UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
Preconditioning Stochastic Galerkin Methods of Diffusion Problems with Random Data

Permalink
https://escholarship.org/uc/item/7m64r1nw

Author
Wang, Dongwu

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Preconditioning Stochastic Galerkin Methods of Diffusion Problems with Random Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Dongwu Wang

Dissertation Committee:
Professor Long Chen, Chair
Chancellor's Professor Hongkai Zhao
Professor Jack Xin

2018

© 2018 Dongwu Wang

DEDICATION

To my family

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **vi**

ACKNOWLEDGMENTS **viii**

CURRICULUM VITAE **ix**

ABSTRACT OF THE DISSERTATION **x**

1 Introduction **1**

2 Model Problems **5**
2.1 Problem Setting 5
2.2 Karhunen-Loève Expansion 7
2.2.1 Derivation 7
2.2.2 Analytical Examples of Covariance Functions 8
2.2.3 Numerical Methods to Compute the Covariance Function 10
2.2.4 Truncated Karhunen-Loève Expansion 16

3 Matrix Form of Stochastic Galerkin Methods **18**
3.1 Spectral Stochastic Finite Element Methods 18
3.2 Matrix Representation 20
3.3 Sparsity in Stochastic Matrices 22
3.3.1 Stochastic matrices 22
3.3.2 Generalized Polynomials Chaos 23
3.3.3 Tri-diagonal Blocks in Stochastic Matrices 24
3.3.4 Simplex Lattice Representation 26

4 Preconditioning **31**
4.1 A Block Diagonal Preconditioner using Mean Field 31
4.1.1 A tighter bound for the preconditioned system 32
4.1.2 Symmetry for the Spectrum of the Preconditioned System 36
4.2 A New Block Diagonal Preconditioner 38
4.3 A Block Triangular Preconditioner 41

4.3.1 Block-triangular preconditioner 41
4.3.2 Convergence Analysis 42
4.4 A Kronecker Product Preconditioner 46

5 Numerical Experiments **48**
5.1 Homogeneous Dirichlet Boundary Condition 49
5.1.1 Gaussian Random Variables and Hermite Polynomial Chaos 50
5.1.2 Uniform Random Variables and Legendre Polynomial Chaos 53
5.1.3 Maximal and Minimal Eigenvalues of the Preconditioned System 54
5.2 Mixed Boundary Condition 57
5.2.1 Gaussian Random Variables and Hermite Polynomial Chaos 58
5.2.2 Uniform Random Variables and Legendre Polynomial Chaos 59
5.3 A 3-D Homogeneous Dirichlet Boundary Condition Problem 61

6 Stochastic Galerkin Methods for Mixed Formulation **63**
6.1 Stochastic Galerkin Mixed Formulation 63
6.1.1 Model Problem 64
6.1.2 Karhunen-Loève Expansion 65
6.2 Matrix Form of Stochastic Galerkin Methods 66
6.2.1 Mixed Finite Element Method 66
6.2.2 Matrix Form 67
6.3 Preconditioning 70
6.3.1 A Mean-Based Diagonal Preconditioner 70
6.3.2 A Mean-Based Triangular Preconditioner 71
6.3.3 A Block Triangular Preconditioner 72
6.4 Numerical Examples 73
6.4.1 Hermite Polynomials with Gaussian Random Variables 74
6.4.2 Legendre Polynomials with Uniform Random Variables 75

7 Conclusion **77**

Bibliography **78**

LIST OF FIGURES

Page

2.1 The first 20 eigenvalues of the 1-D exponential covariance function with different correlation lengths. 11
2.2 The 4th, 16th and 64th eigenvectors of the 2-D exponential covariance function. . . 12
2.3 Realization of the truncated random field $a_L(\boldsymbol{x}, \xi)$ with $L = 4,\ 16,\ 64$. 12

3.1 Eigenvalues of T (Legendre) 25
3.2 Eigenvalues of T (Hermite) 26
3.3 The simplex lattice and corresponding graph of stochastic matrix G with parameters $p = 5$ and $L = 2$. 28
3.4 Block sparsity structure of the Galerkin matrix $\mathcal{A}$ under the odd-even ordering of their total degree ($L = 4$, $p = 4$). 29
3.5 A simplex representation with $L = 3$ and $p = 3$ (left) and the corresponding G_1 matrix (right) 30

5.1 Mean (left) and variance (right) of pressure on a 32×32 grid with $L = 4$ and $p = 4$ 49
5.2 Mean (left) and variance (right) of pressure on a 32×32 grid with $L = 4$ and $p = 4$ 58
5.3 Mean (left) and variance (right) of pressure on a $16 \times 16 \times 16$ grid with $L = 10$ and $p = 4$ 61

6.1 Mean (left) and variance (right) of $\boldsymbol{u}$ on a 64×64 grid with $L = 4$ and $p = 4$ . . . 74

LIST OF TABLES

Page

5.1 Iteration steps with different mesh size h for all preconditioners (Hermite polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.3$) 50
5.2 Time costs (in seconds) with different mesh size h for all preconditioners (Hermite polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.3$) 50
5.3 Iteration steps with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$) 51
5.4 Time costs with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$) 51
5.5 Iteration steps with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3,\ c_1 = c_2 = 10$) 52
5.6 Time costs with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3,\ c_1 = c_2 = 10$) 52
5.7 Iteration steps with various σ (Hermite polynomials with $\mu = 1,\ h = 1/16,\ p = 4$ and $L = 6$) 52
5.8 Iteration steps with different mesh size h for all preconditioners (Legendre polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.7$) 53
5.9 Time costs with different mesh size h for all preconditioners (Legendre polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.7$) 53
5.10 Iteration steps with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$) 54
5.11 Time costs with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$) 54
5.12 Example 5.1.1: Extremal Eigenvalues with $\sigma = 0.1,\ \mu = 1,\ c = 1,\ h = 1/8$ 55
5.13 Example 5.1.2: Extremal Eigenvalues with $\sigma = 0.3,\ \mu = 1,\ c = 1,\ h = 1/8$ 56
5.14 Example 5.1.3: Extremal Eigenvalues with $\sigma = 0.1,\ \mu = 1,\ c = 10,\ h = 1/8$ . . . 56
5.15 Example 5.1.4: Extremal Eigenvalues with $\sigma = 0.3,\ \mu = 1,\ c = 10,\ h = 1/8$ . . . 57
5.16 Iteration steps with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$) 58
5.17 Time costs with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$) 59
5.18 Iteration Steps with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$) 59
5.19 Time costs with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$) 60

5.20 Iteration steps with various σ (Hermite polynomials with $\mu = 1,\ h = 1/16,\ p = 4$ and $L = 6$) 60
5.21 Iteration steps and time costs with different L and p for both preconditioners (Legendre polynomials with $h = 1/16,\ \mu = 1,\ \sigma = 0.7$) 62

6.1 Iteration steps with various $h,\ L$ and p for all preconditioners (Hermite polynomials with $\mu = 1,\ \sigma = 0.3$) 75
6.2 Time costs with various $h,\ L$ and p for all preconditioners (Hermite polynomials with $\mu = 1,\ \sigma = 0.3$) 75
6.3 Iteration steps with various $h,\ L$ and p for all preconditioners (Legendre polynomials with $\mu = 1,\ \sigma = 0.3$) 76
6.4 Time costs with various $h,\ L$ and p for all preconditioners (Legendre polynomials with $\mu = 1,\ \sigma = 0.3$) 76

ACKNOWLEDGMENTS

I would like to express the deepest gratitude to my Ph.D. advisor, Professor Long Chen, for his wise advice, continuous support and endless patience on my academic research as well as life beyond school. He has set up a great role model to me as a mathematician and it is a great pleasure to work with him in the past few years.

I would like to thank Chancellor's Professor Hongkai Zhao, Professor Jack Xin, Professor Yifeng Yu and Professor Yun Wang for serving as my committee members and share their insights and expertise on my research. I would like to thank my collaborators Dr. Bin Zheng, Dr. Xiaozhe Hu and Dr. Xuehai Huang for all their enlightening discussions together. I would also like to thank all Long's group members and my friends at the Mathematics Department in University of California, Irvine for all the time and memories we spent together in the past five years.

