




# The Helmholtz Equation with Uncertainties in the Wavenumber

Roland Pulch<sup>1</sup> · Olivier Sète<sup>1</sup> 

Received: 3 January 2023 / Revised: 11 October 2023 / Accepted: 2 January 2024 /  
Published online: 5 February 2024  
© The Author(s) 2024

## Abstract

We investigate the Helmholtz equation with suitable boundary conditions and uncertainties in the wavenumber. Thus the wavenumber is modeled as a random variable or a random field. We discretize the Helmholtz equation using finite differences in space, which leads to a linear system of algebraic equations including random variables. A stochastic Galerkin method yields a deterministic linear system of algebraic equations. This linear system is high-dimensional, sparse and complex symmetric but, in general, not hermitian. We therefore solve this system iteratively with GMRES and propose two preconditioners: a complex shifted Laplace preconditioner and a mean value preconditioner. Both preconditioners reduce the number of iteration steps as well as the computation time in our numerical experiments.

**Keywords** Helmholtz equation · Polynomial chaos · Stochastic Galerkin method · GMRES · Complex shifted Laplace preconditioner · Mean value preconditioner

**Mathematics Subject Classification** 65N30 · 65C20 · 35R60

## 1 Introduction

The Helmholtz equation is a linear partial differential equation (PDE), whose solutions are time-harmonic states of the wave equation, see [15, 20]. Important applications of this model are given in acoustics and electromagnetics [2]. The Helmholtz equation includes a wavenumber, which is either a constant parameter or a space-dependent function. Furthermore, boundary conditions are imposed on the spatial domain.

We consider uncertainties in the wavenumber. Thus the wavenumber is replaced by a random variable or a spatial random field to quantify the uncertainties. The solution of the Helmholtz equation changes into a random field, which can be expanded into the (gener-

---

✉ Olivier Sète  
olivier.sete@uni-greifswald.de

Roland Pulch  
roland.pulch@uni-greifswald.de

<sup>1</sup> Institute of Mathematics and Computer Science, Universität Greifswald, Walther-Rathenau-Straße 47, 17489 Greifswald, Germany

alized) polynomial chaos, see [31]. We employ the stochastic Galerkin method to compute approximations of the unknown coefficient functions. Stochastic Galerkin methods were used for linear PDEs of different types including random variables, for example, see [13, 33] on elliptic type, [14, 24] on hyperbolic type, and [22, 32] on parabolic type. Wang et al. [30] applied a multi-element stochastic Galerkin method to solve the Helmholtz equation including random variables. We investigate the ordinary stochastic Galerkin method, which is efficient if the wavenumbers are not close to resonance.

The stochastic Galerkin method transforms the random-dependent Helmholtz equation into a deterministic system of linear PDEs. Likewise, the original boundary conditions yield boundary conditions for this system. We examine the system of PDEs in one and two space dimensions. A finite difference method, see [16], produces a high-dimensional linear system of algebraic equations. When considering absorbing boundary conditions, the coefficient matrices are complex-valued and non-hermitian.

We focus on the numerical solution of the linear systems of algebraic equations. The dimension of these linear systems rapidly grows for increasing numbers of random variables. Hence we use iterative methods like GMRES [27] in the numerical solution. The efficiency of an iterative method strongly depends on an appropriate preconditioning of the linear systems. We propose two preconditioners in the general case where the wavenumber can depend on space and on multiple random variables: a complex shifted Laplace preconditioner, see [6, 8], and a mean value preconditioner, see [10, 30]. Statements on the location of spectra and estimates of matrix norms are shown. Furthermore, results of numerical computations are presented for both settings.

The article is organized as follows. The stochastic Helmholtz equation is introduced in Sect. 2 and discretized in Sect. 3. We discuss the complex shifted Laplace preconditioner in Sect. 4 and the mean value preconditioner in Sect. 5. Sections 6 and 7 contain numerical experiments in one and two spatial dimensions, respectively, which show the effectiveness of the preconditioners.

## 2 Problem Definition

We illustrate the stochastic problem associated to the Helmholtz equation.

### 2.1 Helmholtz Equation

The Helmholtz equation is a PDE of the form

$$-\Delta u - k^2 u = f \quad \text{in } Q \quad (1)$$

with an (open) spatial domain  $Q \subseteq \mathbb{R}^d$  and given source term  $f : Q \rightarrow \mathbb{R}$ . The wavenumber  $k$  is either a positive constant or a function  $k : \overline{Q} \rightarrow \mathbb{R}_+$ . The unknown solution is  $u : \overline{Q} \rightarrow \mathbb{K}$  with either  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{K} = \mathbb{C}$ . Here  $\Delta = \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}$  denotes the Laplace operator with respect to  $x = [x_1, \dots, x_d]^\top \in \mathbb{R}^d$ .

Often homogeneous Dirichlet boundary conditions, i.e.,

$$u = 0 \quad \text{on } \partial Q, \quad (2)$$

are applied for simplicity. Alternatively, absorbing boundary conditions read as

$$\partial_n u - iku = 0 \quad \text{on } \partial Q, \quad (3)$$

where  $\partial_n$  denotes the derivative with respect to the outward normal of  $Q$  and  $i = \sqrt{-1}$  is the imaginary unit.

### 2.2 Stochastic Modeling

We consider uncertainties in the wavenumber. A simple model to include a variation of the wavenumber is to replace the constant  $k$  by a random variable on a probability space  $(\Omega, \mathcal{A}, P)$ . We write  $k = k(\xi)$ , where  $\xi : \Omega \rightarrow \mathbb{R}$  is some random variable with a traditional probability distribution. More generally, the wavenumber can be a space-dependent function on  $\overline{Q}$  including a multidimensional random variable  $\xi : \Omega \rightarrow \Xi$  with  $\Xi \subseteq \mathbb{R}^s$ . We assume  $\xi = (\xi_1, \dots, \xi_s)^\top$  with independent random variables  $\xi_\ell$  for  $\ell = 1, \dots, s$ . Now the wavenumber becomes a random field

$$k(x, \xi) = k_0(x) + \sum_{\ell=1}^s \xi_\ell k_\ell(x) \tag{4}$$

with given functions  $k_\ell : \overline{Q} \rightarrow \mathbb{R}$  for  $\ell = 0, 1, \dots, s$ , as in [30]. A truncation of a Karhunen-Loève expansion, see [11, p. 17], also yields a random input of the form (4). Consequently, the solution of the deterministic Helmholtz equation (1) changes into a random field  $u : \overline{Q} \times \Xi \rightarrow \mathbb{K}$ . We write  $u(x, \xi)$  to indicate the dependence of the solution on space as well as the random variables.

We assume that each random variable  $\xi_\ell$  has a probability density function  $\rho_\ell$ . Since the random variables are independent, the product  $\rho = \rho_1 \cdots \rho_s$  is the joint probability density function. Without loss of generality, let  $\rho(\xi) > 0$  for almost all  $\xi \in \Xi$ . The expected value of a measurable function  $f : \Xi \rightarrow \mathbb{K}$  depending on the random variables is

$$\mathbb{E}(f) = \int_{\Omega} f(\xi(\omega)) \, dP(\omega) = \int_{\Xi} f(\xi) \rho(\xi) \, d\xi,$$

if the integral is finite. The inner product of two square-integrable functions  $f, g$  is

$$\langle f, g \rangle = \int_{\Xi} f(\xi) \overline{g(\xi)} \rho(\xi) \, d\xi. \tag{5}$$

In the following,  $\mathcal{L}^2(\Xi, \rho)$  denotes the Hilbert space of square-integrable functions. The associated norm is  $\|f\|_{\mathcal{L}^2(\Xi, \rho)} = \sqrt{\langle f, f \rangle}$ .

Later we will focus on uniformly distributed random variables  $\xi_\ell : \Omega \rightarrow [-1, 1]$ . In this case, the joint probability density function is constant, i.e.,  $\Xi = [-1, 1]^s$  and  $\rho \equiv 2^{-s}$ .

### 2.3 Polynomial Chaos Expansions

We assume that there is an orthonormal polynomial basis  $(\phi_i)_{i \in \mathbb{N}_0}$  in  $\mathcal{L}^2(\Xi, \rho)$ . Thus it holds that

$$\langle \phi_i, \phi_j \rangle = \delta_{i,j} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

with the inner product (5). In the case of uniform probability distributions, the multivariate functions  $\phi_i$  are products of the (univariate) Legendre polynomials. We assume that  $\phi_0 \equiv 1$ . The number  $m + 1$  of multivariate polynomials in  $s$  variables up to a total degree  $r$  is

$$m + 1 = \frac{(s + r)!}{s! r!}, \tag{6}$$

see [31, p. 65]. This number grows fast for increasing  $r$  or  $s$ .

Let  $u(x, \cdot) \in \mathcal{L}^2(\Xi, \rho)$  for each  $x \in \overline{Q}$ . The polynomial chaos (PC) expansion is

$$u(x, \xi) = \sum_{i=0}^{\infty} v_i(x) \phi_i(\xi) \tag{7}$$

with (a priori unknown) coefficient functions

$$v_i(x) = \langle u(x, \xi), \phi_i(\xi) \rangle \quad \text{for } i \in \mathbb{N}_0. \tag{8}$$

The series (7) converges in  $\mathcal{L}^2(\Xi, \rho)$  pointwise for  $x \in \overline{Q}$ . If the wavenumber  $k$  is an analytic function of the random variables, then the rate of convergence is exponentially fast for traditional probability distributions.

### 3 Discretization of the Stochastic Helmholtz Equation

We consider the stochastic Helmholtz equation

$$-\Delta u(x, \xi) - k(x, \xi)^2 u(x, \xi) = f(x), \quad x \in Q \subseteq \mathbb{R}^d, \tag{9}$$

with given source term  $f : Q \rightarrow \mathbb{R}$  and random wavenumber  $k : \overline{Q} \times \Xi \rightarrow \mathbb{R}_+$ , together with either homogeneous Dirichlet boundary conditions

$$u(x, \xi) = 0, \quad x \in \partial Q, \quad \xi \in \Xi, \tag{10}$$

or with absorbing boundary conditions

$$\partial_n u(x, \xi) - ik(x, \xi)u(x, \xi) = 0, \quad x \in \partial Q, \quad \xi \in \Xi. \tag{11}$$

All derivatives are taken with respect to  $x$ . We discretize this boundary value problem in two steps, with a finite difference method (FDM) in space and the stochastic Galerkin method in the random-dependent part. The steps can be done in any order. We first give an overview of the procedure when beginning with the FDM in Sect. 3.1. In Sect. 3.2, we discuss the discretization when beginning with the stochastic Galerkin method.

