,1,\dots,N$ and the overall sparsity of the coefficient matrix $\boldsymbol{A}$.

To assess the number of the nonzero blocks of $\boldsymbol{A}$ (see (2.22)), the following matrix is defined

$$\hat{\boldsymbol{G}}(j,k) = \begin{cases} 1, \text{ if there exist } l,m \in \{0,1,\dots N\} \text{ such that } \boldsymbol{G}_{lm}(j,k) \neq 0 \\ 0, \text{ otherwise} \end{cases},$$

where $j,k = 1,\dots,N_\xi$. It is clear that the number of nonzero blocks of $\boldsymbol{A}$ is less or equal to the number of nonzero entries of $\hat{\boldsymbol{G}}$ above, and we next count nonzero entries of $\hat{\boldsymbol{G}}$. Without loss of generality, we hereafter suppose the univariate basis functions are orthonormal, i.e.,

$$\int \pi_i(\xi_i)\phi_j(\xi_i)\phi_k(\xi_i) = \delta_{jk},\ i = 1,\dots,N,$$

where δ_{jk} is the Kronecker delta function.

For the case $l = m = 0$, $\boldsymbol{G}_{00}(j,k) = \prod_{i=1}^N \delta_{j_i k_i}$ for $j,k = 1,\dots N_\xi$, where $(j_1,\dots,j_N)$ and $(k_1,\dots,k_N)$ are the multi-indices corresponding to j and k. The nonzero terms are those with indices j,k satisfying $|j_i - k_i| = 0$ for $i \in \{1,\dots N\}$.

When either l or m is zero, we consider the case $\boldsymbol{G}_{0l}(j,k) = \boldsymbol{G}_{l0}(j,k) = \mathbb{E}\Big[\xi_l \phi_{j_l}^{(l)}(\xi_l)\phi_{k_l}^{(l)}(\xi_l)\Big]\prod_{i=1,i\neq l}^N \delta_{j_i k_i}$, with $l > 0$. Since $\phi_{j_l}^{(l)}(\xi_l)$ is orthogonal to the polynomials of degree less than j_l, $\boldsymbol{G}_{0l}(j,k) \neq 0$ holds, i.e. $\hat{\boldsymbol{G}}(j,k) = 1$, only if the indices j,k satisfying $|j_l - k_l| \leq 1$ and $j_i = k_i$ for $i \in \{1,\dots,N\} \setminus \{l\}$, where $(j_1,\dots,j_N)$ and $(k_1,\dots,k_N)$ are the multi-indices corresponding to j and k.

When $l = m > 0$, $\boldsymbol{G}_{ll}(j,k) = \mathbb{E}\Big[\xi_l^2 \phi_{j_l}^{(l)}(\xi_l)\phi_{k_l}^{(l)}(\xi_l)\Big]\prod_{i=1,k\neq\{l\}}^N \delta_{j_i k_i}$. In this case, $\boldsymbol{G}_{ll}(j,k) \neq 0$ holds, i.e. $\hat{\boldsymbol{G}}(j,k) = 1$, only if $|j_l - k_l| \leq 2$ and $j_i = k_i$ for $i \in \{1,\dots,N\} \setminus \{l\}$.

When $l \neq m$ and $lm \neq 0$, $\boldsymbol{G}_{lm}(j,k) = \mathbb{E}\Big[\xi_l \phi_{j_l}^{(l)}(\xi_l)\phi_{k_l}^{(l)}(\xi_l)\Big]\mathbb{E}\Big[\xi_m \phi_{j_m}^{(m)}(\xi_m)\phi_{k_m}^{(m)}(\xi_m)\Big]\prod_{i=1,i\neq\{l,m\}}^N \delta_{j_i k_i}$. In this case, $\boldsymbol{G}_{lm}(j,k) \neq 0$ holds, i.e. $\hat{\boldsymbol{G}}(j,k) = 1$, only if $|j_l - k_l| \leq 1$ and $|j_m - k_m| \leq 1$ and $j_i = k_i$ for $i \in \{1,\dots,N\} \setminus \{l,m\}$.

In summary, $\hat{\boldsymbol{G}}(j,k) \neq 0$ holds if and only if one of the following three statements holds true

(a) $j_i = k_i$ for $i \in \{1,\dots N\}$;

(b) for each $l \in \{1,\dots,N\}$, $|j_l - k_l| = 1,2$, and $j_i = k_i$ for $i \in \{1,\dots,N\} \setminus \{l\}$;

(c) for each pair $l,m \in \{1,\dots,N\}$ with $l \neq m$, $|j_l - k_l| = 1$, $|j_m - k_m| = 1$ and $j_i = k_i$ for $i \in \{1,\dots,N\} \setminus \{l,m\}$.

The number of indices that satisfy case (a) equals to the number of solutions of the following problem: find non-negative integers $j_1,\dots,j_N$, such that

$$j_1 + \cdots + j_N \leq p.$$

The number of solutions for the above equation can be computed by the stars and bars method [20], and equals $\binom{N+p}{p}$.

For case (b), the situation that $|j_l - k_l| = 1$ and $j_i = k_i$ ($i \in \{1,\dots,N\} \setminus \{l\}$) is studied in [19]. To simplify the analysis, we use a different counting method. In the following, we take $j_l = k_l + 1$ and $j_i = k_i$ ($i \in \{1,\dots,N\} \setminus \{l\}$) as an example to demonstrate the method. Since the total degree of each gPC basis function is equal to or smaller than p, the multi-index of j satisfies

$$j_1 + \cdots + j_l + \cdots + j_N \le p,$$

or equivalently

$$j_1 + \cdots + (k_l + 1) + \cdots + j_N \le p,$$

where $j_1,\dots,k_l,\dots,j_N$ are non-negative integers. Thus the number of index pairs j,k satisfying $j_l = k_l + 1$ and $j_i = k_i$ ($i \in \{1,\dots,N\} \setminus \{l\}$) equals to the number of solutions of the following problem: find non-negative integers $j_1,\dots,k_l,\dots,j_N$, such that

$$j_1 + \cdots + k_l + \cdots + j_N \le p - 1.$$

By the stars and bars method, this number is $\binom{N+p-1}{p-1}$. As discussed above, the total number of indices that satisfy (b) is

$$2N\binom{N+p-1}{p-1} + 2N\binom{N+p-2}{p-2},$$

For case (c), the counting method is similar and we take $j_l = k_l + 1$, $j_m = k_m + 1$ and $j_i = k_i$ ($i \in \{1,\dots,N\} \setminus \{l,m\}$) as an example. Since the total degree of each gPC basis function is equal to or smaller than p, the multi-index of j satisfies

$$j_1 + \cdots + j_l + \cdots + j_N \le p,$$

or equivalently

$$j_1 + \cdots + (k_l + 1) + \cdots + (k_m + 1) + j_N \le p,$$

where $j_1,\dots,k_l,\dots,k_m,\dots,j_N$ are non-negative integers. Thus the number of index-pairs j,k satisfying $j_l = k_l + 1$, $j_m = k_m + 1$ and $j_i = k_i$ ($i \in \{1,\dots,N\} \setminus \{l,m\}$) equals to the number of solutions of the following problem: find non-negative integers $j_1,\dots,k_l,\dots,k_m,\dots,j_N$, such that

$$j_1 + \cdots + k_l + \cdots + k_m + \cdots + j_N \le p - 2.$$

By the stars and bars method, this number is $\binom{N+p-2}{p-2}$. As discussed above, the total number of indices that satisfy (c) is[†]

$$(N^2 - N)\binom{N+p-2}{p-2} + (N^2 - N)\binom{N+p-1}{p-1}$$

[†]When $p = 1$, we define $\binom{N+p-2}{p-2} = 0$.