Last but not least, I would like to thank my wife Zhe Chen and my parents for their unconditional supports and love all the time, which is the deepest strength to push me forward in pursuit of my Ph.D. degree.

CURRICULUM VITAE

Dongwu Wang

EDUCATION

Doctor of Philosophy in Mathematics	**2018**
University of California, Irvine	*Irvine, CA*
Bachelor of Science in Computational Mathematics	**2013**
Peking University	*Beijing, China*
Bachelor of Science in Psychology	**2013**
Peking University	*Beijing, China*

ABSTRACT OF THE DISSERTATION

Preconditioning Stochastic Galerkin Methods of Diffusion Problems with Random Data

By

Dongwu Wang

Doctor of Philosophy in Mathematics

University of California, Irvine, 2018

Professor Long Chen, Chair

When solving stochastic partial differential equations with random coefficients, the stochastic Galerkin method results in a large single system via a traditional finite element discretization in the spatial domain and generalized polynomial chaos in the stochastic space. The sparse pattern on the stochastic Galerkin matrix can be described by a simplex lattice structure. With the help of this simplex structure, it is shown that the spectrum of the mean-based preconditioning system is symmetric with respect to one and a tighter eigen-bound for the mean-based preconditioned system is developed. Furthermore, a new variance-involved block diagonal preconditioner and a new block triangular preconditioner based on the simplex lattice structure are proposed and their corresponding spectral analyses are studied. Numerical experiments show that the new preconditioners are more robust and efficient compared to the mean-based preconditioner, especially when the variance and the total polynomial degree in the complete stochastic space are large. The block triangular preconditioner is extended to a stochastic diffusion problem with mixed formulation together with a new mean-based block triangular preconditioner. Both preconditioners outperform the traditional mean-based diagonal preconditioner on a benchmark problem.

Chapter 1

Introduction

Nowadays engineering problems involving a large amount of uncertainties are drawing more and more attentions [2]. The uncertainties usually appear in the input data, including model coefficients, forcing terms, boundary conditions and geometry of the domains, and it is usually treated as random fields in the modeling [40, 2].

A traditional way to solve the uncertainty problem is by Monte Carlo Methods (MCM) [12, 3, 24]. Randomly sampling the uncertainty repeatedly results in a large amount of deterministic systems and each can be solved by existing numerical solvers. The statistical properties are then generated by the results of each deterministic system. However, large numbers of realizations have to be performed in order to ensure the proper sampling of the random fields as well as the convergence of MCM. Moreover, it is always worth optimizing the computational cost of the numerical solver for the individual deterministic system.

Alternatively, Spectral Stochastic Finite Element Method (SSFEM) [23] is another popular technique in dealing with stochastic PDEs by parametrizing the random fields by a Karhunen-Loève expansion [28] and discretizing the space with traditional finite element basis in the spatial domain and generalized polynomial chaos [45] in the stochastic domain. A single linear system is then to

be solved and the results are post-processed to generate the statistical properties of interest. Although much larger than the scales of each deterministic problem in MCM, the single linear system from SSFEM usually has nicer structures to employ fast solvers.

SSFEM can be divided into two categories according to the choice of stochastic basis functions, namely Stochastic Collocation (SC) methods and Stochastic Galerkin (SG) methods. Stochastic Collocation methods [2] apply doubly-orthogonal polynomials with maximal degree p in each individual direction while Stochastic Galerkin methods [23] employ orthogonal polynomials with maximal total degree p. The dimension of the former space $(1+p)^L$ grows exponentially with respect to the number of Karhunen-Loève expansion terms L, while that of the latter space $\binom{L+p}{p}$ only grows algebraically. However, the former case results in a decoupled system with the introduction of doubly orthogonal polynomials [2] while the latter one, together with orthogonal polynomials, leads to a fully coupled system that requires sophisticated algorithms to solve [33, 30, 19, 35, 39].

We shall focus on the Stochastic Galerkin methods in this thesis. Consider a steady-state diffusion problem with random diffusion coefficients. The resulting fully coupled linear system can be written as $\mathcal{A}\boldsymbol{u} = \boldsymbol{f}$ after applying Stochastic Galerkin methods with Karhunen-Loève expansions for the random field. The matrix $\mathcal{A}$ is the sum of tensor products of matrices

$$\mathcal{A} = \sum_{s=0}^{L} G_s \otimes K_s,$$

where K_s are the stiffness matrices in space using traditional finite element basis functions with truncated Karhunen-Loève expansion, and G_s are the stochastic matrices corresponding to generalized orthogonal polynomials. Pellissetti [30] and Ghanem [22] provided a natural mean-based preconditioner $\mathcal{P}_0 = G_0 \otimes K_0$ containing the mean of the random coefficient for the system discretized by SSFEM. Powell and Elman [33] then gave detailed spectral analysis on this mean-based preconditioner.

We shall explore more on the sparse pattern of the stochastic matrix $\mathcal{A}$ by introducing a simplex

lattice structure. Based on that one can show the spectrum of $\mathcal{P}_0^{-1}\mathcal{A}$ is symmetric with respect to 1 and a tighter bound for the preconditioned system is provided using the Gershgorin's theorem for block matrices. More importantly, a new variance-involved preconditioner $\mathcal{P}_1 = G_0 \otimes K_0 + G_1 \otimes K_1$ is presented which contains the information on not only the mean but also part of the variance of the random coefficient and a corresponding spectral analysis is also given. A block triangular preconditioner $\mathcal{P}_T$ associated with the simplex structure and odd-even indices is then introduced together with its convergence analysis. In addition, a Kronecker product preconditioner $\mathcal{P}_K$ [42] is reviewed, which minimized the Frobenius norm of a single tensor product to the matrix $\mathcal{A}$. The performances of all preconditioners are compared on several benchmark problems and the new preconditioners are more robust and efficient in terms of iteration steps as well as computational time. The block triangular preconditioner $\mathcal{P}_T$ outperforms the other three preconditioners in all problems, and it converges at least twice faster compared to the mean-based preconditioner $\mathcal{P}_0$, especially when the variance σ and total degree p are large. The variance-involved preconditioner $\mathcal{P}_1$ outperforms $\mathcal{P}_0$ when the total degree p and the correlation length c are large.

The stochastic Galerkin finite element method is then extended to the mixed formulation of stochastic diffusion problems [18, 16]. Other than the original mean-based diagonal preconditioner $\mathcal{P}_D$ introduced in [18, 16], two new preconditioners $\mathcal{P}_0$, a mean-based triangular preconditioner, and $\mathcal{P}_T$, a block triangular preconditioner, are introduced. Both $\mathcal{P}_0$ and $\mathcal{P}_T$ are direct extensions from the corresponding preconditioners in the primal formulation, and they are inspired by the block triangular preconditioner for the deterministic saddle-point system [8]. The performances of all three preconditioners are compared on a benchmark mixed formulation problem, and both $\mathcal{P}_0$ and $\mathcal{P}_T$ outperform $\mathcal{P}_D$ by twice faster, especially when the variance is large.

The outline of the thesis is as follows.

- In Chapter 2 we introduce the steady-state diffusion model problem (primal formulation). The Karhunen-Loève expansion, which can be seen as a finite approximation to the random

fields, is reviewed and several analytical and numerical examples are presented to compute the truncated Karhunen-Loève expansion.

- In Chapter 3 we discretize the system using SSFEM with generalized polynomial chaos in the random space and traditional linear finite element basis in the spatial space. The sparsity of stochastic matrices are then discussed. A simplex representation is presented with odd-even indices, which will be helpful in designing new preconditioners as well as in the corresponding spectral analysis.

- In Chapter 4 we first provide a tighter bound to the preconditioned system as well as a symmetric analysis for the eigenvalues of the mean-based preconditioner $\mathcal{P}_0$. Then we present a new variance-involved block diagonal preconditioner $\mathcal{P}_1$, a new block triangular preconditioner $\mathcal{P}_T$, a Kronecker product preconditioner $\mathcal{P}_K$ and their corresponding spectral analyses.