#### 3.1 FDM and Stochastic Galerkin Method

A spatial discretization of the boundary value problem with a second order FDM on an equispaced grid leads to a (stochastic) linear algebraic system

$$S(\xi)U(\xi) = F_0 \tag{12}$$

with constant vector  $F_0 \in \mathbb{R}^n$  and stochastic matrix  $S(\xi) \in \mathbb{K}^{n,n}$  for  $\xi \in \Xi$ , depending on the boundary conditions. In the case of homogeneous Dirichlet boundary conditions, it follows that

$$S(\xi) = T - D_2(\xi), \tag{13}$$

with  $T$  and  $D_2(\xi)$  symmetric positive definite, and in case of absorbing boundary conditions,

$$S(\xi) = T - iD_1(\xi) - D_2(\xi), \tag{14}$$

with  $T$  and  $D_1(\xi)$  symmetric positive semidefinite and  $D_2(\xi)$  symmetric positive definite. (For details of this discretization, see the appendix of [23].) In a second step, we consider a PC approximation of  $U(\xi)$  of the form

$$\tilde{U}_m(\xi) = \sum_{i=0}^m \phi_i(\xi) V_i, \quad \text{where } V_i = [v_{\ell,i}]_{\ell=1}^n \in \mathbb{R}^n \text{ for } i = 0, 1, \dots, m, \quad (15)$$

and  $\phi_i$  are polynomials as in Sect. 2.3. The coefficient vectors  $V_i$  are determined by the orthogonality of the residual

$$R_m(\xi) = S(\xi)\tilde{U}_m(\xi) - F_0. \quad (16)$$

to the subspace  $\text{span}\{\phi_0, \phi_1, \dots, \phi_m\}$  with respect to the inner product  $\langle \cdot, \cdot \rangle$  in (5), i.e., by  $\langle R_m(\xi), \phi_i(\xi) \rangle = 0$  for  $i = 0, 1, \dots, m$ . Here the inner product is taken component-wise. The orthogonality condition is equivalent to

$$\langle S(\xi)\tilde{U}_m(\xi), \phi_i(\xi) \rangle = \langle 1, \phi_i(\xi) \rangle F_0 = \delta_{i,0} F_0, \quad i = 0, 1, \dots, m, \quad (17)$$

due to  $\phi_0 \equiv 1$ . This leads to a (deterministic) linear algebraic system

$$AV = F, \quad V = \begin{bmatrix} V_0 \\ \vdots \\ V_m \end{bmatrix}, \quad F = \begin{bmatrix} F_0 \\ \vdots \\ F_m \end{bmatrix}, \quad (18)$$

where the stochastic Galerkin projection  $A \in \mathbb{K}^{(m+1)n, (m+1)n}$  is a block matrix with  $m + 1$  blocks of size  $n \times n$ , and  $F_i = 0 \in \mathbb{R}^n$  for  $i = 1, \dots, m$ .

**Remark 1** The Galerkin approximation (15) can be interpreted as a spatial discretization of a Galerkin approximation  $\tilde{u}_m(x, \xi) = \sum_{i=0}^m v_i(x)\phi_i(\xi)$  of  $u(x, \xi)$ . Evaluating  $\tilde{u}_m$  at discretization points  $x_1, \dots, x_n$  yields

$$\begin{bmatrix} \tilde{u}_m(x_1, \xi) \\ \vdots \\ \tilde{u}_m(x_n, \xi) \end{bmatrix} = \sum_{i=0}^m \phi_i(\xi) \begin{bmatrix} v_i(x_1) \\ \vdots \\ v_i(x_n) \end{bmatrix}. \quad (19)$$

Hence  $V_i$  in (15) can be interpreted as a discretization of  $v_i(x)$  by  $v_{\ell,i} = v_i(x_\ell)$ .

The matrix  $S(\xi)$  in (13) and (14) is a (complex) linear combination of real symmetric positive (semi-)definite matrices. The following lemma shows that this structure is preserved in the stochastic Galerkin method; see [21, Lem. 1] and its proof. These properties of the matrix  $S$  and thus  $A$  will be essential for our analysis of shifted Laplace preconditioners in Sect. 4.

**Lemma 2** Let  $A(\xi) = [a_{\mu,v}(\xi)]_{\mu,v} \in \mathbb{R}^{n,n}$  with  $a_{\mu,v} \in \mathcal{L}^2(\Xi, \rho)$ , and  $V \in \mathbb{R}^n$ . Define

$$A_{ij} := [\langle a_{\mu,v}(\xi)\phi_i(\xi), \phi_j(\xi) \rangle]_{\mu,v} \in \mathbb{R}^{n,n}, \quad i, j = 0, 1, \dots, m, \quad (20)$$

and the stochastic Galerkin projection

$$A := [A_{ij}]_{i,j} \in \mathbb{R}^{(m+1)n, (m+1)n}. \quad (21)$$

We then obtain for  $i, j = 0, 1, \dots, m$

$$\langle A(\xi)\phi_i(\xi)V, \phi_j(\xi) \rangle = A_{ij}V, \quad (22)$$

where the inner product is taken component-wise. Additionally,  $A_{ij} = A_{ji}$ . Moreover, if  $A(\xi)$  is symmetric, then  $A$  is symmetric, and if  $A(\xi)$  is symmetric positive (semi-)definite for almost all  $\xi \in \Xi$ , then  $A$  is symmetric positive (semi-)definite.

**Corollary 3** *In the notation of Lemma 2, if  $A(\xi) = A_0$  is independent of  $\xi$ , then  $A_{ij} = \delta_{ij} A_0$  and  $A = I_{m+1} \otimes A_0$ , with the identity matrix  $I_{m+1} \in \mathbb{K}^{m+1, m+1}$  and the Kronecker product.*

Finally, we obtain the following result on the structure of the matrix  $A$  in (18).

**Theorem 4** *Let the spatial dimension be  $d \in \{1, 2\}$ . A finite difference and stochastic Galerkin approximation of the Helmholtz equation (9) on  $Q = ]0, 1[^d$  with either homogeneous Dirichlet or absorbing boundary conditions leads to a linear system (18) with coefficient matrix*

$$A = L - iB - K \tag{23}$$

and real-valued matrices  $L, B, K$ . The matrix  $K$  is symmetric positive definite,  $B, L$  are symmetric positive semidefinite. In case of homogeneous Dirichlet boundary conditions,  $L$  is symmetric positive definite and  $B = 0$ .

**Proof** The statement of the theorem follows in each case by applying the stochastic Galerkin approximation as described above to (12) and using Lemma 2 as well as Corollary 3 separately for each term composing  $S(\xi)$ ; see (13) and (14). □

The matrix  $L$  results essentially from the discretization of the Laplacian,  $B$  from the (absorbing) boundary conditions, and  $K$  is the discretization of the term including the wavenumber.

### 3.2 Stochastic Galerkin Method and FDM

Alternatively, we can begin with the stochastic Galerkin method. This leads to a system of deterministic PDEs, which are subsequently discretized by a FDM. The PC expansion (7) suggests a stochastic Galerkin approximation of  $u(x, \xi)$  of the form

$$\tilde{u}_m(x, \xi) = \sum_{i=0}^m v_{i,m}(x) \phi_i(\xi). \tag{24}$$

The coefficient functions  $v_{i,m}$  in the stochastic Galerkin method are in general distinct from the coefficients  $v_i$  in (8). Nevertheless, we will usually write  $v_i$  instead of  $v_{i,m}$  in the sequel for notational convenience. The coefficients in the Galerkin approach are determined by the orthogonality of the residual

$$\begin{aligned} R_m(x, \xi) &= -\Delta \tilde{u}_m(x, \xi) - k(x, \xi)^2 \tilde{u}_m(x, \xi) - f(x) \\ &= -\sum_{i=0}^m \Delta v_i(x) \phi_i(\xi) - k(x, \xi)^2 \sum_{i=0}^m v_i(x) \phi_i(\xi) - f(x) \end{aligned}$$

to the subspace  $\text{span}\{\phi_0, \phi_1, \dots, \phi_m\}$ , i.e., by  $\langle R_m(x, \xi), \phi_j(\xi) \rangle = 0$  for  $j = 0, 1, \dots, m$  and each  $x \in Q$ . The latter is equivalent to

$$-\Delta v_j(x) - \sum_{i=0}^m \langle k(x, \xi)^2 \phi_i(\xi), \phi_j(\xi) \rangle v_i(x) = \langle 1, \phi_j(\xi) \rangle f(x) = \delta_{j,0} f(x) \tag{25}$$

for  $j = 0, 1, \dots, m$  in  $Q$ . Thus we obtain a system of PDEs for the unknown coefficient functions  $v_0, v_1, \dots, v_m$ . Define  $C(x) = [c_{ij}(x)] \in \mathbb{R}^{m+1, m+1}$  for  $x \in Q$  by

$$c_{ij}(x) = \langle k(x, \xi)^2 \phi_i(\xi), \phi_j(\xi) \rangle = \int_{\Xi} \phi_i(\xi) \phi_j(\xi) k(x, \xi)^2 \rho(\xi) \, d\xi, \quad i, j = 0, 1, \dots, m. \tag{26}$$

Since by assumption  $k(x, \xi) > 0$  for all  $x$  and  $\xi$ , the matrix  $C(x)$  is symmetric positive definite (as Gramian of an inner product with weight function  $k(x, \xi)^2 \rho(\xi)$ ). Setting

$$v(x) = [v_0(x) \ v_1(x) \ \cdots \ v_m(x)]^\top, \quad F(x) = [f(x) \ 0 \ \cdots \ 0]^\top, \tag{27}$$

we write the system of PDEs (25) as

$$-\Delta v(x) - C(x)v(x) = F(x) \quad \text{in } Q, \tag{28}$$

which is a larger deterministic system of linear PDEs. Still we require boundary conditions for the system (28).

The homogeneous Dirichlet boundary condition (10) implies  $v_j(x) = 0$  for  $x \in \partial Q$  and  $j = 0, 1, \dots, m$ , hence

$$v(x) = 0 \quad \text{on } \partial Q. \tag{29}$$

Inserting the Galerkin approximation (24) into the absorbing boundary conditions (11) yields the residual

$$R_m(x, \xi) = \sum_{i=0}^m (\partial_n v_i)(x) \phi_i(\xi) - ik(x, \xi) \sum_{i=0}^m v_i(x) \phi_i(\xi). \tag{30}$$

By the orthogonality  $\langle R_m(x, \xi), \phi_j(\xi) \rangle = 0$  in the Galerkin approach, we obtain

$$\partial_n v_j(x) - i \sum_{i=0}^m \langle k(x, \xi) \phi_i(\xi), \phi_j(\xi) \rangle v_i(x) = 0, \quad j = 0, 1, \dots, m. \tag{31}$$

The matrix  $B(x) = [b_{ij}(x)] \in \mathbb{R}^{m+1, m+1}$  with

$$b_{ij}(x) = \langle k(x, \xi) \phi_i(\xi), \phi_j(\xi) \rangle = \int_{\Xi} \phi_i(\xi) \phi_j(\xi) k(x, \xi) \rho(\xi) \, d\xi, \quad i, j = 0, 1, \dots, m, \tag{32}$$

is symmetric and positive definite (since  $k(x, \xi) > 0$  by assumption). The boundary condition (31) can be written with  $B(x)$  as

$$(\partial_n v)(x) - iB(x)v(x) = 0 \quad \text{on } \partial Q. \tag{33}$$

Discretizing the boundary value problem (28) with (29) or (33) with a second order FDM yields the same linear algebraic system as in Sect. 3.1.

### 4 Complex Shifted Laplace Preconditioner

Following the investigation in [9], we consider the Helmholtz equation (9) with a *complex shift* in the wavenumber

$$-\Delta u(x, \xi) - (1 + i\beta)k(x, \xi)^2 u(x, \xi) = f(x), \quad x \in Q, \tag{34}$$

with  $\beta \in \mathbb{R}$ , together with either homogeneous Dirichlet boundary conditions (10) or absorbing boundary conditions (11). We discretize this boundary value problem as described in Sect. 3.1. For  $\beta = 0$ , we have the matrix (23) in Theorem 4, and for  $\beta \in \mathbb{R}$  we obtain

$$M := M(\beta) := L - iB - (1 + i\beta)K = A - i\beta K, \tag{35}$$

since only the term with the wavenumber is multiplied by  $1 + i\beta$ . Motivated by [12, p. 1945], we call  $M$  a *complex shifted Laplace preconditioner* (CSL preconditioner).