Thus the total number of nonzero entries in the matrices $\hat{\boldsymbol{G}}$, i.e. the number of nonzero blocks in coefficient matrix (see (2.22)), is at most

$$\begin{aligned}
&(N^2+N)\left[\binom{N-1+p}{p-1}+\binom{N-1+p-1}{p-2}\right]+\binom{N+p}{p}\\
&=\left[(N^2+N)\frac{N+2p-2}{N+p-1}\frac{p}{N+p}+1\right]\binom{N+p}{p} \qquad (2.25)\\
&\triangleq C_\xi\binom{N+p}{p}=C_\xi N_\xi,
\end{aligned}$$

where

$$C_\xi=\left[(N^2+N)\frac{N+2p-2}{N+p-1}\frac{p}{N+p}+1\right]<2(N^2+N)+1.$$

It is clear that, C_ξ is typically much smaller than $N_\xi=\binom{N+p}{p}$ when N and p are not too small. Values of the ratio C_ξ/N_ξ are shown in Figure 1, where it can be seen that the ratio values decrease quickly as N and p increase.

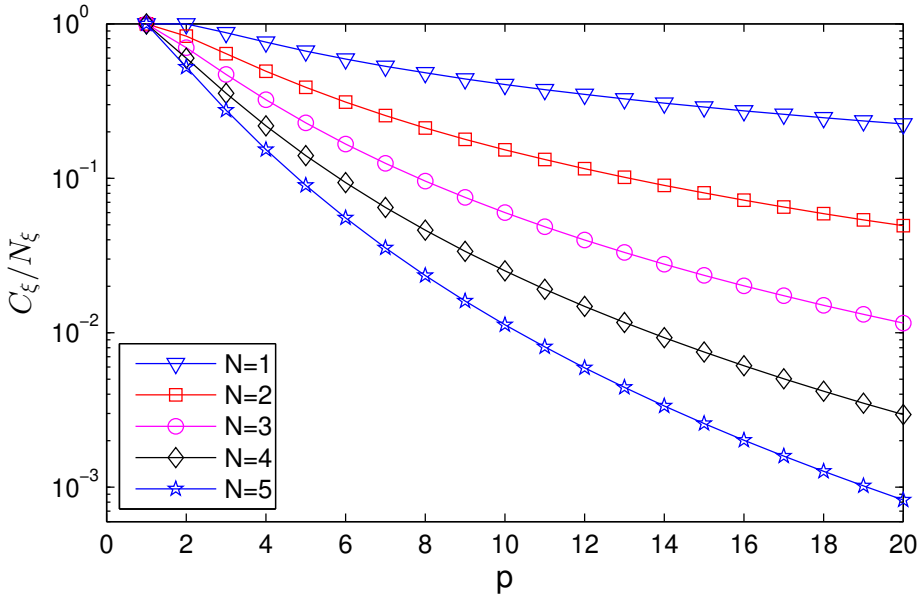


Figure 1: The sparsity of blocks.

Next, as discussed in [43], each nonzero block of $\boldsymbol{A}$ in (2.22) has the same sparsity pattern as the corresponding deterministic problem. When using standard finite element methods to discretize the physical space D, the number of nonzero entries of each block can typically be written as C_xN_x with $C_x \ll N_x$ and C_x is independent of finite element degrees of freedom, for

example, $C_x = 9$ for bilinear rectangular finite elements [15]. Thus, the number of nonzero entries of $\boldsymbol{A}$ is at most $C_x C_\xi N_x N_\xi$ and can be written as $O(N_x N_\xi)$ (since $C_x \ll N_x$ and $C_\xi \ll N_\xi$ as discussed above). Considering its size ($N_x N_\xi \times N_x N_\xi$), the matrix $\boldsymbol{A}$ is sparse, and it is of interest to develop iterative linear solvers to solve (2.16) with a cost $O(N_x N_\xi)$ [15,44], and we will discuss this again in Section 3.

2.4. Detailed discrete formulation for uniform inputs

For any distribution, once the coefficients of the three-term recurrence relation (**??**), i.e. α_j, β_j, is known, the matrices $\boldsymbol{G}_{lm}$ and vectors $\boldsymbol{h}$ can be calculated analytically. Specifically, in this section, we give the formulation for independent identically distributed uniform random inputs.

For a uniform distribution ξ in $[-1,1]$, the probability density function is $\pi(\xi) = 1/2$. It is well known that the Legendre polynomials form an orthogonal basis in $[-1,1]$ with respect to the probability density function $\pi(\xi) = 1/2$. Normalizing the Legendre polynomials, we obtain the three-term recurrence relation for the orthonormal polynomial bases, i.e,

$$\phi_{i+1}(\xi) = \frac{\sqrt{(2i+1)(2i+3)}}{i+1}\xi\phi_i(\xi) - \frac{i\sqrt{2i+3}}{(i+1)\sqrt{2i-1}}\phi_{i-1}(\xi),$$

where $\phi_0(\xi) = 1$ and $\phi_1(\xi) = \sqrt{3}\xi$.

By the definition of $\boldsymbol{h}$ and $\boldsymbol{G}_{lm}$, we have

$$\boldsymbol{h}(j) = \mathbb{E}\big[\Phi_j(\boldsymbol{\xi})\big] = \left(\prod_{i=1}^{N}\mathbb{E}\big[\phi_{j_i}(\xi_i)\big]\right) = \begin{cases} 1, \text{ if } j_i = 0, \\ 0, \text{ otherwise;} \end{cases}$$

if $l = m = 0$,

$$\boldsymbol{G}_{00} = \boldsymbol{I};$$

if either l or m equals zero (i.e., $lm = 0$ and $l + m > 0$),

$$\begin{aligned} \boldsymbol{G}_{0l}(j,k) = \boldsymbol{G}_{l0}(j,k) &= \mathbb{E}[\xi_l \Phi_j(\boldsymbol{\xi})\Phi_k(\boldsymbol{\xi})] \\ &= \left(\prod_{i=1, i\neq l}^{N}\mathbb{E}\big[\phi_{j_i}(\xi_i)\phi_{k_i}(\xi_i)\big]\right)\mathbb{E}\big[\xi_l\phi_{j_l}(\xi_l)\phi_{k_l}(\xi_l)\big] \\ &= \begin{cases} \dfrac{j_l}{\sqrt{4j_l^2-1}}\displaystyle\prod_{i=1,i\neq l}^{N}\delta_{j_i k_i}, & \text{if } k_l = j_l - 1, \\ \dfrac{k_l}{\sqrt{4k_l^2-1}}\displaystyle\prod_{i=1,i\neq l}^{N}\delta_{j_i k_i}, & \text{if } j_l = k_l - 1, \\ 0, & \text{otherwise;} \end{cases} \end{aligned}$$