- In Chapter 5 we compare the performances of all four preconditioners on two benchmark problems with various parameters and polynomials.

- In Chapter 6 we extend the stochastic Galerkin finite element method to the mixed formulation, transfer all the preconditioners into the mixed system and test the performances on a benchmark problem with mixed formulation.

- We conclude this thesis with some remarks in Chapter 7.

Chapter 2

Model Problems

In this chapter we first present the steady-state diffusion problem with random coefficients and suitable boundary conditions. Then we derive a finite term approximation to the random field with given covariance function via Karhunen-Loève expansion and present several analytical and numerical examples on the kernel function. We end this chapter by a deterministic weak formulation with a truncated Karhunen-Loève expansion.

2.1 Problem Setting

Let $D \subset \mathbb{R}^2$ be a bounded domain. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space where Ω is the set of outcomes, $\mathcal{F}$ is the σ-algebra of the subsets of Ω, and P is a probability measure on $\mathcal{F}$. Consider the stochastic steady-state diffusion problem with homogeneous Dirichlet boundary conditions: find a random field $u(\boldsymbol{x}, \omega) : \ D \times \Omega \to \mathbb{R}$ such that for P-almost surely, the following

equations hold:

$$\begin{cases} -\nabla \cdot (a(\boldsymbol{x},\omega)\nabla u(\boldsymbol{x},\omega)) = f(\boldsymbol{x}), & \text{in } D \times \Omega, \\ u(\boldsymbol{x},\omega) = 0, & \text{on } \partial D \times \Omega. \end{cases} \tag{2.1}$$

Here we assume the randomness comes from the diffusive coefficient $a(\boldsymbol{x},\omega) : D \times \Omega \to \mathbb{R}$, while the source term f and boundary conditions are deterministic. It is not hard to generate the problems in which the source term f also contains randomness.

Let $L^2(D) \otimes L^2(\Omega) = \{u(\boldsymbol{x},\omega) : D \times \Omega \to \mathbb{R} \mid \|u\| < \infty\}$ be the space consists of all random functions with finite second moments, where the norm is defined as

$$\|u(\boldsymbol{x},\omega\|^2 = \int_\Omega \int_D u^2(\boldsymbol{x},\omega)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\omega.$$

Then define $V = H_0^1(D) \otimes L^2(\Omega) = \{u(\boldsymbol{x},\omega) : D \times \Omega \to \mathbb{R} \mid \|u\|_V < \infty,\ u|_{\partial D \times \Omega} = 0\}$ with the corresponding V-norm defined as

$$\|u(\boldsymbol{x},\omega\|_V^2 = \int_\Omega \int_D a(\boldsymbol{x},\omega)|\nabla u(\boldsymbol{x},\omega)|^2\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\omega.$$

The weak formulation to (2.1) is to find $u(\boldsymbol{x},\omega) \in H_0^1(D) \otimes L^2(\Omega)$ such that

$$\int_\Omega \int_D a(\boldsymbol{x},\omega)\nabla u(\boldsymbol{x},\omega) \cdot \nabla v(\boldsymbol{x},\omega)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\omega = \int_\Omega \int_D f(\boldsymbol{x})v(\boldsymbol{x},\omega)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\omega, \quad \forall v \in V. \tag{2.2}$$

To be well posed, we assume

$$0 < c_{\min} \leqslant a(\boldsymbol{x},\omega) \leqslant c_{\max} < +\infty, \quad \text{a.e. in } D \times \Omega. \tag{2.3}$$

Then the V-norm is well-defined. The weak solution in $V = H_0^1(D) \otimes L^2(\Omega)$ exists and is unique for (2.2) by the Lax-Milgram Lemma.

2.2 Karhunen-Loève Expansion

We consider the colored noise scenario here [2], where the input data vary randomly between any two points in the physical domain. The uncertainty is described as a random field with given covariance structure. To convert the stochastic equation (2.2) to a deterministic one, one has to parametrize the input random fields by a finite number random variables with a given accuracy. A popular choice is to represent $a(\boldsymbol{x}, \omega)$ by Karhunen-Loève expansion [28].

2.2.1 Derivation

Let $\mu(\boldsymbol{x}) = \mathbf{E}[a(\boldsymbol{x}, \cdot)]$ be the expected value of random field $a(\boldsymbol{x}, \omega)$ and $K(\boldsymbol{x}, \boldsymbol{y})$ be the continuous covariance function of the random field $a(\boldsymbol{x}, \omega)$, i.e.

$$K(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{E}\left[(a(\boldsymbol{x}, \omega) - \mathbf{E}\left[a(\boldsymbol{x}, \omega)\right])(a(\boldsymbol{y}, \omega) - \mathbf{E}\left[a(\boldsymbol{y}, \omega)\right])\right].$$

Due to its boundedness, symmetry and positive definiteness, the covariance function admits a spectral decomposition [13]

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{s=0}^{\infty} \lambda_s \alpha_s(\boldsymbol{x}) \alpha_s(\boldsymbol{y})$$

where the pairs $\{\lambda_s, \alpha_s(\boldsymbol{x})\}_{s=0}^{\infty}$ are the s-th eigenvalue and eigenvector of the covariance kernel, i.e. they are the solutions to the Fredholm integral equation [13]

$$\int_D K(\boldsymbol{x}, \boldsymbol{y}) \alpha_s(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \lambda_s \alpha_s(\boldsymbol{x}). \tag{2.4}$$

Therefore, we have the following Karhunen-Loève expansion theorem [28]

Theorem 2.2.1 (Karhunen-Loève)**.** *Let $a : D \times \Omega \to \mathbb{R}$ be a mean-square continuous stochastic*

process with $a \in L^2(\Omega \times D)$. There exists an eigen-pair basis $\{\lambda_s, \alpha_s(\boldsymbol{x})\}_{s=0}^{\infty}$ from (2.4) *such that*

$$a(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sum_{s=1}^{\infty} \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\xi_s(\omega) = \sum_{s=0}^{\infty} \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\xi_s(\omega), \tag{2.5}$$

where $\mu(\boldsymbol{x}) = \mathbf{E}[a(\boldsymbol{x}, \omega)]$ is the mean of random field and random variables ξ_s given by

$$\xi_s(\omega) = \frac{1}{\sqrt{\lambda_s}} \int_D a(\boldsymbol{x}, \omega)\alpha_s(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

have zero mean $\mathbf{E}[\xi_s] = 0$ and unit variance $\mathrm{Var}(\xi_s) = 1$. Furthermore they are mutually uncorrelated and thus the variance of $a(\cdot, \omega)$ is just the sum of the eigenvalues $Var(a) = \sum_{s=1}^{\infty} \lambda_s$. Note that in the second summation of (2.5)*, to keep the notation simple, we denote $\sqrt{\lambda_0}\alpha_0(\boldsymbol{x}) = \mu(\boldsymbol{x})$ and $\xi_0(\omega) = 1$.*

Remark 2.2.2 (Convergence)**.** *For continuous covariance functions $K(\boldsymbol{x}, \boldsymbol{y})$, the above convergence is uniform on D in the mean squared error sense, i.e.*

$$\mathbf{E}[(a(\boldsymbol{x}, \omega) - \sum_{s=0}^{\infty} \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\xi_s(\omega))^2] \to 0.$$

2.2.2 Analytical Examples of Covariance Functions

A list of popular kernels with analytical expressions of eigen-pairs are presented below in 1-D case (see Chapter 1.2 in [44] for more details):

- Brownian kernel: The covariance function $K(x, y)$ for the Brownian motion is $\min(x, y)$. The corresponding Fredholm integral equation can be written as

$$\int_D \min(x, y)\alpha(y) \, \mathrm{d}y = \lambda\alpha(x),$$

and the resulting eigen-pairs can be formulated as

$$\lambda_i = \frac{1}{(i+1/2)^2\pi^2}, \quad \alpha_i(x) = \sqrt{2}\sin((i+\frac{1}{2})\pi x), \quad i = 0, 1, \cdots .$$