For the deterministic Helmholtz equation, preconditioning with the CSL preconditioner is a widely studied and successful technique for solving the discretized Helmholtz equation; see, e.g., [1, 3, 6, 7, 25] and [9], as well as references therein. See also [5] for a survey and [18] for recent developments. In the deterministic case, the spectrum of the preconditioned matrix  $AM^{-1}$  lies in the disk (36), and the improved localization of the spectrum typically leads to a faster convergence of Krylov solvers. The CSL preconditioner  $M$  can be approximately inverted efficiently, for example, by multigrid techniques.

Here, we focus on locating the spectrum of the preconditioned matrix in the stochastic case, in analogy to [8, 9, 12] for the deterministic Helmholtz equation.

**Theorem 5** *Let the notation be as in Theorem 4, let  $\beta > 0$ , let  $A$  be the discretization (23) of the stochastic Helmholtz equation (9) and  $M$  be the discretization (35) of the shifted Helmholtz equation (34).*

1. *In the case of absorbing boundary conditions (11), the spectrum of the preconditioned matrix  $AM^{-1}$  is contained in the closed disk*

$$\mathcal{D} = \{z \in \mathbb{C} : |z - 1/2| \leq 1/2\}. \tag{36}$$

2. *In the case of homogeneous Dirichlet boundary conditions (10), the spectrum of the preconditioned matrix  $AM^{-1}$  lies on the circle*

$$\mathcal{C} = \{z \in \mathbb{C} : |z - 1/2| = 1/2\}. \tag{37}$$

**Proof** We begin with the case of absorbing boundary conditions. The proof closely follows [12, Sect. 3] with minor modifications. We have

$$A = L - iB - z_1K, \quad M = L - iB - z_2K \tag{38}$$

with  $z_1 = 1$  and  $z_2 = 1 + i\beta$  and where  $L, B, K$  are symmetric,  $K$  is positive definite and  $L, B$  are positive semidefinite; see Theorem 4. Then  $A$  and  $M$  are of the form in [12, Sect. 3], except for the opposite sign of  $B$ . The opposite sign affects the positive semidefiniteness, but not the overall strategy of the proof. Nevertheless, we give a full proof here.

Step 1: Observe first that  $AM^{-1}$  and  $M^{-1}A$  have the same spectrum, and that  $M^{-1}Ax = \sigma x$  is equivalent to the generalized eigenproblem  $Ax = \sigma Mx$ .

Step 2:  $x$  is an eigenvector of  $Ax = \sigma Mx$  if and only if  $(L - iB)x = \lambda Kx$ , which can be seen as follows:

$$(L - iB - z_1K)x = \sigma(L - iB - z_2K)x \Leftrightarrow (1 - \sigma)(L - iB)x = (z_1 - \sigma z_2)Kx. \tag{39}$$

For  $\sigma \neq 1$ , we obtain  $(L - iB)x = \lambda Kx$  with  $\lambda = (z_1 - \sigma z_2)/(1 - \sigma)$ . (Note that  $\sigma = 1$  is equivalent to  $z_1 = z_2$ , i.e., to  $A = M$ , which is excluded since  $\beta > 0$ .) Conversely, if  $(L - iB)x = \lambda Kx$ , then  $(L - iB - z_1K)x = (\lambda - z_1)Kx = \frac{\lambda - z_1}{\lambda - z_2}(L - iB - z_2K)x$  and  $\sigma = \frac{\lambda - z_1}{\lambda - z_2}$ , provided that  $\lambda \neq z_2$ . (Note that  $(L - iB)x = z_2Kx$ , i.e.,  $\lambda = z_2$ , implies that  $M$  is singular and thus not eligible as preconditioner.)

Step 3: Location of  $\lambda$  in the generalized eigenvalue problem  $(L - iB)x = \lambda Kx$ . Since  $K$  is real, symmetric positive definite, it has a Cholesky factorization  $K = UU^T = UU^H$  and the generalized eigenvalue problem is equivalent to

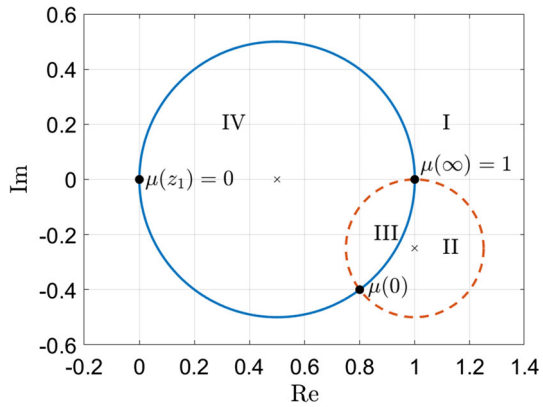
$$U^{-1}(L - iB)U^{-H}y = \lambda y, \tag{40}$$

where  $y = U^Hx$ . Multiplication of (40) by  $y^H$  and division by  $y^Hy$  yields

$$\lambda = \frac{y^H U^{-1} L U^{-H} y}{y^H y} - i \frac{y^H U^{-1} B U^{-H} y}{y^H y}. \tag{41}$$



**Fig. 1** Images under  $\mu$  of the real and imaginary axis (solid and dashed circles, respectively) and of the four quadrants; see Remark 6 and the proof of Theorem 5



This shows  $\text{Re}(\lambda) \geq 0$  and  $\text{Im}(\lambda) \leq 0$  since  $L$  and  $B$  are symmetric positive semidefinite.

Step 4: Estimate of the eigenvalues  $\sigma$  of  $M^{-1}A$ . Since it holds that  $z_1 \neq z_2$ ,

$$\mu(z) = \frac{z - z_1}{z - z_2} \tag{42}$$

is a Möbius transformation. By step 2,  $\sigma = \mu(\lambda)$  where  $\lambda$  is an eigenvalue of the generalized eigenvalue problem  $(L - iB)x = \lambda Kx$  which satisfies  $\text{Im}(\lambda) \leq 0$ . To determine  $\mu(\mathbb{R})$ , we compute

$$\mu(0) = \frac{z_1}{z_2} = \frac{1}{1 + i\beta} = \frac{1 - i\beta}{1 + \beta^2}, \quad \mu(z_1) = 0, \quad \mu(\infty) = 1 \tag{43}$$

and

$$\left| \mu(0) - \frac{1}{2} \right|^2 = \left| \frac{1}{1 + \beta^2} - \frac{1}{2} - i \frac{\beta}{1 + \beta^2} \right|^2 = \frac{(1 - \beta^2)^2}{4(1 + \beta^2)^2} + \frac{\beta^2}{(1 + \beta^2)^2} = \frac{1}{4}. \tag{44}$$

Hence  $\mu$  maps the real line onto the circle  $C$  in (37) (for any  $\beta \neq 0$ ). For  $\beta > 0$ , the lower half-plane is mapped by  $\mu$  onto the interior of  $C$  (for  $\beta < 0$  onto the exterior); see Fig. 1. This completes the proof in case of absorbing boundary conditions.

The proof in the case of Dirichlet boundary conditions is very similar. The only difference is in the location of the eigenvalues  $\lambda$  in step 3. Since  $L$  is symmetric positive definite and  $B = 0$ , (40) implies  $\lambda > 0$ , hence  $\sigma = \mu(\lambda)$  lies on the circle (37).  $\square$

**Remark 6** In the proof of Theorem 5, we additionally have  $\text{Re}(\lambda) \geq 0$ . Hence  $\sigma$  is located in the image of the (closed) fourth quadrant under  $\mu$  in (42). To determine this image, note that  $\mu$  maps the imaginary axis onto the circle

$$C_\beta = \{z \in \mathbb{C} : |z - (1 - i(\beta/2))| = |\beta|/2\}, \tag{45}$$

which intersects  $\mu(\mathbb{R}) = C$  perpendicularly in  $\mu(0)$  and  $\mu(\infty) = 1$ . Considering the orientations shows that  $\mu$  maps the right half-plane onto the exterior of  $C_\beta$ ; see Fig. 1. Thus the spectrum satisfies

$$\sigma(AM^{-1}) \subseteq \{z \in \mathbb{C} : |z - 1/2| \leq 1/2\} \setminus \{z \in \mathbb{C} : |z - (1 - i(\beta/2))| < |\beta|/2\}. \tag{46}$$

In case of Dirichlet boundary conditions, the eigenvalues of  $AM^{-1}$  lie on the arc of the circle  $C$  from  $\mu(0)$  to  $\mu(\infty) = 1$  that contains the origin.

This observation further tightens the inclusion set of  $\sigma(AM^{-1})$ , also in the case of a deterministic wavenumber. This tighter inclusion set is already visible in [8, Figs. 1, 2] and [9, Fig. 2.1] but we are not aware of a proof in the literature.

### 5 Mean Value Preconditioner

We consider the discretization from Sect. 3.1. Let  $S(\xi) \in \mathbb{K}^{n,n}$  be the coefficient matrix of a linear system resulting from a spatial discretization of the Helmholtz equation (1) including boundary conditions and wavenumber  $k(x, \xi)$ . We assume that  $S(\xi)$  is non-singular for almost all realizations  $\xi \in \Xi$ . Let  $\bar{\xi} \in \Xi$  be the expected value of the multidimensional random variable  $\xi$ . It holds that

$$S(\xi) = S(\bar{\xi}) + (S(\xi) - S(\bar{\xi})) =: S(\bar{\xi}) + \Delta S(\xi).$$

The stochastic Galerkin method applied to  $S(\xi)$  yields a matrix  $A \in \mathbb{K}^{(m+1)n, (m+1)n}$  as shown in Sect. 3.1. Furthermore, we define the constant matrix

$$\bar{A} = I_{m+1} \otimes S(\bar{\xi}). \tag{47}$$

This matrix allows for the construction

$$A = \bar{A} + (A - \bar{A}) =: \bar{A} + \Delta A. \tag{48}$$

We employ the Frobenius matrix norm  $\|\cdot\|_F$  in the following.

**Theorem 7** *Using the Frobenius norm, it holds that*

$$\|\bar{A}^{-1}A - I_{(m+1)n}\|_F \leq C_m \|S(\bar{\xi})^{-1}\|_F \|\Delta S(\xi)\|_F \Big\|_{\mathcal{L}^2(\Xi, \rho)} \tag{49}$$

with the constants

$$C_m = \sqrt{m+1} \left( \sum_{i,j=0}^m \|\phi_i(\xi)\phi_j(\xi)\|_{\mathcal{L}^2(\Xi, \rho)}^2 \right)^{\frac{1}{2}}$$

provided that the  $\mathcal{L}^2$ -norm of the matrix norm is finite.