if $l = m > 0$,

$$\begin{aligned}
\boldsymbol{G}_{ll}(j,k) &= \mathbb{E}[\xi_l^2 \Phi_j(\boldsymbol{\xi})\Phi_k(\boldsymbol{\xi})] \\
&= \left(\prod_{i=1,i\neq l}^{N} \mathbb{E}\left[\phi_{j_i}(\xi_i)\phi_{k_i}(\xi_i)\right]\right) \mathbb{E}\left[\xi_l^2 \phi_{j_l}(\xi_l)\phi_{k_l}(\xi_l)\right] \\
&= \begin{cases}
\left(\dfrac{(j_l+1)^2}{(2j_l+1)(2j_l+3)} + \dfrac{j_l^2}{4j_l^2-1}\right) \displaystyle\prod_{i=1,i\neq l}^{N} \delta_{j_i k_i}, & \text{if } j_l = k_l, \\
\left(\dfrac{1}{\sqrt{(2j_l+1)(2j_l-3)}} \dfrac{j_l(j_l-1)}{2j_l-1}\right) \displaystyle\prod_{i=1,i\neq l}^{N} \delta_{j_i k_i}, & \text{if } k_l = j_l - 2, \\
\left(\dfrac{1}{\sqrt{(2k_l+1)(2k_l-3)}} \dfrac{k_l(k_l-1)}{2k_l-1}\right) \displaystyle\prod_{i=1,i\neq l}^{N} \delta_{j_i k_i}, & \text{if } j_l = k_l - 2, \\
0, & \text{otherwise;}
\end{cases}
\end{aligned}$$

if $l \neq m$ and $lm \neq 0$,

$$\begin{aligned}
\boldsymbol{G}_{lm}(j,k) &= \boldsymbol{G}_{ml}(j,k) = \mathbb{E}[\xi_l \xi_m \Phi_j(\boldsymbol{\xi})\Phi_k(\boldsymbol{\xi})] \\
&= \left(\prod_{i=1,i\neq\{l,m\}}^{N} \mathbb{E}\left[\phi_{j_i}(\xi_i)\phi_{k_i}(\xi_i)\right]\right) \\
&\quad \times \mathbb{E}\left[\xi_l \xi_m \phi_{j_l}(\xi_l)\phi_{j_m}(\xi_m)\phi_{k_l}(\xi_l)\phi_{k_m}(\xi_m)\right], \\
&= \begin{cases}
\left(\dfrac{j_l}{\sqrt{4j_l^2-1}} \dfrac{j_m}{\sqrt{4j_m^2-1}}\right) \displaystyle\prod_{i=1,i\neq\{l,m\}}^{N} \delta_{j_i k_i}, \text{ if } \begin{cases} k_l = j_l - 1 \\ k_m = j_m - 1 \end{cases}, \\
\left(\dfrac{j_l}{\sqrt{4j_l^2-1}} \dfrac{k_m}{\sqrt{4k_m^2-1}}\right) \displaystyle\prod_{i=1,i\neq\{l,m\}}^{N} \delta_{j_i k_i}, \text{ if } \begin{cases} k_l = j_l - 1 \\ j_m = k_m - 1 \end{cases}, \\
\left(\dfrac{k_l}{\sqrt{4k_l^2-1}} \dfrac{j_m}{\sqrt{4j_m^2-1}}\right) \displaystyle\prod_{i=1,i\neq\{l,m\}}^{N} \delta_{j_i k_i}, \text{ if } \begin{cases} j_l = k_l - 1 \\ k_m = j_m - 1 \end{cases}, \\
\left(\dfrac{k_l}{\sqrt{4k_l^2-1}} \dfrac{k_m}{\sqrt{4k_m^2-1}}\right) \displaystyle\prod_{i=1,i\neq\{l,m\}}^{N} \delta_{j_i k_i}, \text{ if } \begin{cases} j_l = k_l - 1 \\ j_m = k_m - 1 \end{cases}, \\
0, \text{ otherwise.}
\end{cases}
\end{aligned}$$

Finally, combining the above results with (2.12) and (2.17)–(2.18), the linear system (2.16) can be formed for uniform random inputs.

3. Iterative methods for the linear system

As discussed in Section 2.2, discretization of the Helmholtz problem (2.1)–(2.3) results in the sparse linear system $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}$ (see (2.16)), where $\boldsymbol{A}$ and $\boldsymbol{b}$ are defined through (2.17) and (2.18). When high solution accuracy is required, the size of the matrix $\boldsymbol{A}$ can be large. In this section, we discuss efficient iterative methods to solve this kind of large sparse linear systems. In particular, we focus on Krylov subspace methods [15,17,31], of which the main methodology is to project the linear system (2.16) into a consecutively constructed Krylov subspace defined as

$$\mathscr{K}_m(\boldsymbol{A}, \boldsymbol{r}^{(0)}) = \text{span}\{\boldsymbol{r}^{(0)}, \boldsymbol{A}\boldsymbol{r}^{(0)}, \boldsymbol{A}^2\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{A}^{m-1}\boldsymbol{r}^{(0)}\},$$

where

$$\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}^{(0)} \tag{3.1}$$

with $\boldsymbol{u}^{(0)}$ a given initial guess.

When the system matrix is symmetric and positive definite, the conjugate gradient (CG) method (a kind of Krylov subspace method) [32] is a popular choice, which in general only uses three vectors in memory and minimizes the error in the $\boldsymbol{A}$-norm. However, the linear system considered in this paper (see (2.16)) is complex-symmetric (not Hermitian), such that directly using CG method may not lead to a convergent algorithm. As discussed in [44], methods that are based on the Lanczos bi-orthogonalization procedure can be used for non-symmetric and nonsingular matrices, such as the bi-conjugate gradient (Bi-CG) method [24] and its variants [46,50]. Since the Bi-CG methods may not be stable for non-Hermitian linear systems [16,25], Freund and Nachtigal proposed the quasi-minimal residual (QMR) method [25] which is more robust for these linear systems. In this paper, we focus on the QMR method [25, 26], and a bi-conjugate gradient stabilized (Bi-CGSTAB) method (a variant of Bi-CG method [50]) for comparison. Our implementation is based on the MATLAB functions `qmr` and `bicgstab` for the QMR method and the Bi-CGSTAB method respectively.

To reduce the number of iterations, preconditioners are typically required when using iterative methods. As the preconditioning framework discussed in [15, 44], instead of solving the original problem (2.16), we solve the following problem

$$\boldsymbol{P}_1^{-1}\boldsymbol{A}\boldsymbol{P}_2^{-1}\tilde{\boldsymbol{u}} = \boldsymbol{P}_1^{-1}\boldsymbol{b}, \quad \text{where } \tilde{\boldsymbol{u}} = \boldsymbol{P}_2\boldsymbol{u}. \tag{3.2}$$

The nonsingular matrices $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are called preconditioners, and the linear systems $\boldsymbol{P}_i\boldsymbol{x} = \boldsymbol{b}$ for $i = 1, 2$ are expected to be inexpensive to solve. An efficient preconditioner corresponds to well-clustered eigenvalues that are not too close to the origin [18].