- Brownian bridge: The Brownian bridge is a standardized Brownian motion with the endpoint equals to 0. The covariance function $K(x, y)$ is therefore $\min(x, y) - xy$. The corresponding Fredholm integral equation can be written as

$$\int_D (\min(x, y) - xy)\alpha(y)\, \mathrm{d}y = \lambda\alpha(x),$$

and the resulting eigen-pairs can be formulated as

$$\lambda_i = \frac{1}{i^2\pi^2}, \quad \alpha_i(x) = \sqrt{2}\sin(i\pi x), \quad i = 1, 2, \cdots .$$

Another example of a kernel function with analytical eigen-pair solutions to the Fredholm integral equation (2.4) is discussed in details below, and we will use this exponential kernel example in the numerical experiments. Consider the following exponential kernel as a continuous symmetric kernel example

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\boldsymbol{L}_c}\right), \tag{2.6}$$

where $D = [-a, a] \times [-a, a]$ and $\boldsymbol{L}_c$ is the correlation length. The eigen-pair solution corresponding to (2.6) exists since the 2-D kernel can be separated into two 1-D exponential kernels and the 2-D eigen-pair solution is simply the multiplication of the two 1-D eigen-pair solutions. The corresponding 1-D Fredholm integral equation is

$$\int_{-a}^{a} \sigma^2 \exp\left(-\frac{|x - y|}{L_c}\right) \alpha_s(y)\, \mathrm{d}y = \lambda_s\alpha_s(x) \tag{2.7}$$

and the eigen-pair solution $\{\lambda_s, \alpha_s(x)\}_{s=0}^{\infty}$ is given in [23]

$$\lambda_s = \begin{cases} \dfrac{2\sigma^2 c}{w_s^2 + c^2}, & s = 1, 3, 5, \cdots \\ \dfrac{2\sigma^2 c}{(w_s^*)^2 + c^2}, & s = 2, 4, 6, \cdots \end{cases} \qquad \alpha_s(x) = \begin{cases} \dfrac{\cos(w_s x)}{\sqrt{a + \frac{\sin(2w_s a)}{2w_s}}}, & s = 1, 3, 5, \cdots \\ \dfrac{\sin(w_s^* x)}{\sqrt{a - \frac{\sin(2w_s^* a)}{2w_s^*}}}, & s = 2, 4, 6, \cdots \end{cases} \tag{2.8}$$

where $c = 1/L_c$, and w and w^* are the solutions of

$$c - w\tan(wa) = 0, \quad w^* + c\tan(w^* a) = 0.$$

Let $D = [-0.5, -0.5] \times [-0.5, 0.5]$ and $\mu = 1$. Figure 2.1 shows the first 20 eigenvalues of the 1-D exponential covariance function with various correlation lengths L_c. The eigenvalues decay, and the larger the correlation length is, the faster the eigenvalues decay. Figure 2.2 shows the 4th, 16th and 64th eigenvectors of the 2-D exponential covariance function with $\boldsymbol{L}_c = (1, 1)$ on D. The eigenvectors tend to be more oscillatory as s increases. Figure 2.3 gives three realizations of the truncated random field $a_L(\boldsymbol{x}, \xi)$ on D with $L = 4,\ 16,\ 64$, $\sigma = 0.6$ and $\boldsymbol{L}_c = (1, 1)$. In the last case ($L = 64$) the approximated random field is not even strictly positive $a.e.$ on $D \times \Omega$. If the variance σ is not that large, however, the resulting deterministic SSFEM system could still be positive definite with proper choice of L and p.

2.2.3 Numerical Methods to Compute the Covariance Function

The existence of such eigen-pair solutions for Fredholm integral equation (2.4) is guaranteed for continuous symmetric kernels [13], however, the computation of such eigen-pair solutions is not that straightforward. Only few kernels have analytic solutions [23, 44] while most of the kernels do not. A few numerical methods have been proposed to solve the Fredholm integral equation (2.4)

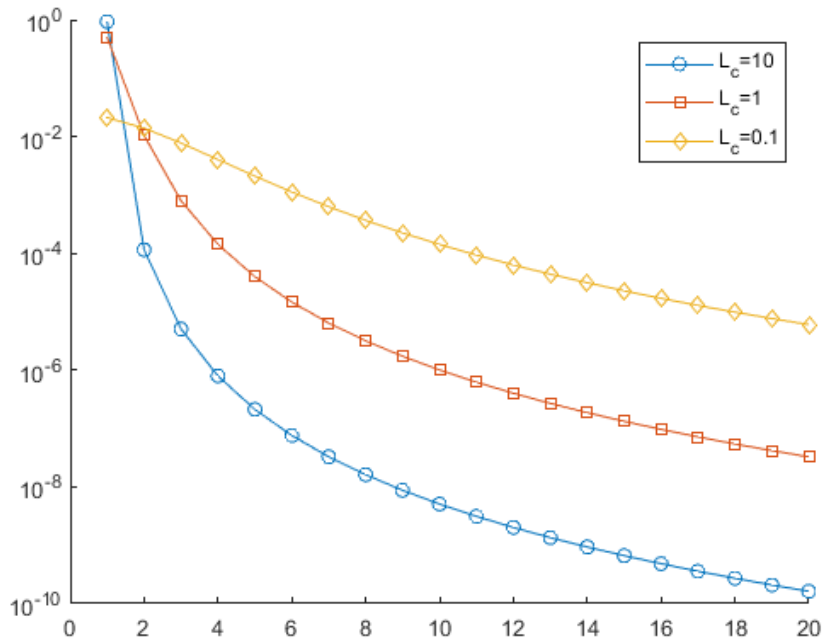


Figure 2.1: The first 20 eigenvalues of the 1-D exponential covariance function with different correlation lengths.

including a mixed wavelet-Galerkin method [31] and a fast multiple method [21]. In this subsection we discuss two general numerical approaches to compute the Fredholm integral equation in (2.4) including integral method and expansion method [44]. And we conclude this subsection with an efficient randomized SVD decomposition [26] for solving the resulting linear system of eigenvalue problem.

Integral Method

Integral method applies numerical integration to approximate the left-hand side integral with a weighted finite sum. Given a set of points $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\} \subset D$ and their corresponding weights $\{w_j\}_{j=1}^N$ associated to the numerical integration scheme, we can rewrite the Fredholm integral equation as

$$\int_D K(\boldsymbol{x}, \boldsymbol{y})\alpha(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} = \sum_{j=1}^{N} K(\boldsymbol{x}, \boldsymbol{y}_j)\alpha(\boldsymbol{y}_j)w_j = \lambda\alpha(\boldsymbol{x}).$$

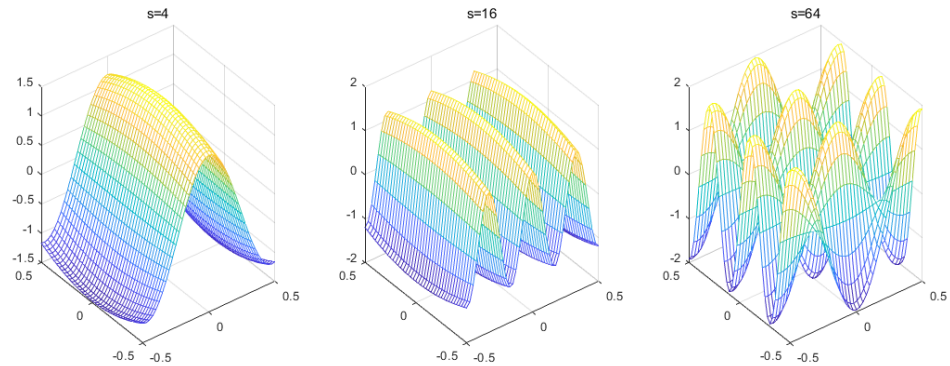


Figure 2.2: The 4th, 16th and 64th eigenvectors of the 2-D exponential covariance function.