**Proof** The definition (48) directly yields

$$\bar{A}^{-1}A - I_{(m+1)n} = I_{(m+1)n} + \bar{A}^{-1}\Delta A - I_{(m+1)n} = \bar{A}^{-1}\Delta A.$$

We obtain  $\|\bar{A}^{-1}\Delta A\|_F \leq \|\bar{A}^{-1}\|_F \|\Delta A\|_F$ . The properties of the Kronecker product and (47) imply  $\|\bar{A}^{-1}\|_F^2 = (m+1)\|S(\bar{\xi})^{-1}\|_F^2$ . We estimate  $\|\Delta A\|_F$  using the Cauchy-Schwarz inequality with respect to the inner product (5)

$$\begin{aligned} \|\Delta A\|_F^2 &= \sum_{i,j=0}^m \sum_{\mu,v=1}^n |(\phi_i(\xi)\phi_j(\xi), \Delta S_{\mu,v}(\xi))|^2 \\ &\leq \sum_{i,j=0}^m \sum_{\mu,v=1}^n \|\phi_i(\xi)\phi_j(\xi)\|_{\mathcal{L}^2(\Xi, \rho)}^2 \|\Delta S_{\mu,v}(\xi)\|_{\mathcal{L}^2(\Xi, \rho)}^2 \\ &= \left( \sum_{i,j=0}^m \|\phi_i(\xi)\phi_j(\xi)\|_{\mathcal{L}^2(\Xi, \rho)}^2 \right) \|\Delta S(\xi)\|_F^2 \Big\|_{\mathcal{L}^2(\Xi, \rho)}. \end{aligned}$$

In the last step, we used that the square of an  $\mathcal{L}^2$ -norm is an integral and thus summation (with respect to  $\mu, \nu$ ) and integration can be interchanged. Applying the square root to the above estimate yields the statement (49).  $\square$

**Remark 8** Rough estimates are used in the proof of Theorem 7. Thus the true matrix norms of  $\bar{A}^{-1}A - I_{(m+1)n}$  are often much smaller than the upper bounds in (49).

**Remark 9** If the random variable  $\Delta S(\xi)$  is essentially bounded, then it follows that

$$\|\Delta S(\xi)\|_F \Big|_{\mathcal{L}^2(\Xi, \rho)} \leq \sup_{\xi \in \Xi \setminus \Upsilon} \|\Delta S(\xi)\|_F < \infty$$

with a set  $\Upsilon \subseteq \Xi$  of measure zero due to the normalization  $\|1\|_{\mathcal{L}^2(\Xi, \rho)} = 1$ .

**Remark 10** The bound of Theorem 7 also holds true for the Frobenius norm of  $A\bar{A}^{-1} - I_{(m+1)n}$ .

Theorem 7 together with Remark 8 demonstrate that the matrix  $\bar{A}$  is a good preconditioner for solving linear systems with coefficient matrix  $A$ . In this context,  $\bar{A}$  is called the *mean value preconditioner*, as in [30] for the multi-element method. When  $\bar{A}$  is used as a preconditioner (left-hand or right-hand), linear systems with coefficient matrix  $\bar{A}$  have to be solved. The matrix  $\bar{A}$  from (47) is block-diagonal with  $m + 1$  identical blocks in this application. Thus just a single  $LU$ -decomposition of the matrix  $S(\xi)$  is required. Many linear systems with different right-hand sides are solved using this  $LU$ -decomposition in an iterative method like GMRES, for example.

**Theorem 11** Let  $S(\xi) = S_0 + \theta T(\xi)$  with a non-singular constant matrix  $S_0$ , a matrix  $T = [t_{\mu, \nu}]_{\mu, \nu}$  depending on a random variable  $\xi$  with components  $t_{\mu, \nu} \in \mathcal{L}^2(\Xi, \rho)$  and a real parameter  $\theta > 0$ . Using  $A_0 = I_{m+1} \otimes S_0$ , the Frobenius norm exhibits the asymptotic behavior

$$\|A_0^{-1}A - I_{(m+1)n}\|_F = O(\theta). \tag{50}$$

**Proof** Since the entries of  $T(\xi)$  are assumed to be square-integrable, also the expected values are finite. Let  $\bar{T}$  be the constant matrix containing the expected values of  $T(\xi)$ . We apply the decomposition

$$S(\xi) = (S_0 + \theta\bar{T}) + \theta(T(\xi) - \bar{T}).$$

The matrix  $S_0 + \theta\bar{T}$  is non-singular for sufficiently small  $\theta$ . Moreover, we obtain the relation  $(S_0 + \theta\bar{T})^{-1} = S_0^{-1} + O(\theta)$ . Theorem 7 yields

$$\|\bar{A}^{-1}A - I_{(m+1)n}\|_F \leq C_m \|(S_0 + \theta\bar{T})^{-1}\|_F \|\theta(T - \bar{T})\|_F \Big|_{\mathcal{L}^2(\Xi, \rho)}$$

with  $\bar{A} = I_{m+1} \otimes (S_0 + \theta\bar{T})$ . It holds that  $\bar{A} = A_0 + O(\theta)$  and thus  $\bar{A}^{-1} = A_0^{-1} + O(\theta)$ . We conclude

$$\|A_0^{-1}A - I_{(m+1)n}\|_F \leq \left( C_m \left( \|S_0^{-1}\|_F + O(\theta) \right) \theta \|T - \bar{T}\|_F \Big|_{\mathcal{L}^2(\Xi, \rho)} \right) + O(\theta) = O(\theta),$$

which confirms (50).  $\square$

An important case of Theorem 11 is  $\bar{T} = 0$ , i.e., these expected values are zero. Then  $A_0 = \bar{A}$  is the mean value preconditioner.

**Corollary 12** *Under the assumptions of Theorem 7, the Frobenius norm satisfies the estimate*

$$\|\bar{A}^{-1}A - I_{(m+1)n}\|_F < 1 \tag{51}$$

for all sufficiently small  $\Delta S$ .

Likewise, the Frobenius norm using  $A_0$  instead of  $\bar{A}$  is smaller than one if the parameter  $\theta$  is sufficiently small in the context of Theorem 11.

A stationary iterative scheme for solving a linear system  $Ax = b$  reads as

$$Bx^{(i+1)} = b - (A - B)x^{(i)} \quad \text{for } i = 0, 1, 2, \dots \tag{52}$$

with a non-singular matrix  $B$  which should approximate  $A$ , see [29, p. 621]. In each iteration step, we have to solve a linear system with coefficient matrix  $B$ . The property (51) is sufficient for the global convergence of the iteration (52) using  $B = \bar{A}$ . The computational costs of an iteration step are much less than the steps in GMRES using  $\bar{A}$  as preconditioner, because the construction of Krylov subspaces is avoided. In practice, we do not know if  $\Delta S$  is sufficiently small such that the bound (51) is guaranteed. Nevertheless, it is worth to try this stationary iteration, as we will observe in Sect. 7.

## 6 Numerical Experiments in 1D

Our model problem in one space dimension is the stochastic Helmholtz equation (9) on  $Q = ]0, 1[$  with absorbing boundary conditions. The right-hand side is the point source  $f(x) = \delta(x - \frac{1}{2})$ , similarly to, e.g., [9, 12, 19, 28], where the right-hand side is a (possibly scaled) point source. We consider a random wavenumber  $k(x, \xi) = k(\xi)$  constant in space, which is uniformly distributed in some interval  $[k_{\min}, k_{\max}]$  with  $0 < k_{\min} < k_{\max}$ . Equivalently, we define

$$k(\xi) = (1 + \theta\xi)\bar{k} \tag{53}$$

with a random variable  $\xi$  that is uniformly distributed in  $[-1, 1]$ , a mean value  $\bar{k}$ , and a real parameter  $\theta \in ]0, 1[$ . It follows that  $k_{\min} = (1 - \theta)\bar{k}$  and  $k_{\max} = (1 + \theta)\bar{k}$ .

In our numerical experiments in one and two spatial dimensions, we compute the mesh-size  $h = \frac{1}{q+1}$  in the FD discretization by

$$q = 2^\ell - 1, \quad \text{where } \ell = \max \left\{ \left\lceil \log_2 \left( \frac{15}{2\pi} k_{\max} \right) \right\rceil, 1 \right\}. \tag{54}$$

Then the relation  $\frac{2\pi}{kh} \approx \text{constant}$ , advocated in [17, Sect. 4.4.1], is satisfied. Indeed, the estimate  $x \leq \lceil x \rceil \leq x + 1$  for  $x \in \mathbb{R}$  implies  $\frac{15k}{2\pi} \leq q + 1 \leq 2\frac{15k}{2\pi}$  for large  $k$ . In particular,  $q$  grows linearly with  $k$  and thus the size of the matrix  $A$  grows with  $k$ ; see, e.g., Fig. 3. Our choice for  $q$  can be adapted for a future use of a multigrid method (as in [9]).

Discretizing the model problem yields the linear algebraic system of equations

$$Ax = b \tag{55}$$

in Theorem 4. This one-dimensional problem can be solved by a direct method, since the computational work is not too large. Nevertheless we also consider its solution with the GMRES method [27] and investigate the application of CSL and mean value preconditioners introduced in Sects. 4 and 5, respectively.

**Table 1** Different discretizations of the Helmholtz and shifted Helmholtz equation, with or without uncertainties

$A$	Discretized Helmholtz equation with uncertainties ( $\theta > 0$ )
$A_0$	Discretized Helmholtz equation without uncertainties ( $\theta = 0$ )
$M$	Discretized shifted Helmholtz equation with uncertainties ( $\theta > 0$ )
$M_0$	Discretized shifted Helmholtz equation without uncertainties ( $\theta = 0$ )

The matrix  $A$  in (55) has the form

$$A = I_{m+1} \otimes T - i[B_{ij}] - [C_{ij}]. \tag{56}$$

If needed, we write  $A_\theta$  to indicate the dependence of  $A$  on  $\theta$ , and in particular  $A_0$  for  $\theta = 0$ , which corresponds to the mean value preconditioner. Since the wavenumber in (53) is constant in space, the matrices  $[B_{ij}]$  and  $[C_{ij}]$  have the form

$$[B_{ij}] = [(k(\xi)\phi_j(\xi), \phi_i(\xi))]_{ij} \otimes D_1, \quad D_1 = \frac{1}{h} \text{diag}(1, 0, \dots, 0, 1), \tag{57}$$

$$[C_{ij}] = [(k(\xi)^2\phi_j(\xi), \phi_i(\xi))]_{ij} \otimes D_2, \quad D_2 = \text{diag}\left(\frac{1}{2}, 1, \dots, 1, \frac{1}{2}\right), \tag{58}$$

and, moreover,  $B_{ij} = 0$  for  $|i - j| > 1$  and  $C_{ij} = 0$  for  $|i - j| > 2$ . In other words, the matrices  $[(k(\xi)\phi_j(\xi), \phi_i(\xi))]_{ij}$  and  $[(k(\xi)^2\phi_j(\xi), \phi_i(\xi))]_{ij}$  are tridiagonal and pentadiagonal, respectively, due to the orthogonality properties of the polynomials  $\phi_i(\xi), i = 0, 1, \dots$

**Remark 13** In the deterministic case  $k(\xi) = \bar{k}$  in (53), i.e.,  $\theta = 0$ , the matrices  $[B_{ij}] = \bar{k}I_{m+1} \otimes D_1$  and  $[C_{ij}] = \bar{k}^2I_{m+1} \otimes D_2$  are diagonal, and

$$A_0 = I_{m+1} \otimes (T - i\bar{k}D_1 - \bar{k}^2D_2) = I_{m+1} \otimes S(0) \tag{59}$$

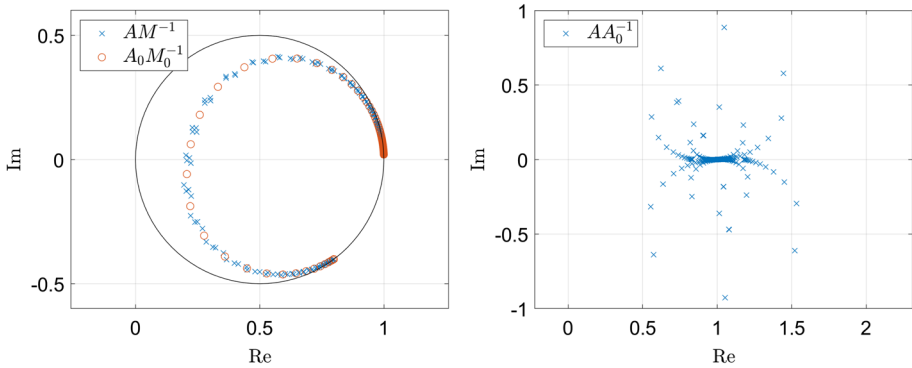
with  $S$  from (14). This shows that the mean value preconditioner  $A_0$  is block-diagonal with  $m + 1$  identical diagonal blocks. The latter are the FD-discretization of the deterministic Helmholtz equation with wavenumber  $\bar{k}$  (associated to  $\xi = 0$ ).