Following the framework introduced in [29, 41, 43], we in the following construct a mean-based preconditioner for the stochastic Helmholtz equation. To start with, we denote the mean value of ξ by $\xi^{(0)}$. We next construct the following matrix

$$\boldsymbol{P} = \boldsymbol{G}_{00} \otimes (\boldsymbol{K} + \boldsymbol{M}_p + \mathbf{i}\boldsymbol{L}_p) \tag{3.3}$$

where $\boldsymbol{G}_{00}$, $\boldsymbol{K}$ are defined in (2.21) and for $s,t=1,\ldots,N_x$,

$$\boldsymbol{M}_p(s,t)=\int_D \kappa^2\left(\boldsymbol{\xi}^{(0)}\right)v_s v_t\,\mathrm{d}\boldsymbol{x},\ \boldsymbol{L}_p(s,t)=\int_{\partial D_R}\kappa\left(\boldsymbol{\xi}^{(0)}\right)v_s v_t\,\mathrm{d}s. \tag{3.4}$$

The mean-based preconditioners herein are defined through setting $\boldsymbol{P}_1=\boldsymbol{P}$ and $\boldsymbol{P}_2=\boldsymbol{I}$ in (3.2). In addition, to start the iterative solving procedure, we set an initial guess $\boldsymbol{u}^{(0)}=\boldsymbol{P}^{-1}\boldsymbol{b}$ in (3.1).

At each iteration step, the following equation needs to be solved

$$\boldsymbol{P}\hat{\boldsymbol{x}}=\hat{\boldsymbol{y}}, \tag{3.5}$$

where

$$\hat{\boldsymbol{x}}=\begin{bmatrix}\hat{\boldsymbol{x}}_1\\ \hat{\boldsymbol{x}}_2\\ \vdots\\ \hat{\boldsymbol{x}}_{N_\xi}\end{bmatrix},\quad \hat{\boldsymbol{y}}=\begin{bmatrix}\hat{\boldsymbol{y}}_1\\ \hat{\boldsymbol{y}}_2\\ \vdots\\ \hat{\boldsymbol{y}}_{N_\xi}\end{bmatrix},\ \text{where } \hat{\boldsymbol{x}}_i,\hat{\boldsymbol{y}}_i\in\mathbb{C}^{N_x}.$$

Since $\boldsymbol{G}_{00}=\boldsymbol{I}$ is symmetric, solving (3.5) is equivalent to solving the following problem

$$(\boldsymbol{K}+\boldsymbol{M}_p+\mathbf{i}\boldsymbol{L}_p)\hat{\boldsymbol{X}}\boldsymbol{G}_{00}=\hat{\boldsymbol{Y}}, \tag{3.6}$$

where

$$\hat{\boldsymbol{X}}=\left[\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_{N_\xi}\right]\text{ and }\hat{\boldsymbol{Y}}=\left[\hat{\boldsymbol{y}}_1,\ldots,\hat{\boldsymbol{y}}_{N_\xi}\right].$$

To compute the solution of (3.6), we only need to solve N_ξ linear systems with size $N_x\times N_x$, which is much cheaper than directly solving (2.16) (whose size is $N_xN_\xi\times N_xN_\xi$).

4. Numerical results

In this section, two test problems are considered. The first one considers the refractive index to be a random field, and the second one considers a problem close to a resonant frequency. In both test problems, we discretize in physical space using a bilinear finite element approximation [6, 15], and our implementation is based on the IFISS [13] and the S-IFISS [4] packages.

4.1. Test problem 1 (the refractive index modeled by a random field)

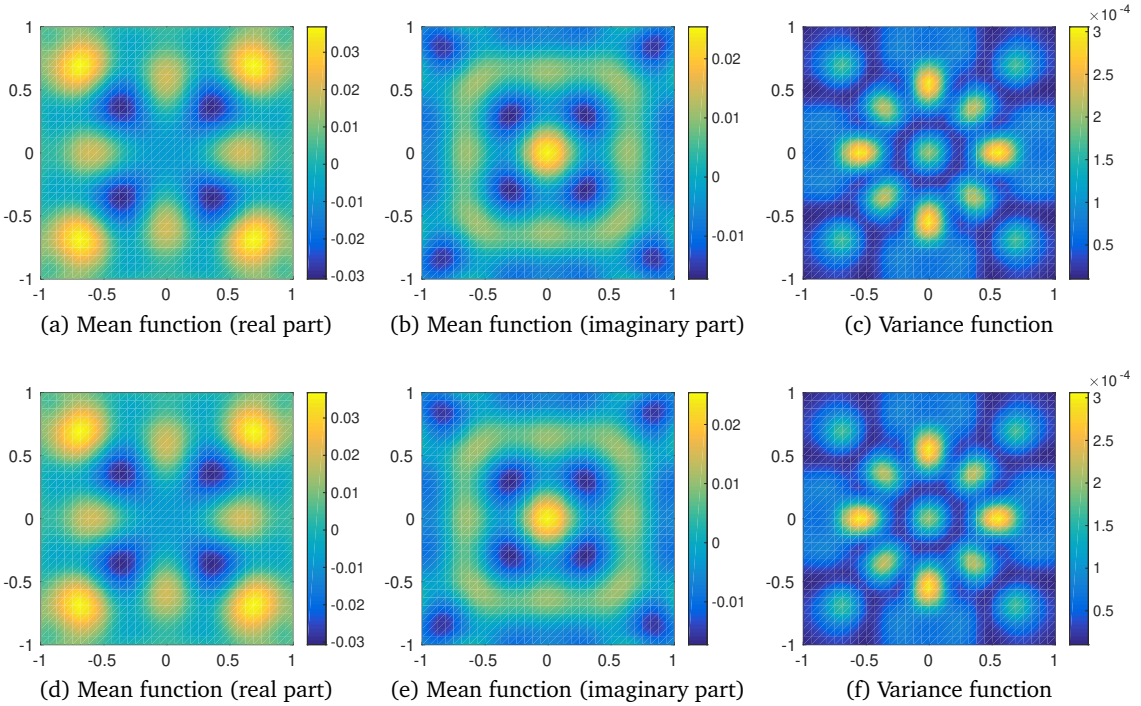
In this test problem, the physical domain considered is $[-1,1]\times[-1,1]$, the boundary conditions in (2.1)–(2.3) are set to $\partial D_R=\partial D$ and $\partial D_D=\emptyset$. The refractive index in this test problem is set to a truncated Karhunen–Loève (KL) expansion [11, 30] of a random field with mean function $\kappa_0(\boldsymbol{x})$, standard deviation σ and covariance function $Cov(\boldsymbol{x},\boldsymbol{y})$,

$$Cov(\boldsymbol{x},\boldsymbol{y})=\sigma^2\exp\left(-\frac{|x_1-y_1|}{c}-\frac{|x_2-y_2|}{c}\right),$$

where $\boldsymbol{x} = [x_1, x_2]^T$, $\boldsymbol{y} = [y_1, y_2]^T$ and the correlation length is set to $c = 4$. The KL expansion is expressed as

$$\kappa(\boldsymbol{x}, \xi) = \kappa_0(\boldsymbol{x}) + \sum_{i=1}^{N} \kappa_i(\boldsymbol{x})\xi_i = \kappa_0(\boldsymbol{x}) + \sum_{i=1}^{N} \sqrt{\lambda_i} c_i(\boldsymbol{x})\xi_i, \tag{4.1}$$

where $\{\lambda_i, c_i(\boldsymbol{x})\}_{i=1}^N$ are eigenpairs of $Cov(\boldsymbol{x}, \boldsymbol{y})$, N is the number of KL modes retained, and $\{\xi_i\}_{i=1}^N$ are uncorrelated random variables. The error associated with truncation of the KL expansion depends on the amount of total variance captured, $\delta_{KL} := (\sum_{i=1}^N \lambda_i)/(|D|\sigma^2)$, where $|D|$ denotes the area of D [30, 43]. In the following, we set $\kappa_0(\boldsymbol{x}) = 10$, $\sigma = 1$. In addition, we set $N = 4$ such that $\delta_{KL} > 89\%$, and set the random variables $\{\xi_i\}_{i=1}^N$ to be independent uniform distributions with range $[-1, 1]$. What is more, the source term in (2.1) is specified as $f(\boldsymbol{x}) = 2(0.5 - x_1^2 - x_2^2)$.