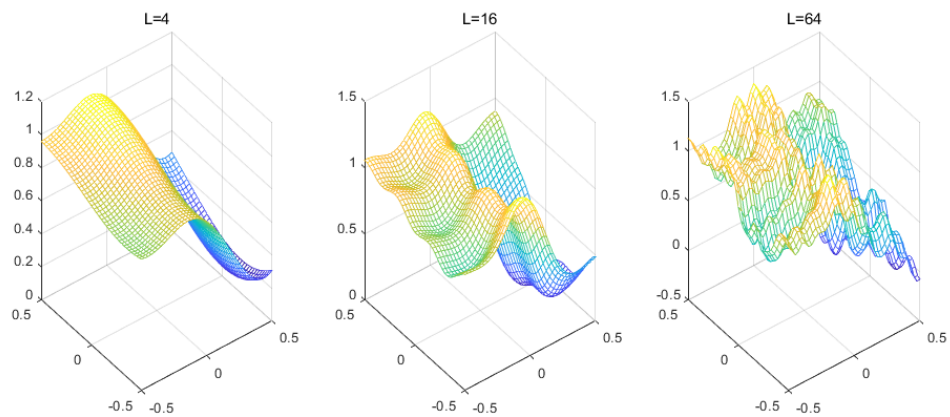


Figure 2.3: Realization of the truncated random field $a_L(\boldsymbol{x}, \xi)$ with $L = 4,\ 16,\ 64$.

Plug in $\boldsymbol{x} = \boldsymbol{y}_i,\ i = 1, 2, \cdots, N$, we have a matrix form

$$\tilde{K} W \boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}, \tag{2.9}$$

where $\tilde{K} = (K(\boldsymbol{y}_i, \boldsymbol{y}_j))_{N\times N},\ W = \mathrm{diag}(w_1, w_2, \cdots, w_N),\ \boldsymbol{\alpha} = (\alpha(\boldsymbol{y}_1), \alpha(\boldsymbol{y}_2), \cdots, \alpha(\boldsymbol{y}_N))^\intercal$. Since W is a diagonal matrix with positive diagonal elements, let $\boldsymbol{\beta} = W^{1/2}\boldsymbol{\alpha}$ and we can get

$$(W^{1/2}\tilde{K}W^{1/2})\boldsymbol{\beta} = \lambda\boldsymbol{\beta}. \tag{2.10}$$

The eigen-pairs $(\lambda, \boldsymbol{\alpha})$ can be solved on the grid points $(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N)$ by simply doing SVD-decomposition to equation (2.10) and compute $\boldsymbol{\alpha} = W^{-1/2}\boldsymbol{\beta}$.

Midpoint rule and trapezoidal rule are among most popular numerical integration schemes. Take an 1-D integral as an example. Let $a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$ be a uniform discretization in $[a, b]$ with $h = (b - a)/N$ and each $x_i = a + ih$. For midpoint rule, $y_i = (x_{i-1} + x_i)/2$ with $w_i = h$ for $i = 1, 2, \cdots, N$. For trapezoidal rule, $y_i = x_i$ for $i = 1, 2, \cdots, N$, and $W = \mathrm{diag}(h/2, h, h, \cdots, h, h/2)$. Both midpoint rule and trapezoidal rule can be easily extended to higher dimensions.

Expansion Method

The integral method results in only the nodal values of the eigenfunction $\alpha(\boldsymbol{x})$ based on the discretization, while the expansion method treats the eigenfunction as a function of $\boldsymbol{x}$ with certain basis functions. Assume we have a set of basis functions $\{\beta_i(\boldsymbol{x})\}_{i=1}^N$ and the eigenfunction $\alpha(\boldsymbol{x})$ can be expanded as

$$\alpha(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \beta_i(\boldsymbol{x}).$$

Insert into the Fredholm integral equation (2.4) and get

$$\sum_{i=1}^{N} \alpha_i \int_D K(\boldsymbol{x}, \boldsymbol{y}) \beta_i(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \lambda \sum_{i=1}^{N} \alpha_i \beta_i(\boldsymbol{x}).$$

Multiply $\beta_j(\boldsymbol{x})$ on both sides and integrate over D we have

$$\sum_{i=1}^{N} \alpha_i \int_{D \times D} K(\boldsymbol{x}, \boldsymbol{y}) \beta_i(\boldsymbol{y}) \beta_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{x} = \lambda \sum_{i=1}^{N} \alpha_i \int_D \beta_i(\boldsymbol{x}) \beta_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{2.11}$$

Let

$$\tilde{K} = \int_{D\times D} K(\boldsymbol{x},\boldsymbol{y})\beta_i(\boldsymbol{y})\beta_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x},$$
$$\tilde{B} = \int_{D} \beta_i(\boldsymbol{x})\beta_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$
$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)^\intercal.$$

Then we get a generalized eigen-value problem

$$\tilde{K}\boldsymbol{\alpha} = \lambda\tilde{B}\boldsymbol{\alpha}. \tag{2.12}$$

Since $\tilde{B}$ is a diagonal matrix (the basis are orthogonal) with positive diagonal elements, let $\boldsymbol{\gamma} = \tilde{B}^{1/2}\boldsymbol{\alpha}$ we have

$$(\tilde{B}^{-1/2}\tilde{K}\tilde{B}^{-1/2})\boldsymbol{\gamma} = \lambda\boldsymbol{\gamma}. \tag{2.13}$$

The eigen-pairs $(\lambda, \boldsymbol{\alpha})$ can be solved by doing SVD-decomposition to equation (2.13) and compute $\boldsymbol{\alpha} = \tilde{B}^{-1/2}\boldsymbol{\gamma}$. Trigonometrical functions [23] and wavelet functions [31] are among the most popular choices for the basis functions.

Expansion method with wavelet basis functions performs best in numerical experiments [44], however it is not trivial to generate the wavelet basis from 1-D to 2-D or even higher dimensions. Thus we shall use integral method with midpoint rule to compute 2-D Fredholm integral equation (2.4) here. Let $\mathcal{T}_h = \cup_{i=1}^{N_T} K_i$ be a triangular discretization of D with uniform refinement and each K_i is an element (triangle in 2D) in $\mathcal{T}_h$ and denote $\boldsymbol{x}_c^i$ as the center of element K_i. By simple calculation, in (2.10) $N = N_T$, $\tilde{K} = (K(\boldsymbol{x}_c^i, \boldsymbol{x}_c^j))_{N\times N}$, $W = \mathrm{diag}(|K_1|, |K_2|, \cdots, |K_N|)$, $\boldsymbol{\alpha} = (\alpha(\boldsymbol{x}_c^1), \alpha(\boldsymbol{x}_c^2), \cdots, \alpha(\boldsymbol{x}_c^N))^\intercal$. The advantage of integral method discussed here is that the discretization $\mathcal{T}_h$ coincides with the mesh used later on in the discretization of spatial domain. The nodal values $\boldsymbol{\alpha}$ on the mesh can be used directly in assembling the stiffness matrices.

Randomized SVD Decomposition

Only the first few eigenvalues with largest magnitudes and their corresponding eigenvectors need to be solved through (2.10) or (2.13), which requires a partial SVD decomposition. Here we use an efficient randomized SVD decomposition to compute the first k eigen-pairs of the problem $A\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$ with $A \in \mathbb{R}^{n\times n}$. The general procedure is given below (cf. Algorithm 4.1 and 4.5 in [26])

- Computes an $n \times l$ orthonormal matrix Q whose range approximates the range of A.
  - Generate an $n \times l$ Gaussian random matrix Ω.
  - Form the $n \times l$ matrix $Y = A\Omega$.
  - Compute QR factorization $Y = QR$ to get an $n \times l$ matrix Q whose columns form an orthonormal basis for $\mathrm{range}(Y)$.
- Compute an approximate SVD factorization $A \approx U\Sigma V^*$, where U and V are orthonormal and Σ is a nonnegative diagonal matrix.
  - Form the $l \times n$ matrix $B = Q^*A$.
  - Compute SVD decomposition for the small matrix $B = \tilde{U}\Sigma V^*$.
  - Form the orthonormal matrix $U = Q\tilde{U}$.

Here $l = k + p$ for some oversampling parameter p and in practice $p = 5$ or $p = 10$ usually results in good performance. The randomized SVD decomposition has the advantage that it only need to solve a SVD problem with a much smaller scale $(l + p) \times (l + p)$ compared to $n \times n$ originally. We refer to [26] for more details on randomized SVD decomposition.