If not specified otherwise, we use  $m = 3$  in the stochastic Galerkin method and  $\theta = 0.1$  in (53). Finally, we also consider the shifted Helmholtz equation (34) with shift  $\beta = \frac{1}{2}$  and denote the CSL preconditioner by  $M = M(\frac{1}{2})$ , see (35). As for  $A$ , we write  $M_\theta$  if we wish to emphasize the dependence on  $\theta$ . Table 1 summarizes the four kinds of matrices involved in our computations in Sect. 6 and Sect. 7.

The numerical experiments have been performed in the software package MATLAB R2020b on an i7-7500U @ 2.70GHz CPU with 16 GB RAM.

### 6.1 Spectra

By Theorem 5, the eigenvalues of the CSL preconditioned matrix  $AM^{-1}$  lie in the closed disk (36). This is illustrated in the left panel of Fig. 2, which displays the spectra of  $AM^{-1}$  with  $\theta = 0.1$  and of  $A_0M_0^{-1}$ , i.e., with  $\theta = 0$  (without uncertainties); see also Table 1 for an overview of the different matrices. Each eigenvalue of  $A_0M_0^{-1}$  is  $(m + 1)$ -fold, since  $A_0 = I_{m+1} \otimes S(0)$  is block-diagonal with identical diagonal blocks, see Remark 13, and similarly for  $M_0$ . For  $\theta \neq 0$ , the matrix  $AM^{-1}$  is not block-diagonal, and  $AM^{-1}$  has clusters



**Fig. 2** Left: Spectrum of  $AM^{-1}$  for  $\bar{k} = 50, m = 3, \theta = 0.1$  (crosses) and  $\theta = 0$  (circles). The large solid circle illustrates (37). Right: Spectrum of  $AA_0^{-1}$

of  $m + 1$  eigenvalues close to each  $(m + 1)$ -fold eigenvalue of  $A_0M_0^{-1}$ . This can be observed in the figure with  $m + 1 = 4$ . The right panel in Fig. 2 displays the spectrum of  $AA_0^{-1}$  with the mean value preconditioner  $A_0$ . The eigenvalues are clustered at 1, which suggests a fast convergence of GMRES. If the eigenvalues satisfy  $|\lambda - 1| < 1$ , then the stationary method (52) with  $B = A_0$  converges.

### 6.2 Condition Numbers

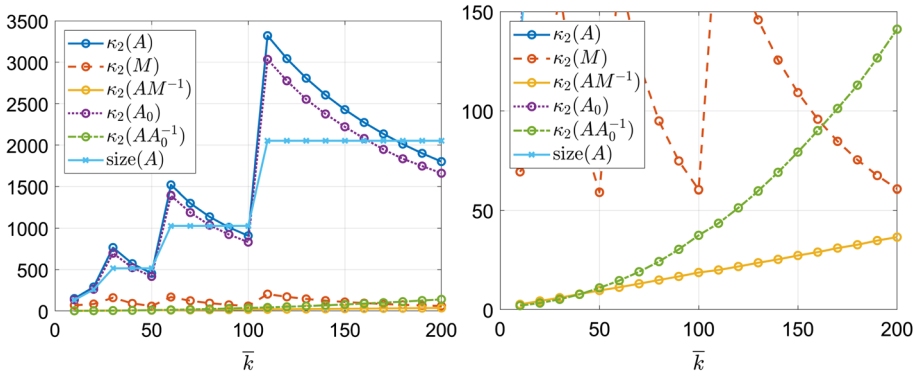
Recall that  $A, M, A_0$  and  $M_0$  denote the discretizations of the (shifted) Helmholtz equation with or without uncertainties; see Table 1. Figure 3 displays the 2-norm condition numbers of  $A, M, AM^{-1}, A_0$  and  $AA_0^{-1}$  as functions of  $\bar{k}$  (with  $\theta = 0.1$ ). Clearly, the condition numbers of  $M$  and  $AM^{-1}$  are much smaller than the condition number of  $A$ . A smaller condition number is beneficial when solving linear systems, since then a small relative residual of an approximate solution implies a small relative error of the approximate solution. (This follows from the following well-known residual-based forward error bound: Let  $A \in \mathbb{C}^{n,n}$  be non-singular,  $b \in \mathbb{C}^n \setminus \{0\}, x = A^{-1}b$ , and let  $\hat{x}$  be an approximate solution of the linear algebraic system  $Ax = b$ . Then the relative error satisfies  $\|x - \hat{x}\|_2 / \|x\|_2 \leq \kappa_2(A) \|r\|_2 / \|b\|_2$  with the residual  $r = b - A\hat{x}$ .) In this example,  $\kappa_2(M) \leq 205$  for all  $\bar{k}$ , which is very moderate, and  $\kappa_2(AM^{-1})$  grows linearly in  $\bar{k}$  from 2.6485 when  $\bar{k} = 10$  to only 36.5190 when  $\bar{k} = 200$ . In contrast,  $\kappa_2(A)$  is roughly 50 to 160 times larger than  $\kappa_2(AM^{-1})$ . The observed spikes of  $\kappa_2(A)$  occur when more discretization points are used which leads to a larger size of  $A$ , compare the curve of `size(A)`. The condition number of the mean value preconditioned matrix  $AA_0^{-1}$  is also moderate, growing from 2 to 141, which is beneficial for solving the preconditioned linear system, while  $\kappa_2(A_0)$  is of the order of  $\kappa_2(A)$ .

### 6.3 GMRES

We solve the unpreconditioned system (55) and the right and left preconditioned systems

$$AM^{-1}y = b, \quad x = M^{-1}y, \quad \text{and} \quad M^{-1}Ax = M^{-1}b \tag{60}$$

with full GMRES (no restarts), using MATLAB’s built-in `gmres` command. The residual in the  $i$ th step is  $r^{(i)} = b - Ax^{(i)}$  for unpreconditioned and right preconditioned GMRES, and



**Fig. 3** 2-norm condition numbers as functions of  $\bar{k}$  (left) and zoom-in (right). The matrices  $A$  and  $A_0$  arise from the Helmholtz equation with and without uncertainties, respectively, and  $M$  and  $M_0$  from the shifted Helmholtz equation; see also Table 1

$M^{-1}r^{(i)}$  for left preconditioned GMRES. The stopping criterion is that the relative residual norm is less than  $10^{-12}$ , i.e.,  $\|r^{(i)}\|_2/\|r^{(0)}\|_2 < 10^{-12}$  for unpreconditioned and right preconditioned GMRES, and  $\|M^{-1}r^{(i)}\|_2/\|M^{-1}r^{(0)}\|_2 < 10^{-12}$  for left preconditioned GMRES. In particular, the stopping criterion is in general different for left and right preconditioning; see [26, Ch. 9.3] for a detailed discussion. We will consider the following three preconditioners:

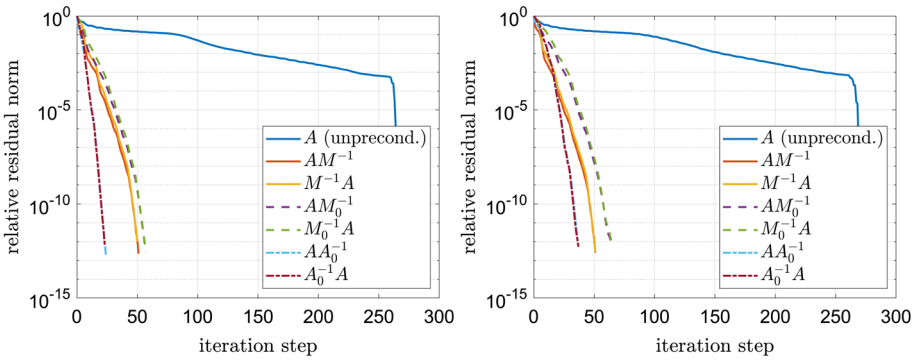
1. the CSL preconditioner  $M$ ,
2. the mean value preconditioner  $A_0$ ,
3. the mean value CSL preconditioner  $M_0$ .

In preconditioned GMRES, we need to solve linear systems with the preconditioner, for which we use an  $LU$ -decomposition. In one spatial dimension, this is not competitive with the direct solution (see the end of Sect. 6.3), but in two spatial dimensions the block structure of the preconditioners  $A_0$  and  $M_0$  leads to a competitive method. In MATLAB, the  $LU$ -decomposition of the sparse matrix  $M$  calls the associated routine from UMFPACK; see [4]. The decomposition has the form

$$PMQ = LU \tag{61}$$

with a lower triangular matrix  $L$ , upper triangular matrix  $U$ , and two permutation matrices  $P, Q$ . By Remark 13,  $A_0 = I_{m+1} \otimes S(0) \in \mathbb{K}^{(m+1)n, (m+1)n}$  is block-diagonal with equal diagonal blocks so that, for fixed  $\bar{k}$ , only a single  $LU$ -decomposition of  $S(0) \in \mathbb{K}^{n, n}$  is necessary to compute  $A_0^{-1}x$  for any vector  $x \in \mathbb{K}^{(m+1)n}$ . In our implementation, we partition and reshape  $x$  so that only one linear system with  $S(0)$  (using the  $LU$ -factors) is solved. The preconditioner  $M_0$  is implemented in the same way.