(a) Mean function (real part) (b) Mean function (imaginary part) (c) Variance function

(d) Mean function (real part) (e) Mean function (imaginary part) (f) Variance function

Figure 2: Test problem 1: gPC method with order $p = 10$ (top) and Monte Carlo method with 10^6 samples (bottom).

Figure 2(a), Figure 2(b) and Figure 2(c) show gPC approximations of the mean and variance functions of this test problem, where the order of gPC expansion is $p = 10$, a uniform 33×33 grid is used to discretize the physical space, and the Bi-CGSTAB method with the

mean-based preconditioner (3.3) is used to solve the corresponding linear system (see (2.16)). In this paper, the stopping criterion for the iterative methods is based on the relative residual $\|\boldsymbol{A}\boldsymbol{u}^{(k)} - \boldsymbol{b}\|/\|\boldsymbol{b}\|$, where $\|\cdot\|$ denotes the vector L^2 norm and the superscript k denotes the iteration number, and iteration terminates when the relative residual is smaller than 10^{-8}. In addition, the Monte Carlo method (see [54] for a formal presentation) is tested for comparison. Figure 2(d), Figure 2(e) and Figure 2(f) show the results generated by the Monte Carlo methods, in which the number of samples is 10^6. From these figures, the results generated by the two methods are visually indistinguishable.

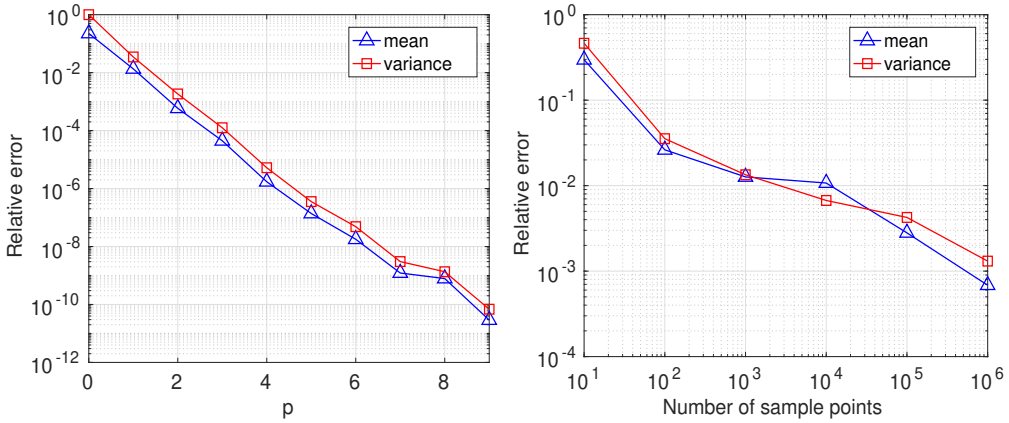
To assess the accuracy of the gPC finite element approximation (2.15), we consider the relative mean and variance errors, which are defined through

$$error_{\text{mean}} := \frac{\|\mathbb{E}(u^{ph}) - \mathbb{E}(u_{\text{ref}})\|_2}{\|\mathbb{E}(u_{\text{ref}})\|_2},\ error_{\text{variance}} := \frac{\|\text{Var}(u^{ph}) - \text{Var}(u_{\text{ref}})\|_2}{\|\text{Var}(u_{\text{ref}})\|_2}, \tag{4.2}$$

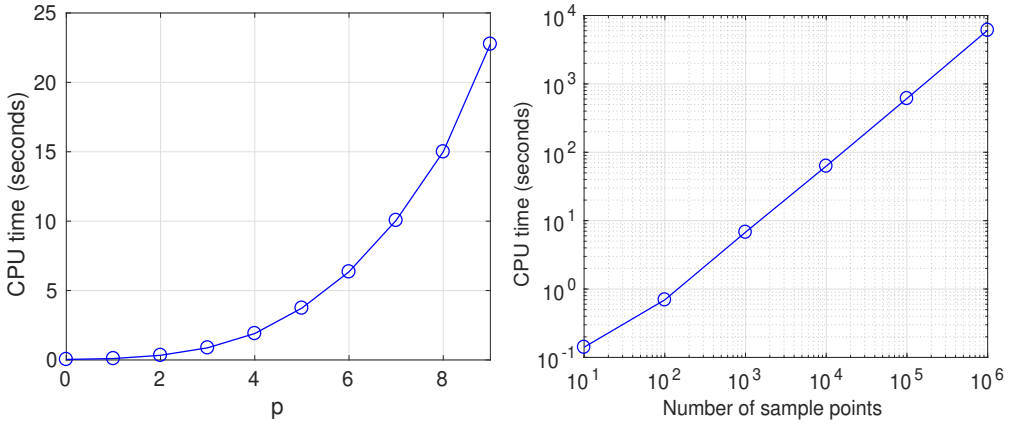
where u_{ref} is a reference solution, and $\|\cdot\|_2$ is defined in (2.7). For the Monte Carlo method, the relative mean and variance errors are defined through replacing u^{ph} in (4.2) by the mean and variance function estimates generated by the Monte Carlo method.

Figure 3(a) and Figure 3(b) show the mean and variance errors of the gPC method and the Monte Carlo method for this test problem, where a uniform 33×33 spatial grid is used. To generate the reference solution, the gPC method with $p = 10$ is used, and the corresponding linear system is solved by preconditioned Bi-CGSTAB. Figure 3(a) shows that the errors of the gPC approximation decrease quickly as the gPC order increases. Compared with the Monte Carlo method, of which the errors are shown Figure 3(b), the gPC method is efficient—the errors of the gPC approximation with $p = 2$ are smaller than those of the Monte Carlo method with 10^6 samples. Figure 3(c) and Figure 3(d) show the the CPU times of the gPC method and the Monte Carlo method. For the gPC method, the CPU time includes the time for constructing the linear system and solving it using preconditioned Bi-CGSTAB method. For the Monte Carlo method, the CPU time includes the times for constructing and solving linear systems (using the MATLAB backslash solver) associated with deterministic problems at all input sample points. All results in this paper are obtained in MATLAB on a desktop with 3.60GHz Intel Core i7 CPU. From 3(c) and Figure 3(d), it is clear that to achieve a given accuracy, the gPC method requires significantly less CPU times than the Monte Carlo method.