2.2.4 Truncated Karhunen-Loève Expansion

The random field $a(\boldsymbol{x}, \omega)$ is well approximated by the truncated Karhunen-Loève expansion

$$a_L(\boldsymbol{x}, \omega) = \mu(\boldsymbol{x}) + \sum_{s=1}^{L} \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\xi_s(\omega) = \sum_{s=0}^{L} \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\xi_s(\omega)$$

in order to convert the stochastic weak formulation (2.2) into a deterministic system.

The random variable $\xi_s(\omega) : \Omega \to \Gamma_s \subset \mathbb{R}$ is a real valued function defined on the sample space Ω. We will use it as a variable ξ_s and thus ignore the notation ω. Let $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_L) \in \mathbb{R}^L$. Then the coefficients can be written as $\hat{a}_L(\boldsymbol{x}, \boldsymbol{\xi})$ where $(\boldsymbol{x}, \boldsymbol{\xi}) \in D \times \mathbb{R}^L$.

The modified weak formulation with truncated Karhunen-Loève expansion is given as

$$\int_\Gamma \int_D \hat{a}_L(\boldsymbol{x}, \boldsymbol{\xi}) \nabla u(\boldsymbol{x}, \boldsymbol{\xi}) \cdot \nabla v(\boldsymbol{x}, \boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{x} \, p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} = \int_\Gamma \int_D f(\boldsymbol{x}) v(\boldsymbol{x}, \boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{x} \, p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}, \tag{2.14}$$

where $p(\boldsymbol{\xi})$ is the joint probability density function, and $\Gamma = \Gamma_1 \times \Gamma_2 \times \cdots \times \Gamma_L$ is the range of $\boldsymbol{\xi}(\omega)$ in $\mathbb{R}^L$.

Remark 2.2.3 (Independence)**.** If random variables are jointly independent, then the joint density function is separable

$$p(\boldsymbol{\xi}) = p(\xi_1)p(\xi_2) \cdots p(\xi_L)$$

and the computation can be simplified significantly. In Karhunen-Loève expansion, however, the random variables ξ_i are only mutually uncorrelated. Gaussian random variables are often used in practice because uncorrelated Gaussian random variables imply independence. However, the uniform ellipticity (2.3) is not automatically satisfied and the weak formulation (2.14) may not be well posed. An alternative approach, which we shall follow, is to use random variables satisfying (2.14) and make an additional assumption on independence. To relax the independence assumption, we refer to [2].

Remark 2.2.4 (Ellipticity)**.** The elliptical assumption of the truncated Karhunen-Loève expansion

$$0 < \hat{c}_{\min} \leq \hat{a}_L(\boldsymbol{x}, \boldsymbol{\xi}) \leq \hat{c}_{\max} < +\infty, \quad \text{a.e. in } D \times \Gamma. \tag{2.15}$$

is not automatically inherit from (2.3) as the eigenfunctions α_s may not be pointwise positive. To derive (2.15) from (2.3), we require $\boldsymbol{\xi}$ have uniformly bounded images, which rules out the most popular Gaussian random variables, and the eigenvalues of the covariance function decay sufficiently fast [21]. An example of the violation of (2.15) is given in [33].

Chapter 3

Matrix Form of Stochastic Galerkin Methods

In this chapter we first employ the matrix representation of the weak formulation (2.14). Then we introduce the generalize polynomial chaos (gPC) basis in the complete polynomial space and explore the block structure of the resulting stochastic matrices. In the end, we provide a simplex lattice representation of the stochastic matrices which will be helpful to both the theoretical analysis and the numerical computations.

3.1 Spectral Stochastic Finite Element Methods

Ghanem and Spanos [22] introduced spectral stochastic finite element method (SSFEM) as an extension of the traditional finite element method for partial differential equations with random coefficients. SSFEM discretizes the solution space W onto the deterministic space $H_0^1(D)$ and the

stochastic space $L^2(\Gamma)$ separately, i.e. looking for a solution $u(\boldsymbol{x}, \boldsymbol{\xi}) \in W^h = X_h \otimes S_{p,L}$, where

$$X_h = \mathrm{span}\{\phi_i(\boldsymbol{x})\}_{i=1}^{N_x} \subset H_0^1(D), \quad S_{p,L} = \mathrm{span}\{\Psi_k(\boldsymbol{\xi})\}_{k=1}^{N_\xi} \subset L^2(\Gamma).$$

The spatial space X_h is the standard finite element space, for example, P_1-basis functions with a triangular mesh in the domain D. The stochastic space $S_{p,L}$ can be either tensor product polynomial space

$$\mathfrak{T}_{p,L} \stackrel{\text{def}}{=} \left\{ \prod_{s=1}^{L} \psi_s(\xi_s) \mid \max_{1 \le s \le L} \deg(\psi_s) \leqslant p \right\} \tag{3.1}$$

with univariate polynomials of maximal individual degree p in each direction, or complete polynomial space

$$\mathfrak{C}_{p,L} \stackrel{\text{def}}{=} \left\{ \prod_{s=1}^{L} \psi_s(\xi_s) \mid \sum_{s=1}^{L} \deg(\psi_s) \leqslant p \right\} \tag{3.2}$$

with univariate polynomials of maximal total degree p. The dimension of the former space $(1+p)^L$ grows exponentially with respect to the number of expansions L, while that of the latter space $\binom{L+p}{p}$ only grows algebraically. However, the former case results in a decoupled system with the introduction of doubly orthogonal polynomials [2] while the latter one, together with orthogonal polynomials, only leads to a fully coupled system that requires sophisticated algorithms to solve [33, 30, 19, 35, 39].

We shall apply the complete polynomial space (3.2) with orthonormal basis due to its lower dimensionality, and we shall develop delicate algorithms to solve the coupled system efficiently.

3.2 Matrix Representation

We expand the solution into basis

$$u(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{i=1}^{N_x} \sum_{k=1}^{N_\xi} u_{ik} \phi_i(\boldsymbol{x}) \Psi_k(\boldsymbol{\xi}).$$

and apply the test function in the form of $\phi_j(\boldsymbol{x})\Psi_l(\boldsymbol{\xi})$ on (2.14) to get

$$\begin{aligned}
&\sum_{i=1}^{N_x} \sum_{k=1}^{N_\xi} u_{ik} \sum_{s=0}^{L} \int_D \sqrt{\lambda_s} \alpha_s \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \int_\Gamma \xi_s \psi_k(\boldsymbol{\xi}) \psi_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \\
&= \int_D f(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \int_\Gamma \psi_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}.
\end{aligned}$$

In the corresponding matrix-vector equation

$$\mathcal{A}\boldsymbol{u} = \boldsymbol{f}, \tag{3.3}$$

the vector $\boldsymbol{u}, \boldsymbol{f} \in \mathbb{R}^{N_x N_\xi}$ and the matrix $\mathcal{A} \in \mathbb{R}^{N_x N_\xi \times N_x N_\xi}$. This is really a huge system as $N_x N_\xi$ is very large for a typical choice of N_ξ.

Matrix $\mathcal{A}$ can be written in the tensor product form

$$\mathcal{A} = \sum_{s=0}^{L} G_s \otimes K_s = G_0 \otimes K_0 + \sum_{s=1}^{L} G_s \otimes K_s,$$

where $K_s \in \mathbb{R}^{N_x \times N_x}$ are the stiffness matrices using coefficients $\sqrt{\lambda_s}\alpha_s(\boldsymbol{x})$ and $G_s \in \mathbb{R}^{N_\xi \times N_\xi}$ are

the stochastic matrices. More precisely,

$$
\begin{aligned}
K_0(i,j) &= \int_D \mu \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \\
G_0(k,l) &= \int_\Gamma \psi_k(\boldsymbol{\xi}) \psi_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}, \\
K_s(i,j) &= \int_D \sqrt{\lambda_s} \alpha_s(\boldsymbol{x}) \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \\
G_s(k,l) &= \int_\Gamma \xi_s \psi_k(\boldsymbol{\xi}) \psi_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi},
\end{aligned}
$$

for $i, j \in \{1, 2, \cdots, N_x\}$ and $k, l \in \{1, 2, \cdots, N_\xi\}$.