In a first experiment, we fix  $\bar{k} = 50$ ,  $\theta = 0.1$  and  $m = 3$ . Solving the unpreconditioned system (55) with GMRES suffers from a long delay of convergence; see Fig. 4. In contrast, all three preconditioners  $M, M_0$ , and  $A_0$  lead to a significant decrease in the number of iteration steps from about 250 to 50 for  $M$  and  $M_0$  (factor 5), and to about 25 for  $A_0$  (factor 10); see Fig. 4 (left panel). The differences between computed solutions are very small:  $\|x - x'\|_\infty \leq 1.7 \cdot 10^{-14}$  (and typically of order  $10^{-15}$ ), where  $x$  is the computed direct solution and  $x'$  is a solution computed with GMRES (unpreconditioned or with one of the preconditioners). Left and right preconditioning lead to very similar relative residual norms and timings for each preconditioner. A heuristic explanation why  $A_0$  performs better than  $M$  and  $M_0$ , is that  $A$  is closer to  $A_0$  than to  $M$  or  $M_0$ . Indeed, we have  $\|A - A_0\|_\infty < \|A - M\|_\infty < \|A - M_0\|_\infty$



**Fig. 4** Relative residual norms when solving (55) with GMRES with various preconditioners,  $m = 3$ ,  $\bar{k} = 50$ , and  $\theta = 0.1$  (left) or  $\theta = 0.2$  (right). The matrices  $A$  and  $A_0$  arise from the Helmholtz equation with and without uncertainties, respectively, and  $M$  and  $M_0$  from the shifted Helmholtz equation; see also Table 1

in this example. Repeating this experiment with  $\theta = 0.2$  leads to very similar results, see Fig. 4, so we focus on  $\theta = 0.1$ .

In a second experiment, we let  $\bar{k}$  vary while  $\theta = 0.1$  and  $m = 3$  are fixed. Figure 5 displays the number of GMRES iteration steps (top) and the computation time (bottom), measured as wall clock time with MATLAB’s `tic toc` command, as functions of  $\bar{k}$ . For small  $\bar{k} \in [10, 50]$ , the difference between unpreconditioned and preconditioned GMRES is not so pronounced, since the linear systems are rather small. For  $60 \leq \bar{k} \leq 200$ , the three preconditioners significantly reduce the number of iteration steps and the computation time compared to unpreconditioned GMRES. The number of iteration steps is reduced to 8–15% of the number of iteration steps in unpreconditioned GMRES when using  $M$ , to 9–16% when using  $M_0$  and to only 3–6% when using  $A_0$  as preconditioner. GMRES preconditioned with  $M$  or  $M_0$  needs only 1–4% of the computation time of unpreconditioned GMRES, and the computation time of GMRES preconditioned with  $A_0$  is reduced to 0.5–1.1% of the computation time of unpreconditioned GMRES. The mean value preconditioner  $A_0$  leads to the smallest number of GMRES iteration steps and computation time, which is likely due to the fact that  $A$  is closer to  $A_0$  than to  $M$  or  $M_0$ . Note, however, that the condition number of  $A_0$  (and  $A$ ) is much larger than that of  $M$  and  $M_0$ . For  $\bar{k} = 150$ , we have (rounded to the nearest integer)  $\kappa_2(A) = 2428$ ,  $\kappa_2(A_0) = 2220$ ,  $\kappa_2(M) = 109$ ,  $\kappa_2(M_0) = 91$ ; see also Fig. 3. Thus, if accuracy is an issue, it is preferable to work with the CSL preconditioners  $M$  or  $M_0$ .

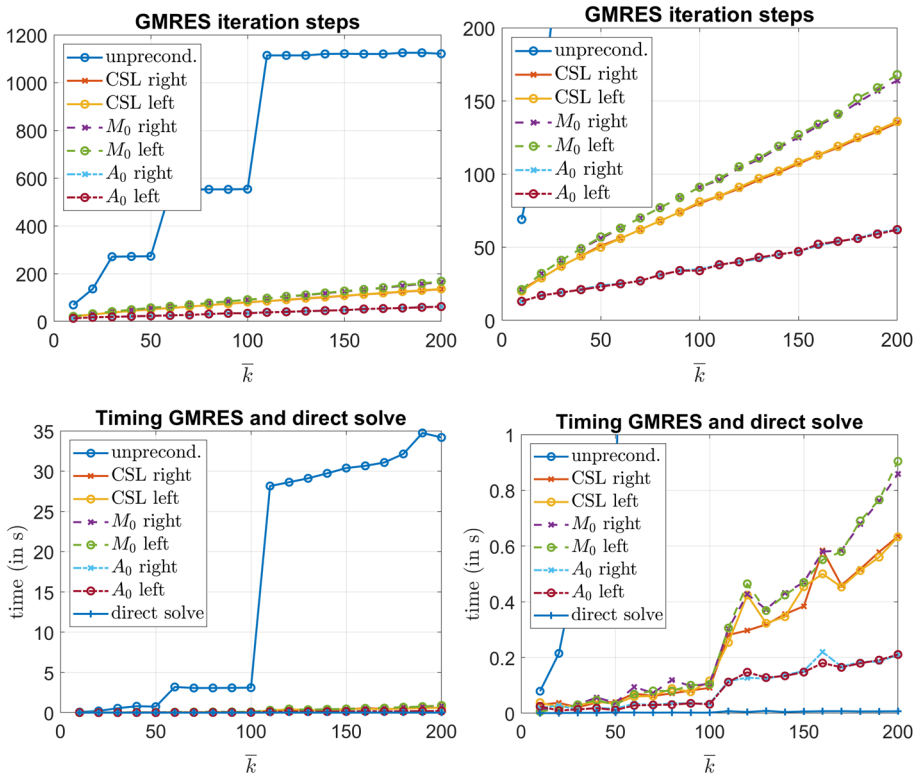
Finally, we note that the direct solution using the ‘backslash’ command with a sparse matrix in MATLAB calls an efficient algorithm from UMFPACK; see [4]. In the above test example, solving the linear system (55) by GMRES (with or without preconditioner) is not competitive with this direct solution, as it is much faster; see the bottom right panel in Fig. 5.

### 6.4 Solutions

Figure 6 displays the real and imaginary parts of the computed coefficients  $v_0, v_1, v_2, v_3$  in the Galerkin approximation for  $\bar{k} = 50$  and  $\theta = 0.1$  in (53). We recognize an effect of the point source at  $x = \frac{1}{2}$  in the real part of  $v_0$ .

We compute the solution for total polynomial degree  $m = 100$ . Figure 7 shows  $\|v_i\|_\infty$  as a function of the polynomial degree  $i$ . In the left panel,  $\theta = 0.1$  is fixed and  $\bar{k}$  varies,





**Fig. 5** Number of GMRES iteration steps (top) and computation time in seconds (bottom) as functions of  $\bar{k}$  for different left and right preconditioners (see Table 1) with fixed  $\theta = 0.1$  and  $m = 3$ . The right panels are zoom-ins

while in the right panel  $\theta$  varies and  $\bar{k} = 100$  is fixed. We observe an exponential decay of the coefficients in all cases, which is related to the exponential convergence of the PC expansion (7). Larger wavenumbers and larger values of  $\theta$  lead to a slower decay of the maximum-norm of the coefficients. The effect of larger  $\theta$  on the convergence/decay is more pronounced, compare, for example, the curve for  $(\bar{k}, \theta) = (150, 0.1)$  in the left panel with the curve for  $(100, 0.5)$  in the right panel.

Next, we vary  $m$  (the maximal degree of the polynomials in the stochastic Galerkin method) and denote by  $x_m$  the solution of (55), which consists of a discretization of the coefficients  $v_{0,m}, \dots, v_{m,m}$  in a Galerkin approximation (24) of the solution  $u$  of the Helmholtz equation; see also Remark 1. The convergence of the stochastic Galerkin method is illustrated by the exponential decay of the norms  $\|x_m - x_{m+1}\|_2$  in Fig. 8.

### 7 Numerical Experiments in 2D

Next, we consider the stochastic Helmholtz equation in two space dimensions, where the wavenumber depends on random variables and on the spatial variables.

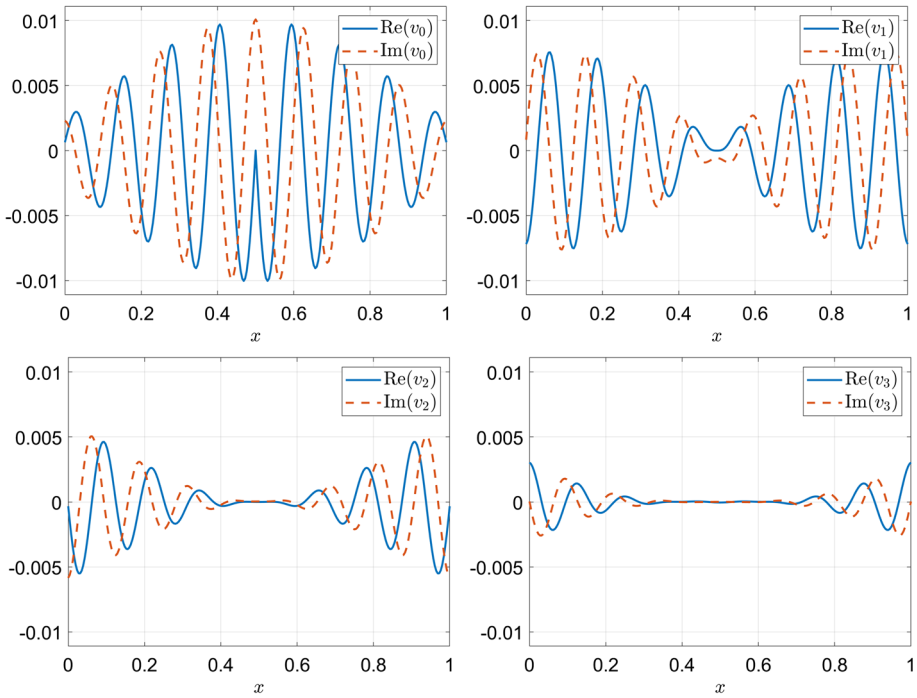


Fig. 6 Plots of the coefficients  $v_0, v_1, v_2, v_3$  for  $\bar{k} = 50$  and  $\theta = 0.1$

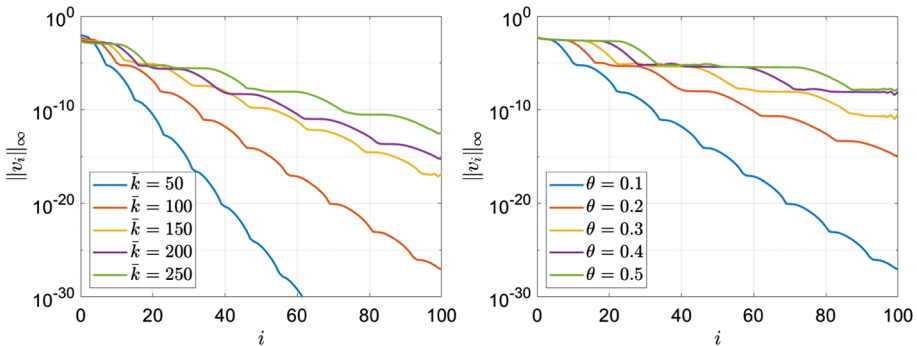
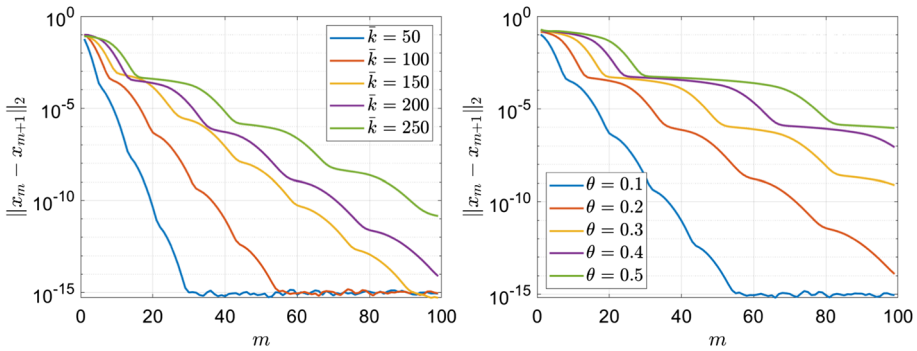