Table 1 shows the number of iterations and the CPU times of the preconditioned iterative methods (including the time for setting up the preconditioners and the time for iterations), associated with different mesh sizes and gPC orders. It can be seen that the numbers of iterations are independent of the mesh size h and the gPC order p, and the numbers are small in general. Figure 4 shows the CPU times versus the number of unknowns ($N_x N_\xi$). Since the solpe of the black line ($y = x$) in the figure is 1, the CPU times (roughly) increase linearly as the number of unknowns increases, which is expected for preconditioned iterative methods.



(a) Errors of the gPC method (b) Errors of the Monte Carlo method



(c) CPU time of the gPC method (d) CPU time of the Monte Carlo method

Figure 3: Errors and CPU times of gPC and Monte Carlo for test problem 1, uniform 33×33 spatial grid.

Table 1: Numbers of iterations and CPU times (shown in brackets) in seconds for preconditioned iterative solvers, test problem 1.

Iterative method	h^{-1}	$p=2$	$p=4$	$p=6$	$p=8$	$p=10$
Bi-CGSTAB	16	8.5(0.36)	10(1.91)	10.5(6.35)	10.5(15.07)	11(31.88)
	32	8.5(1.59)	10.5(9.83)	10.5(29.59)	10.5(68.69)	11(161.58)
	64	8.5(10.78)	10.5(59.55)	10.5(181.17)	11(476.87)	11(975.86)
QMR	16	16(0.72)	19(3.74)	22(13.41)	21(30.13)	22(63.27)
	32	16(3.27)	19(18.58)	21(61.76)	22(148.88)	22(346.83)
	64	16(21.47)	19(110.59)	21(368.79)	22(982.89)	22(1975.09)

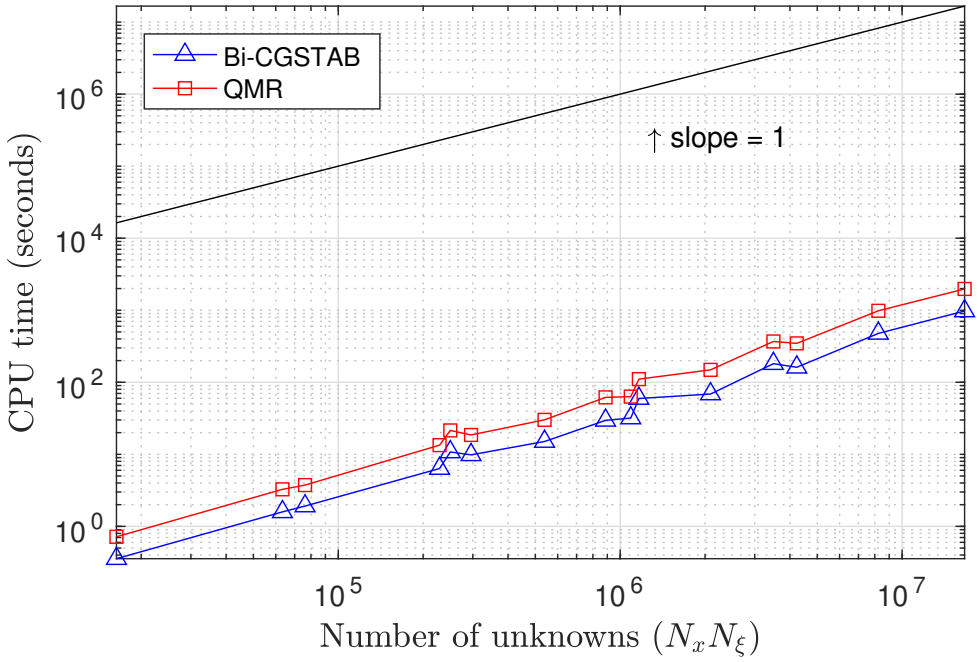


Figure 4: CPU times for preconditioned iterative methods, test problem 1.

4.2. Test problem 2 (a problem close to a resonant frequency)

In this test problem, we consider the refractive index in the form of

$$\kappa(\boldsymbol{x},\xi)=\kappa_0+\kappa_1\xi, \tag{4.3}$$

where ξ is uniformly distributed in $[-1,1]$ and $\kappa_i, i = 0,1$ are two constants which will be specified in the following. A pure Dirichlet boundary condition (2.2) is applied, i.e. $\partial D = \partial D_D$. The eigenvalues of the corresponding eigenvalue problem

$$-\nabla^2 u(\boldsymbol{x},\xi) - \kappa^2(\boldsymbol{x},\xi)u(\boldsymbol{x},\xi) = \lambda(\xi)u(\boldsymbol{x},\xi)$$

with boundary condition (2.2) ($\partial D = \partial D_D$) are

$$\lambda_{i,j}(\xi) = \frac{(i^2+j^2)\pi^2}{4} - \kappa^2(\xi), \tag{4.4}$$

where $i, j = 1, 2, \ldots$.

The right hand side of (2.1) for this test problem is set to the normalized eigenfunction corresponding to $\lambda_{1,1}$, which is

$$f = \cos(\frac{\pi x_1}{2})\cos(\frac{\pi x_2}{2}).$$

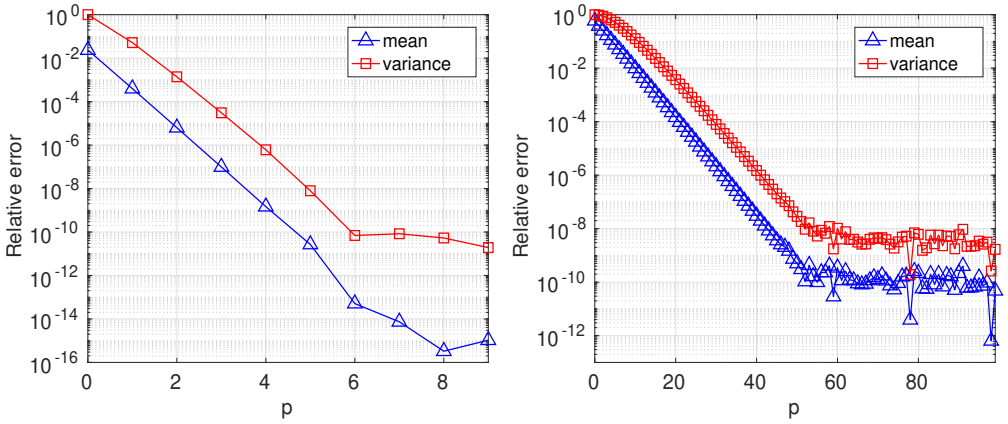
We note that the exact solution of this test problem can be written explicitly, which is $u(\boldsymbol{x},\xi) = f(\boldsymbol{x})/\lambda_{1,1}(\xi)$, but we next solve this problem using the stochastic Galerkin method to test the performance of the gPC approximation and the mean-based preconditioning scheme. When $\kappa(\xi)$ takes the values of $(\sqrt{i^2+j^2}\pi)/2$ (i.e., $\lambda_{i,j} = 0$ in (4.4)) for $i, j = 1, 2, \ldots$, the solution of the deterministic version of this test problem is not unique, which is called resonance. We focus on the situation that κ can be close to the first resonant frequency $\pi/\sqrt{2}$. Specifically, we set $\kappa_0 = \pi/\sqrt{2} + 0.41$ and consider the following two cases of κ_1 (see (4.3) for the definitions of κ_0 and κ_1):

- $\kappa_1 = 0.1$, with $|\lambda_{1,1}(\xi)| \in [1.47, 2.53]$ for $\xi \in [-1,1]$;
- $\kappa_1 = 0.4$, with $|\lambda_{1,1}(\xi)| \in [0.04, 4.26]$ for $\xi \in [-1,1]$.