The right hand side $\boldsymbol{f} = g_0 \otimes f_0$ where

$$
f_0(j) = \int_D f(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \quad g_0(l) = \int_\Gamma \psi_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi},
$$

for $j \in \{1, 2, \cdots, N_x\}$ and $l \in \{1, 2, \cdots, N_\xi\}$.

We single out the first term $G_0 \otimes K_0$ since it has simple and nice structure. With assumption (2.3), the mean $\mu(x)$ is positive and uniformly bounded below and above, and therefore K_0 is elliptic. We can chose L^2-orthonormal basis for polynomials in $\boldsymbol{\xi}$ so that G_0 is the identity matrix.

We now explore the tensor product form by reshaping the solution vector $\boldsymbol{u} \in \mathbb{R}^{N_x N_\xi}$ in (3.3) into a $N_x \times N_\xi$ matrix [17]

$$
U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1N_\xi} \\ u_{21} & u_{22} & \dots & u_{2N_\xi} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_x 1} & u_{N_x 2} & \dots & u_{N_x N_\xi} \end{pmatrix}_{N_x \times N_\xi}, \tag{3.4}
$$

then the matrix-vector equation (3.3) is equivalent to a matrix-matrix equation:

$$\mathcal{A}(U) = F, \tag{3.5}$$

where

$$\mathcal{A}(U) = \sum_{s=0}^{L} K_s U G_s^{\intercal} = K_0 U G_0^{\intercal} + \sum_{s=1}^{L} K_s U G_s^{\intercal}, \quad F = f_0 \cdot g_0^{\intercal}.$$

In the tensor product form, only K_s and G_s are stored and thus save the storage for the whole matrix $\mathcal{A}$. In the following section, we shall explore more on the structure of G_s and show that even G_s does not need to be stored explicitly.

3.3 Sparsity in Stochastic Matrices

3.3.1 Stochastic matrices

Recall that the stochastic matrices are

$$G_s(k,l) = \left(\int_{\Gamma} \xi_s \psi_k(\boldsymbol{\xi}) \psi_l(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \right)_{k,l=1,\cdots,N_\xi}, \qquad s = 1, 2, \cdots, L.$$

Note that from Karhunen-Loève expansion, $\xi_1, \ldots, \xi_L$ are only mutually uncorrelated. Only if ξ_i are Gaussian, uncorrelation implies the independence. Otherwise, we will introduce an additional assumption that $\xi_1, \cdots, \xi_L$ are independent.

With this assumption, the joint density function is in the product type

$$p(\boldsymbol{\xi}) = p_1(\xi_1) p_2(\xi_2) \cdots p_L(\xi_L).$$

By defining a map between global index $\hat{k}$ and multi-index $\boldsymbol{k} = (k_1, k_2, \cdots, k_L)$ based on the

degree of orthogonal polynomials in each ξ_i direction, we can write

$$\psi_{\hat{k}}(\boldsymbol{\xi}) = \psi_{\boldsymbol{k}}(\boldsymbol{\xi}) = \prod_{s=1}^{L} \psi_{k_s}(\xi_s).$$

Therefore G_s can be simplified as

$$\begin{aligned} G_s(\hat{k}, \hat{l}) &= \int_{\Gamma} \xi_s \psi_{\boldsymbol{k}}(\boldsymbol{\xi}) \psi_{\boldsymbol{l}}(\boldsymbol{\xi}) p(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} \\ &= \int_{\Gamma} \xi_s \left(\prod_{i=1}^{L} \psi_{k_i}(\xi_i) \psi_{l_i}(\xi_i) p(\xi_i) \right) \mathrm{d}\xi_1 \, \mathrm{d}\xi_2 \cdots \mathrm{d}\xi_L \\ &= \left(\prod_{i=1,\, i \neq s}^{L} \int_{\Gamma_i} \psi_{k_i}(\xi_i) \psi_{l_i}(\xi_i) p(\xi_i) \, \mathrm{d}\xi_i \right) \int_{\Gamma_s} \xi_s \psi_{k_s}(\xi_s) \psi_{l_s}(\xi_s) p(\xi_s) \, \mathrm{d}\xi_s. \end{aligned} \tag{3.6}$$

3.3.2 Generalized Polynomials Chaos

The generalized polynomial chaos (gPC) basis [45] $\{\psi_n(\xi)\}_{n=0}^{\infty}$ is an orthogonal system [32, 25] with respect to certain real positive measure p if the following inner product property holds:

$$\langle \psi_n, \psi_m \rangle_p \stackrel{\text{def}}{=} \int_{\mathbb{R}} \psi_n(\xi) \psi_m(\xi) p(\xi) \, \mathrm{d}\xi = \gamma_n \delta_{nm}, \quad n, m = 1, 2, \cdots$$

where $\gamma_n = \int_{\mathbb{R}} \psi_n(\xi)^2 p(\xi) \, \mathrm{d}\xi > 0$ is the normalization constant and $\delta_{nm} = 0$ if $m \neq n$ and $\delta_{nm} = 1$ if $m = n$ is the Kronecker delta function. If $\gamma_n = 1$ then $\{\psi_n(\xi)\}_{n=0}^{\infty}$ is called an orthonormal system.

For univariate orthogonal polynomials, the most important property is the three-term recurrence formula [32, 25] with the initial terms $\psi_{-1}(\xi) = 0$ and $\psi_0(\xi) = 1$:

$$\psi_{k+1}(\xi) = (a_k \xi + b_k) \psi_k(\xi) - c_k \psi_{k-1}(\xi), \quad \forall k \geqslant 1, \tag{3.7}$$

where $a_k,\ c_k$ are non-zero and $c_k a_k a_{k-1} > 0$ for all $k \geqslant 1$. This leads to

$$\xi\psi_k(\xi) = \frac{1}{a_k}(\psi_{k+1}(\xi) + c_k\psi_{k-1}(\xi)) - \frac{b_k}{a_k}\psi_k(\xi), \quad \forall k \geqslant 1. \tag{3.8}$$

We shall use orthonormal system in each direction s when computing the stochastic matrix G_s, for $s = 1, 2, \ldots, L$.

3.3.3 Tri-diagonal Blocks in Stochastic Matrices

If orthonormal polynomials are chosen, G_0 is simply identity. To study G_s for $s \geqslant 1$, consider the $(r+1) \times (r+1)$ matrix T_{r+1} with (k, l) entry ($1 \leqslant k, l \leqslant r$) using (3.8)

$$\begin{aligned} T_{r+1}(k,l) &= \int_{\mathbb{R}} \xi\psi_k(\xi)\psi_l(\xi)p(\xi)\,\mathrm{d}\xi \\ &= \int_{\mathbb{R}} \left[\frac{1}{a_k}(\psi_{k+1}(\xi) + c_k\psi_{k-1}(\xi)) - \frac{b_k}{a_k}\psi_k(\xi)\right]\psi_l(\xi)p(\xi)\,\mathrm{d}\xi. \end{aligned} \tag{3.9}$$

By orthogonality, T_{r+1} is a tri-diagonal matrix with

$$T_{r+1} = \begin{pmatrix} -b_0/a_0 & 1/a_0 & & & \\ c_1/a_1 & -b_1/a_1 & 1/a_1 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{r-1}/a_{r-1} & -b_{r-1}/a_{r-1} & 1/a_{r-1} \\ & & & c_r/a_r & -b_r/a_r \end{pmatrix}_{(r+1)\times(r+1)}. \tag{3.10}$$

It is easy to check that T_{r+1} is symmetric which leads to $1/a_k = c_{k+1}/a_{k+1}$ for all $k \geqslant 0$. Moreover, the eigenvalues of T_{r+1} are the distinct roots of the orthonormal polynomials of degree $r+1$ with respect to the density function $p(\xi)$.