Fig. 7 Maximum-norms  $\|v_i\|_\infty$  as a function of  $i = \deg(\phi_i)$ . Left: For fixed  $\theta = 0.1$  and different values of  $\bar{k}$ . Right: For fixed  $\bar{k} = 100$  and different values of  $\theta$

### 7.1 Modeling

We consider the stochastic Helmholtz equation (9) in  $Q = ]0, 1[^2$  with absorbing boundary conditions (11), the point source  $f(x, y) = \delta((x, y) - (\frac{1}{2}, \frac{1}{2}))$  as right-hand side, and space-dependent random wavenumber

$$k(x, y, \xi_1, \xi_2, \xi_3) = \begin{cases} (1 + \theta\xi_1)k_1, & y \leq 0.2 + 0.1x, \\ (1 + \theta\xi_2)k_2, & 0.2 + 0.1x < y < 0.6 - 0.2x, \\ (1 + \theta\xi_3)k_3, & 0.6 - 0.2x \leq y. \end{cases} \quad (62)$$



**Fig. 8** Norms  $\|x_m - x_{m+1}\|_2$  as a function of the maximal degree  $m$  in the stochastic Galerkin method. Left: For fixed  $\theta = 0.1$  and different values of  $\bar{k}$ . Right: For fixed  $\bar{k} = 100$  and different values of  $\theta$

**Table 2** For different total degrees  $r$ , number of basis polynomials (n.basis), size of the matrix  $A$ , number of non-zero elements in  $A$  (nnz), and time to generate  $A$

$r$	n.basis	Size of $A$	nnz	Time (s)
0	1	16641	82689	0.3264
1	4	66564	364038	0.6592
2	10	166410	993300	1.4090
3	20	332820	2119728	3.3421
4	35	582435	3892575	10.3049
5	56	931896	6461094	32.8872
6	84	1397844	9974538	101.1674
7	120	1996920	14582160	291.5063
8	165	2745765	20433213	802.0748

on the wedge-shaped domain from [19, p. 146]; similar domains have been examined in [6, Sect. 6.3] and [8, Sect. 4.4]. The modeling (62) can also be written in the form (4) using spatial indicator functions. The random variables  $\xi_1, \xi_2, \xi_3$  are independent and uniformly distributed in  $[-1, 1]$ . The mean value of the wavenumber is

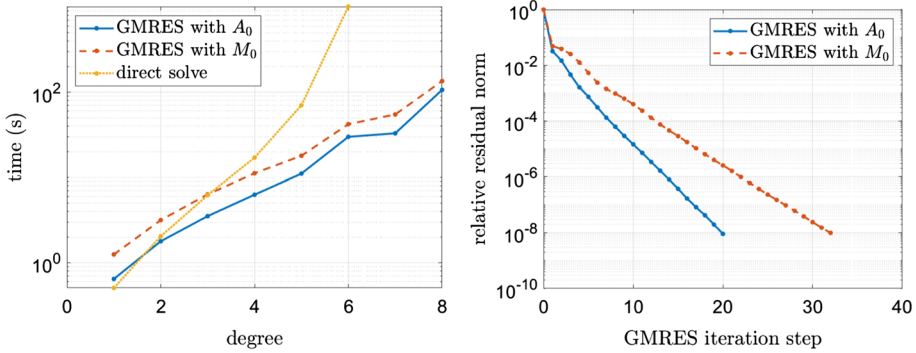
$$\bar{k}(x, y) = \begin{cases} k_1, & y \leq 0.2 + 0.1x, \\ k_2, & 0.2 + 0.1x < y < 0.6 - 0.2x, \\ k_3, & 0.6 - 0.2x \leq y. \end{cases} \tag{63}$$

We discretize the boundary value problem as described in Sect. 3.1 and obtain the linear algebraic system  $Ax = b$  in Theorem 4. The number of polynomials in the three random variables  $\xi_1, \xi_2, \xi_3$  with total degree at most  $r$  is, see (6),

$$m + 1 = \frac{(r+3)!}{r!3!} = \frac{1}{6}(r + 1)(r + 2)(r + 3). \tag{64}$$

Table 2 includes the number of basis polynomials for degrees  $r = 0, 1, \dots, 8$ .

Let  $k_1 = 30, k_2 = 15, k_3 = 20$ , and  $\theta = 0.1$ . Table 2 shows the size of  $A$  and the time (in seconds) for constructing the matrix  $A$  for polynomial degrees up to  $r = 0, 1, \dots, 8$  in the stochastic Galerkin method. As a function of  $r$ , the computation time when solving  $Ax = b$  directly in MATLAB grows much faster than for the iterative solvers; see Fig. 9 (left panel).



**Fig. 9** Solving the linear system directly and with GMRES preconditioned by  $A_0$  and  $M_0$ ; see Sect. 7. Left: Computation time (in seconds) as a function of the polynomial degree  $r$  in the stochastic Galerkin method. Right: Relative residual norms in preconditioned GMRES for polynomial degree  $r = 8$

**Table 3** Total operation count until convergence of preconditioned GMRES with the mean value preconditioner  $A_0 = I_{m+1} \otimes S_0$  for polynomial degrees  $r = 1, \dots, 8$ , as described in Sect. 7.2

polynomial degree $r$	1	2	3	4	5	6	7	8
$LU$ -factorization of sparse matrix $S_0 \in \mathbb{K}^{n,n}$	1	1	1	1	1	1	1	1
Matrix-vector products with $A \in \mathbb{K}^{(m+1)n, (m+1)n}$	13	17	18	19	20	20	20	20
Solves of $S_0 X = B \in \mathbb{K}^{n, m+1}$ with $LU$ -factors of $S_0$	14	18	19	20	21	21	21	21

### 7.2 Iterative Solvers

We solve the linear algebraic system  $Ax = b$  with GMRES using the mean value preconditioner  $A_0 = I_{m+1} \otimes S_0$  from (47) as right preconditioner, that is, we solve

$$AA_0^{-1}y = b, \quad A_0x = y. \tag{65}$$

Here  $S_0$  denotes the FD discretization of the Helmholtz equation with absorbing boundary conditions and deterministic wavenumber (63). The solution of linear systems with the preconditioner  $A_0$  is implemented as described in Sect. 6.3. We solve (65) with full GMRES (no restarts),  $\text{tol}=1\text{e-}8$  and  $\text{maxit}=200$  for polynomial degrees up to  $r = 1, \dots, 8$  in the stochastic Galerkin method. In contrast to the experiments in 1D in Sect. 6, preconditioned GMRES is significantly faster than the direct solution with MATLAB’s ‘backslash’ command; see Fig. 9 (left panel). The computation times for preconditioned GMRES include the computation of the  $LU$ -decomposition of a diagonal block of  $A_0$ . Furthermore, the relative residual norms in GMRES for polynomial degree  $r = 8$  are shown in Fig. 9 (right panel). The mean value CSL preconditioner  $M_0$  has a similar block-diagonal structure to  $A_0$ , which we denote again by  $M_0 = I_{m+1} \otimes S_0$ , and performs similarly well; see Fig. 9. Here  $S_0$  denotes the FD discretization of the Helmholtz equation with absorbing boundary conditions and deterministic wavenumber (63) with a complex shift. Tables 3 and 4 contain the number of operations performed until preconditioned GMRES converges to the prescribed tolerance.

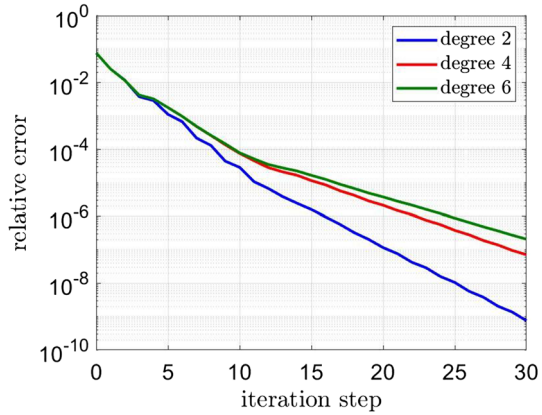
Alternatively to GMRES or a direct solution of the linear system, we also investigate the stationary iteration (52) with  $B = A_0$ , i.e.,

$$A_0x^{(i+1)} = b - (A - A_0)x^{(i)} \quad \text{for } i = 0, 1, 2, \dots \tag{66}$$

**Table 4** Total operation count until convergence of preconditioned GMRES with the mean value CSL preconditioner  $M_0 = I_{m+1} \otimes S_0$  for polynomial degrees  $r = 1, \dots, 8$ , as described in Sect. 7.2

polynomial degree $r$	1	2	3	4	5	6	7	8
$LU$ -factorization of sparse matrix $S_0 \in \mathbb{K}^{n,n}$	1	1	1	1	1	1	1	1
Matrix-vector products with $A \in \mathbb{K}^{(m+1)n,(m+1)n}$	29	31	32	32	32	32	32	32
Solves of $S_0 X = B \in \mathbb{K}^{n,m+1}$ with $LU$ -factors of $S_0$	30	32	33	33	33	33	33	33

**Fig. 10** Relative error norms in the stationary iteration (66) for different polynomial degrees  $r$  in the stochastic Galerkin method



We take the starting vector  $x^{(0)} = A_0^{-1}b$ . Linear systems with the matrix  $A_0$  are solved as described above. For  $\theta = 0.1$ , this iteration converges. Figure 10 displays the relative error norms in the maximum-norm for polynomial degrees  $r = 2, 4, 6$ , where we take the direct solution as the ‘exact’ solution. The slower convergence for larger degree  $r$  in the stochastic Galerkin method is expected, since the matrix size also grows causing higher condition numbers. For  $\theta = 0.2$ , the stationary iteration diverges. This behavior is in agreement to Theorem 11 and Corollary 12.

### 7.3 Solutions

In Fig. 11, the top row displays the expected value of the real and imaginary part of the computed stochastic Galerkin approximation  $\tilde{u}_m$  (with polynomial degree  $r = 5$ ). The variance is displayed in the bottom row of the figure.

Denote by  $x_r$  the solution of  $Ax = b$  when using polynomials of degree up to  $r$  in the stochastic Galerkin method, where the number of basis polynomials is given in (64). The left panel of Fig. 12 displays the differences  $\|x_{r-1} - x_r\|_2$  as a function of  $r$  (the vector  $x_{r-1}$  is padded with zeros at the end to match the size of  $x_r$ ). The observed exponential decay suggests convergence of the stochastic Galerkin method.