It is clear that the stochastic Helmholtz problem (2.1)–(2.3) associated with $\kappa_1 = 0.1$ is away from resonance, while the problem associated with $\kappa_1 = 0.4$ is close to resonance.

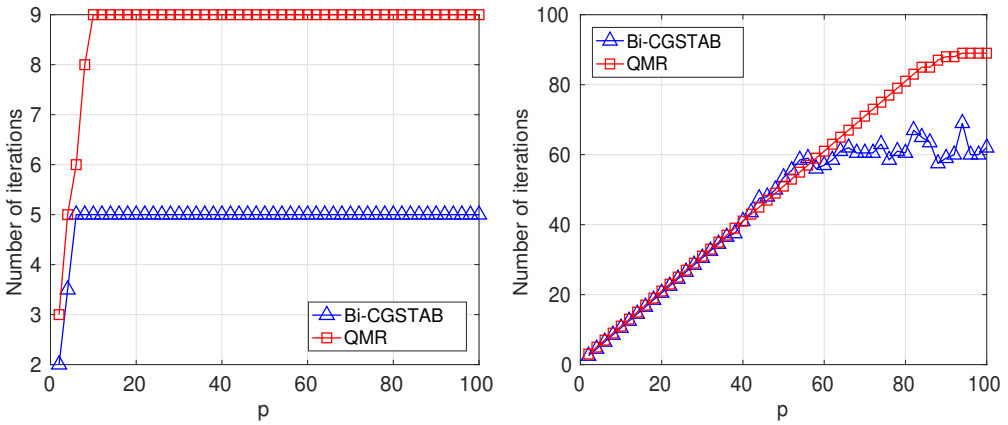
Figure 5 shows the mean and variance errors (defined in (4.2)) for this test problem, where a uniform 33×33 spatial grid is used. To generate the reference solutions for both cases ($\kappa_1 = 0.1$ and $\kappa_1 = 0.4$), the gPC method with order $p = 100$ is used, and the corresponding linear systems are solved by preconditioned Bi-CGSTAB. Figure 5(a) shows that, the mean and variance errors of the gPC approximation decrease quickly as the gPC order increases for the problem that is away from resonance ($\kappa_1 = 0.1$). However, from Figure 5(b), for the problem that is close to resonance ($\kappa_1 = 0.4$), although the errors of the gPC approximation still decrease as the gPC order increases, they decrease much slower than those associated with $\kappa = 0.1$ (shown in Figure 5(a)).

Figure 6 shows the number of iterations for preconditioned solvers for this test problem, where a 33×33 spatial grid is used. From Figure 6(a) ($\kappa_1 = 0.1$ and the problem is away



(a) $\kappa_1 = 0.1$ (b) $\kappa_1 = 0.4$

Figure 5: Errors of the gPC method for test problem 2, uniform 33×33 spatial grid.



(a) $\kappa_1 = 0.1$ (b) $\kappa_1 = 0.4$

Figure 6: Numbers of iterations for preconditioned iterative methods for test problem 2, uniform 33×33 spatial gird.

from resonance), the numbers of iterations of Bi-CGSTAB and QMR are small (only five for Bi-CGSTAB and nine for QMR), and they are independent of gPC order p. However, for the problem close to resonance, of which the results are shown in Figure 6(b), the numbers of iterations for both Bi-CGSTAB and QMR becomes large—for gPC order $p = 60$, both preconditioned Bi-CGSTAB and preconditioned QMR require around sixty iterations to achieve the residual stopping tolerance 10^{-8}. Looking at Figure 6(b) in more detail, as the gPC order p increases,

the number of iterations of Bi-CGSTAB keeps increasing until $p \approx 50$, while that of QMR keeps increasing until $p \approx 90$. Note that the smallest magnitude of $\lambda_{1,1}(\xi)$ is much smaller when $\kappa_1 = 0.4$ than that of $\kappa_1 = 0.1$. Moreover, the solution is given by $u(\boldsymbol{x}, \xi) = f(\boldsymbol{x})/\lambda_{1,1}(\xi)$. These mean that the magnitude of variance function is much bigger when $\kappa_1 = 0.4$ than that of $\kappa_1 = 0.1$. When the variance function becomes large, the efficiency of the mean-based preconditioner can deteriorate.

5. Conclusions

This paper describes the mathematical framework and implementation of spectral stochastic finite element methods for solving the Helmholtz equation with random inputs. The sparsity of the corresponding linear system is analyzed, and iterative methods combined with the mean-based preconditioning scheme are investigated. From all examples considered, it can be seen that the gPC approximation and the mean-based preconditioning scheme are efficient when the stochastic Helmholtz problems are not too close to a resonant frequency, while both the gPC approximation and the mean-based preconditioner become less efficient for problems close to resonance. In addition, the numerical studies in this work only focus on low frequency waves (i.e., the refractive index is small in our setting). We will investigate more efficient gPC-based approximations and fast iterative solvers for both problems close to resonance and problems with high-frequency waves in our future work.

Acknowledgments

The authors thank David Silvester and Catherine Powell for helpful suggestions and discussions. This work is supported by the National Natural Science Foundation of China (No. 11601329) and the science challenge project (No. TZ2018001).

References

[1] R. Askey and M. Ismail, *Recurrence relations, continued fractions and orthogonal polynomials*, American Mathematical Soc. (1984).

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

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

[4] A. Bespalov, C. Powell, and D. Silvester, *S-IFISS version 1.0*, (2013). Available online at http://www.manchester.ac.uk/ifiss/s-ifiss1.0.tar.gz.

[5] J. P. Boyd, *Chebyshev and Fourier spectral methods*, Courier Corporation (2001).

[6] D. Braess, *Finite Elements*, Cambridge University Press, London (1997).

[7] M. Cheng, T. Y. Hou, and Z. Zhang, *A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations I: Derivation and algorithms*, J. Comput. Phys. **242**, 843–868 (2013).

[8] M. Cheng, T. Y. Hou, and Z. Zhang, *A dynamically bi-orthogonal method for time-dependent stochastic partial differential equations II: Adaptivity and generalizations*, J. Comput. Phys. **242**, 753–776 (2013).
[9] M. K. Deb, I. M. Babuška, and J. T. Oden, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Engrg. **190**, 6359–6372 (2001).
[10] M. Eiermann, O. G. Ernst, and E. Ullmann, *Computational aspects of the stochastic finite element method*, Comput. Vis. Sci. **10**, 3–15 (2007).
[11] H. Elman and D. Furnival, *Solving the stochastic steady-state diffusion problem using multigrid*, IMA J. Numer. Anal. **27**, 675–688 (2007).
[12] H. Elman and Q. Liao, *Reduced basis collocation methods for partial differential equations with random coefficients*, SIAM/ASA J. Uncertain. Quantif. **1**, 192–217 (2013).
[13] H. Elman, A. Ramage, and D. Silvester, *IFISS: A computational laboratory for investigating incompressible flow problems*, SIAM Rev. **56**, 261–273 (2014).
[14] H. C. Elman, O. G. Ernst, D. P. OLeary, and M. Stewart, *Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering*, Comput. Methods Appl. Mech. Engrg. **194**, 1037–1055 (2005).
[15] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Oxford University Press (2014).
[16] Y. A. Erlangga, *Advances in iterative methods and preconditioners for the Helmholtz equation*, Arch. Comput. Methods Eng. **15**, 37–66 (2008).
[17] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee, *On a class of preconditioners for solving the Helmholtz equation*, Appl. Numer. Math. **50**, 409–425 (2004).
[18] O. G. Ernst and M. J. Gander, *Why it is difficult to solve Helmholtz problems with classical iterative methods*, in: *Numerical analysis of multiscale problems*, I. Graham, T. Hou, L. O., R. Scheichl (Eds), pp. 325–363, Springer (2012).
[19] O. G. Ernst and E. Ullmann, *Stochastic Galerkin matrices*, SIAM J. Matrix Anal. Appl. **31**, 1848–1872 (2010).
[20] W. Feller, *An introduction to probability theory and its applications: volume I*, John Wiley & Sons London-New York-Sydney-Toronto (1968).
[21] X. Feng, J. Lin, and D. Nicholls, *An efficient Monte Carlo-transformed field expansion method for electromagnetic wave scattering by random rough surface*, Commun. Comput. Phys. **23**, 685-705 (2018).
[22] X. Feng, J. Lin, and C. Lorton, *An efficient numerical method for acoustic wave scattering in random media*, SIAM/ASA J. Uncertain. Quantif. **3**, 790–822 (2015).
[23] Z. Feng, J. Li, T. Tang and T. Zhou, *Efficient stochastic Galerkin methods for Maxwell's equations with random input*, submitted to J. Sci. Comput., (2018).
[24] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in: *Numerical analysis*, pp. 73–89, Springer (1976).
[25] R. W. Freund and N. M. Nachtigal, *QMR: a quasi-minimal residual method for non-hermitian linear systems*, Numer. Math. **60**, 315–339 (1991).
[26] R. W. Freund and N. M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Comput. **15**, 313–337 (1994).
[27] A. Frolov and E. Kartchevskiy, *Integral equation methods in optical waveguide theory*, in: *Inverse Problems and Large-Scale Computations*, pp. 119–133, Springer (2013).
[28] M. J. Gander, I. G. Graham, and E. A. Spence, *Applying GMRES to the Helmholtz equation with shifted laplacian preconditioning: what is the largest shift for which wavenumber-independent con-*

vergence is guaranteed? Numer. Math. **131**, 567–614 (2015).
[29] R. G. Ghanem and R. M. Kruger, *Numerical solution of spectral stochastic finite element systems*, Comput. Methods Appl. Mech. Engrg. **129**, 289–303 (1996).
[30] R. G. Ghanem and P. D. Spanos, *Stochastic finite elements: a spectral approach*, Courier Corporation (2003).
[31] G. H. Golub and C. F. Van Loan, *Matrix computations*, JHU Press (2013).
[32] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, NBS (1952).
[33] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational ocean acoustics*, Springer Science & Business Media (2011).
[34] E. Karchevskii and S. Solovev, *Investigation of a spectral problem for the Helmholtz operator on the plane*, Differ. Equ. **36**, 631–634 (2000).
[35] E. Kartchevski, A. Nosich, and G. Hanson, *Mathematical analysis of the generalized natural modes of an inhomogeneous optical fiber*, SIAM J. Appl. Math. **65**, 2033–2048 (2005).
[36] K. Lee and H. C. Elman, *A preconditioned low-rank projection method with a rank-reduction scheme for stochastic partial differential equations*, SIAM J. Sci. Comput. **39**, S828–S850 (2017).
[37] R. März, *Integrated optics: design and modeling*, Artech House on Demand (1995).
[38] E. Musharbash, F. Nobile and T. Zhou, *Error analysis of the dynamically orthogonal approximation of time dependent random PDEs*, SIAM J. Sci. Comput. **37**, A776–A810 (2015).
[39] L. Ng and K. Willcox, *Multifidelity approaches for optimization under uncertainty*, Internat. J. Numer. Methods Engrg. **100**, 746–772 (2014).
[40] L. W.-T. Ng and M. Eldred, *Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation*, in: *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, (2012).
[41] M. F. Pellissetti and R. G. Ghanem, *Iterative solution of systems of linear equations arising in the context of stochastic finite elements*, Adv. Eng. Softw. **31**, 607–616 (2000).
[42] C. Powell and D. Silvester, *Preconditioning steady-state Navier-Stokes equations with random data*, SIAM J. Sci. Comput. **34**, A2482–A2506 (2012).
[43] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite element systems*, IMA J. Numer. Anal. **29**, 350–375 (2009).
[44] Y. Saad, *Iterative methods for sparse linear systems*, SIAM (2003).
[45] J. Shen and T. Tang, *Spectral and high-order methods with applications*, Science Press (2006).
[46] P. Sonneveld, *CGS, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput. **10**, 36–52 (1989).
[47] B. Sousedík and H. Elman, *Stochastic Galerkin methods for the steady-state Navier-Stokes equations*, J. Comput. Phys. **316**, 435–452 (2016).
[48] B. Sousedík, R. G. Ghanem, and E. T. Phipps, *Hierarchical Schur complement preconditioner for the stochastic Galerkin finite element methods*, Numer. Linear Algebra Appl. **21**, 136–151 (2014).
[49] L. Tamellini, O. L. Maître, and A. Nouy, *Model reduction based on proper generalized decomposition for the stochastic steady incompressible Navier-Stokes equations*, SIAM J. Sci. Comput. **36**, A1089–A1117 (2014).
[50] H. A. Van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput. **13**, 631–644 (1992).
[51] C. Vassallo, *Optical waveguide concepts*, Elsevier (1991).
[52] D. Xiu, *Fast numerical methods for stochastic computations: a review*, Commun. Comput. Phys. **5**, 242–272 (2009).
[53] D. Xiu, *Numerical methods for stochastic computations: a spectral method approach*, Princeton Uni-

versity Press (2010).
[54] D. Xiu and J. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput. **27**, 1118–1139 (2005).
[55] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos*, Comput. Methods Appl. Mech. Engrg. **191**, 4927–4948 (2002).
[56] D. Xiu and G. E. Karniadakis, *The wiener-askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput. **24**, 619–644 (2002).
[57] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in flow simulations via generalized polynomial chaos*, J. Comput. Phys. **187**, 137–167 (2003).
[58] D. Xiu and J. Shen, *An efficient spectral method for acoustic scattering from rough surfaces*, Commun. Comput. Phys. **2**, 54–72 (2007).
[59] D. Xiu and J. Shen, *Efficient stochastic Galerkin methods for random diffusion equations*, J. Comput. Phys. **228**, 266–281 (2009).
[60] T. Tang and T. Zhou, *Convergence analysis for stochastic collocation methods to scalar hyperbolic equations*, Commun. Comput. Phys. **8**, 226–248 (2010).
[61] T. Zhou and T. Tang, *Galerkin methods for stochastic hyperbolic problems using bi-orthogonal polynomials*, J. Sci. Comput. **51**, 274-292 (2012).