Next, we fix the degree  $r = 8$  in the stochastic Galerkin method. Recall from (15) and (18) that the solution of  $Ax = b$  contains the coefficient vectors  $V_0, \dots, V_m$  of the polynomials  $\phi_0, \dots, \phi_m$  in the stochastic Galerkin method. We also examine the largest maximum norm of the coefficients associated to polynomials of total degree (exactly)  $j$ , i.e., the values

$$\gamma_j = \max \{ \|V_i\|_\infty : \deg(\phi_i) = j \}. \tag{67}$$

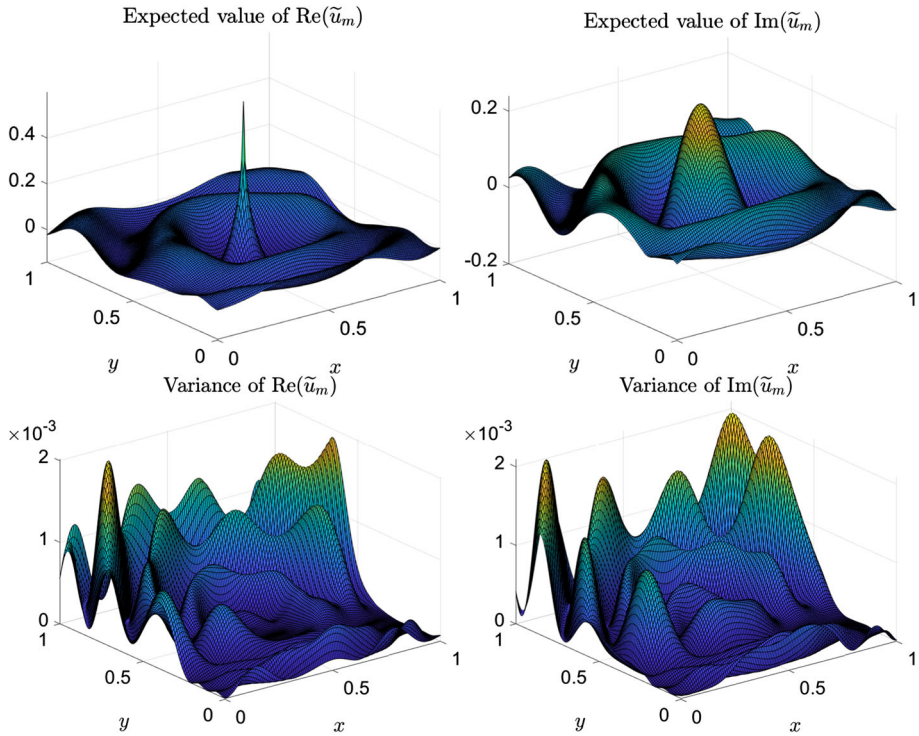


Fig. 11 Expected value (top) and variance (bottom) of  $\text{Re}(\tilde{u}_m)$  and  $\text{Im}(\tilde{u}_m)$

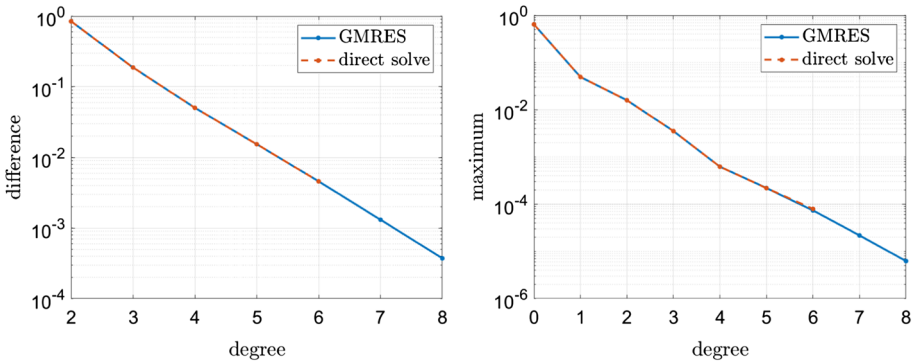


Fig. 12 Left: Euclidean norms  $\|x_{r-1} - x_r\|_2$  as a function of the polynomial degree  $r = 2, \dots, 8$ , where  $x_r$  is the solution of  $Ax = b$  using total degree  $r$  in the stochastic Galerkin method. Right: Magnitudes (67) as a function of the degree  $j$

The right panel of Fig. 12 shows the magnitudes (67) for  $j = 0, 1, \dots, 8$ . The observed exponential decay stems from the exponential convergence of (7), since the wavenumber in (62) is an analytic function of  $\xi_1, \xi_2, \xi_3$ .

We repeat this experiment with  $\theta = 0.2$  instead of  $\theta = 0.1$ . Overall, the behavior is similar as for  $\theta = 0.1$ , but convergence is slower: the relative residual norms reach the

prescribed tolerance in 60 instead of 20 iteration steps, and also  $\|x_{r-1} - x_r\|_2$  as well as the magnitudes (67) converge more slowly.

## 8 Conclusions

We investigated the Helmholtz equation including a random wavenumber. The combination of a stochastic Galerkin method and a finite difference method yielded a high-dimensional linear system of algebraic equations. We examined the iterative solution of these linear systems using three types of preconditioners: a complex shifted Laplace preconditioner, a mean value preconditioner, and a combined variant. Theoretical properties of the preconditioned linear systems were shown. In our numerical experiments, the straightforward mean value preconditioner leads to a more efficient iterative solution of the linear system than the other preconditioners considered here.

**Funding** Open Access funding enabled and organized by Projekt DEAL. No funding was received to assist with the preparation of this manuscript. The authors have no relevant financial or non-financial interests to disclose.

**Data Availability** Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

## Declarations

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Airaksinen, T., Heikkola, E., Pennanen, A., Toivanen, J.: An algebraic multigrid based shifted-Laplacian preconditioner for the Helmholtz equation. *J. Comput. Phys.* **226**(1), 1196–1210 (2007). <https://doi.org/10.1016/j.jcp.2007.05.013>
2. Colton, D., Kress, R.: *Inverse Acoustic and Electromagnetic Scattering Theory*, 3rd edn. Springer, New York (2013)
3. Cools, S., Vanroose, W.: Local Fourier analysis of the complex shifted Laplacian preconditioner for Helmholtz problems. *Numer. Linear Algebra Appl.* **20**(4), 575–597 (2013). <https://doi.org/10.1002/nla.1881>
4. Davis, T.A.: *UMFPACK user guide (version 5.7.7)*. Tech. rep. (2018)
5. Erlangga, Y.A.: Advances in iterative methods and preconditioners for the Helmholtz equation. *Arch. Comput. Methods Eng.* **15**(1), 37–66 (2008). <https://doi.org/10.1007/s11831-007-9013-7>
6. Erlangga, Y.A., Vuik, C., Oosterlee, C.W.: On a class of preconditioners for solving the Helmholtz equation. *Appl. Numer. Math.* **50**(3–4), 409–425 (2004). <https://doi.org/10.1016/j.apnum.2004.01.009>



7. Gander, M.J., Graham, I.G., Spence, E.A.: Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numer. Math.* **131**(3), 567–614 (2015). <https://doi.org/10.1007/s00211-015-0700-2>
8. García Ramos, L., Nabben, R.: On the spectrum of deflated matrices with applications to the deflated shifted Laplace preconditioner for the Helmholtz equation. *SIAM J. Matrix Anal. Appl.* **39**(1), 262–286 (2018). <https://doi.org/10.1137/16M108361X>
9. García Ramos, L., Sète, O., Nabben, R.: Preconditioning the Helmholtz equation with the shifted Laplacian and Faber polynomials. *Electron. Trans. Numer. Anal.* **54**, 534–557 (2021)
10. Ghanem, R.G., Kruger, R.M.: Numerical solution of spectral stochastic finite element systems. *Comput. Meth. Appl. Mech. Engrg.* **129**, 289–303 (1996)
11. Ghanem, R.G., Spanos, P.D.: *Stochastic Finite Elements: A Spectral Method Approach*. Springer, New York (1991)
12. van Gijzen, M.B., Erlangga, Y.A., Vuik, C.: Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM J. Sci. Comput.* **29**(5), 1942–1958 (2007). <https://doi.org/10.1137/060661491>
13. Gittelsohn, C.J.: An adaptive stochastic Galerkin method for random elliptic operators. *Math. Comput.* **82**(283), 1515–1541 (2013)
14. Gottlieb, D., Xiu, D.: Galerkin method for wave equations with uncertain coefficients. *Comm. Comput. Phys.* **3**(2), 505–518 (2008)
15. Griffiths, D.F., Dold, J.W., Silvester, D.J.: *Essential Partial Differential Equations*. Springer Undergraduate Mathematics Series. Springer, Cham (2015)
16. Grossmann, C., Roos, H.G., Stynes, M.: *Numerical Treatment of Partial Differential Equations*. Springer, Berlin (2007)
17. Ihlenburg, F.: *Finite Element Analysis of Acoustic Scattering*, vol. 132. Springer, New York (1998)
18. Lahaye, D., Tang, J., Vuik, K. (eds.): *Modern solvers for Helmholtz problems*. Birkhäuser/Springer, Cham (2017). <https://doi.org/10.1007/978-3-319-28832-1>
19. Livshits, I.: Use of shifted Laplacian operators for solving indefinite Helmholtz equations. *Numer. Math. Theory Methods Appl.* **8**(1), 136–148 (2015). <https://doi.org/10.4208/nmtma.2015.w03si>
20. Polyanin, A.D.: *Handbook of Linear Partial Differential Equations for Engineers and Scientists*. Chapman & Hall/CRC, Boca Raton (2002)
21. Pulch, R.: Stability-preserving model order reduction for linear stochastic Galerkin systems. *J. Math. Ind.* (2019). <https://doi.org/10.1186/s13362-019-0067-6>
22. Pulch, R., van Emmerich, C.: Polynomial chaos for simulating random volatilities. *Math. Comput. Simul.* **80**(2), 245–255 (2009). <https://doi.org/10.1016/j.matcom.2009.05.008>
23. Pulch, R., Sète, O.: The Helmholtz equation with uncertainties in the wavenumber. arXiv preprint: 2209.14740v1 (2022). <https://doi.org/10.48550/arXiv.2209.14740>
24. Pulch, R., Xiu, D.: Generalised polynomial chaos for a class of linear conservation laws. *J. Sci. Comput.* **51**(2), 293–312 (2012)
25. Reps, B., Vanroose, W., Bin Zubair, H.: On the indefinite Helmholtz equation: complex stretched absorbing boundary layers, iterative analysis, and preconditioning. *J. Comput. Phys.* **229**(22), 8384–8405 (2010). <https://doi.org/10.1016/j.jcp.2010.07.022>
26. Saad, Y.: *Iterative Methods for Sparse Linear Systems*, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2003). <https://doi.org/10.1137/1.9780898718003>
27. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.* **7**(3), 856–869 (1986). <https://doi.org/10.1137/0907058>
28. Sheikh, A.H., Lahaye, D., Garcia Ramos, L., Nabben, R., Vuik, C.: Accelerating the shifted Laplace preconditioner for the Helmholtz equation by multilevel deflation. *J. Comput. Phys.* **322**, 473–490 (2016). <https://doi.org/10.1016/j.jcp.2016.06.025>
29. Stoer, J., Bulirsch, R.: *Introduction to Numerical Analysis*, 3rd edn. Springer, New York (2002)
30. Wang, G., Xue, F., Liao, Q.: Localized stochastic Galerkin methods for Helmholtz problems close to resonance. *Int. J. Uncertain. Quantif.* **11**(5), 77–99 (2021)
31. Xiu, D.: *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, Princeton (2010)
32. Xiu, D., Shen, J.: Efficient stochastic Galerkin methods for random diffusion equations. *J. Comput. Phys.* **228**, 266–281 (2009)
33. Youssef, M., Pulch, R.: Poly-Sinc solution of stochastic elliptic differential equations. *J. Sci. Comput.* (2021). <https://doi.org/10.1007/s10915-021-01498-9>