Example 3.3.1. For normalized Legendre polynomials, $p(\xi) = 1/2$ with support $[-1, 1]$, and the

recurrence formula is given by

$$\frac{k+1}{\sqrt{2k+3}}\psi_{k+1}(\xi) = \sqrt{2k+1}\xi\psi_k(\xi) - \frac{k}{\sqrt{2k-1}}\psi_{k-1}(\xi), \quad k = 1, 2, \cdots, \tag{3.11}$$

with $\psi_0(\xi) = 1, \psi_1(\xi) = \sqrt{3}\xi$. So the corresponding matrix T_r is: for $r = 1, 2, \ldots, p$

$$T_{r+1} = \mathrm{diag}\left(\frac{k}{\sqrt{(2k-1)(2k+1)}},\ 0,\ \frac{k+1}{\sqrt{(2k+1)(2k+3)}}\right) \tag{3.12}$$

The eigenvalues of T_{r+1} lie in the interval $[-L_{r+1}^{\max}, L_{r+1}^{\max}]$ with $L_{r+1}^{\max} \leqslant 1$. The eigenvalues of T for different p are shown in Figure 3.1.

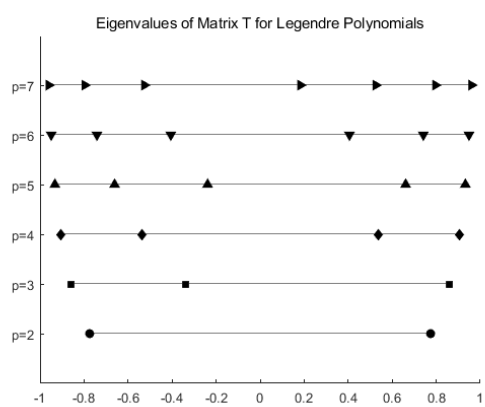


Figure 3.1: Eigenvalues of T (Legendre)

Example 3.3.2. For normalized Hermite polynomials, $p(\xi) = \frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}$, and the recurrence formula is given by

$$\sqrt{k+1}\psi_{k+1}(\xi) = \xi\psi_k(\xi) - \sqrt{k}\psi_{k-1}(\xi), \tag{3.13}$$

with $\psi_0(\xi) = 1, \psi_1(\xi) = \xi$. So the corresponding matrix T_{r+1} is:

$$T_{r+1} = \mathrm{diag}(\sqrt{k}, 0, \sqrt{k+1}) \tag{3.14}$$

The eigenvalues of T_{r+1} lie in the interval $[-H_{r+1}^{\max}, H_{r+1}^{\max}]$ with $H_{r+1}^{\max} \leqslant \sqrt{r} + \sqrt{r-1}$. The

eigenvalues of T for different p are shown in Figure 3.2.

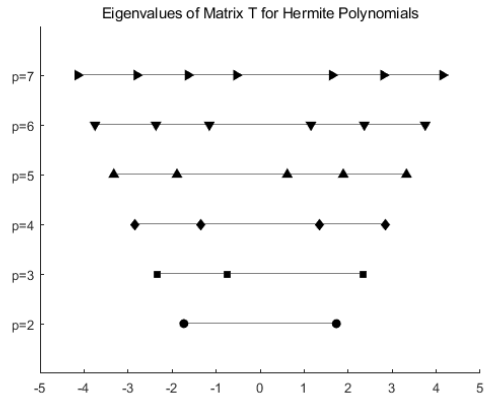


Figure 3.2: Eigenvalues of T (Hermite)

3.3.4 Simplex Lattice Representation

We explore the sparse pattern of the stochastic matrices G_s. Consider a lattice with vertices defined as

$$\mathcal{V}_{p,L}^{\mathfrak{L}} = \left\{ \boldsymbol{k} = (k_1, k_2, \cdots, k_L) \in \mathbb{Z}^L \mid 0 \leqslant k_s \leqslant p,\ s = 1, 2, \cdots, L \right\},$$

where L is the number of truncated terms in the Karhunen-Loève expansion and p is the maximum degree for each univariate orthonormal polynomials. There exists a bijection between the lattice vertices $\mathcal{V}_{p,L}^{\mathfrak{L}}$ and the tensor product polynomial space $\mathfrak{T}_{p,L}$ in (3.1) if consider each k_s in the multi-index $\boldsymbol{k}$ as the degree of univariate polynomial of $\psi_{k_s}(\xi_s)$. Alternatively, define a simplex with vertices $\mathcal{V}_{p,L}^{\mathfrak{S}}$ as a subset of $\mathcal{V}_{p,L}^{\mathfrak{L}}$:

$$\mathcal{V}_{p,L}^{\mathfrak{S}} = \left\{ \boldsymbol{k} = (k_1, k_2, \cdots, k_L) \in \mathbb{Z}^L \mid \sum_{s=1}^{L} k_s \leqslant p,\ k_s \geqslant 0,\ s = 1, 2, \cdots, L \right\}.$$

There also exists a bijection between the simplex vertices $\mathcal{V}_{p,L}^{\mathfrak{S}}$ and the complete polynomial space $\mathfrak{C}_{p,L}$ in (3.2). In view of lattice, stochastic collocation method is corresponding to the lattice vertices $\mathcal{V}_{p,L}^{\mathfrak{L}}$ with tensor product polynomial space $\mathfrak{T}_{p,L}$, while stochastic Galerkin method is corre-

sponding to the simplex vertices $\mathcal{V}_{p,L}^{\mathfrak{S}}$ with complete polynomial space $\mathfrak{C}_{p,L}$.

For any symmetric $N_\xi \times N_\xi$ matrix G, we can define a bijection between G and an undirected graph $\mathfrak{G}_{p,L} = (\mathcal{V}_{p,L}, \mathcal{E}_{p,L})$ on the lattice/simplex vertices $\mathcal{V}_{p,L}$ and their edges $\mathcal{E}_{p,L} = \{(\boldsymbol{k}, \boldsymbol{l}) \ : \ \boldsymbol{k},\ \boldsymbol{l} \in \mathcal{V}_{p,L}\}$ by setting the weight in the graph $w_{\boldsymbol{k},\boldsymbol{l}} = G(\hat{k}, \hat{l})$, where $\hat{k},\ \hat{l}$ are the global index corresponding to the two multi-indices $\boldsymbol{k},\ \boldsymbol{l}$.

Recall the three-term recurrence formula for orthonormal polynomials (3.9) we have

$$G_s(\hat{k}, \hat{l}) = \begin{cases} \dfrac{c_{k_s+1}}{a_{k_s+1}}, & \text{if } k_s = l_s - 1 \text{ and } k_i = l_i,\ i \in \{1, 2, \cdots, L\} \setminus \{s\}, \\ \dfrac{c_{k_s}}{a_{k_s}}, & \text{if } k_s = l_s + 1 \text{ and } k_i = l_i,\ i \in \{1, 2, \cdots, L\} \setminus \{s\}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.15}$$

The result above indicates that the stochastic matrices G_s are sparse. The $(\hat{k}, \hat{l})$-th entry of G_s is not zero only when their corresponding multi-index points on $\mathcal{V}_{p,L}^{\mathfrak{S}}$ are adjacent in the s-th coordinate, i.e. $|k_s - l_s| = 1$ and $k_i = l_i,\ i \in \{1, 2, \cdots, L\} \setminus \{s\}$.

In the sum $G = \sum_{s=1}^{L} G_s$, $G(\hat{k}, \hat{l})$ is non-zero only when $\|\boldsymbol{k} - \boldsymbol{l}\|_1 = 1$, where $\|\cdot\|_1$ is the l_1-norm of vectors. Therefore, we can define a simplex graph $\mathfrak{S}_{p,L} = \left(\mathcal{V}_{p,L}^{\mathfrak{S}}, \mathcal{E}_{p,L}^{\mathfrak{S}}\right)$ on the simplex vertices $\mathcal{V}_{p,L}^{\mathfrak{S}}$ and the edges

$$\mathcal{E}_{p,L}^{\mathfrak{S}} = \left\{(\boldsymbol{k}, \boldsymbol{l}) \mid \boldsymbol{k},\ \boldsymbol{l} \in \mathcal{V}_{p,L}^{\mathfrak{S}},\ \|\boldsymbol{k} - \boldsymbol{l}\|_1 = 1\right\}.$$

Figure 3.3 is an example for the simplex graph $\mathfrak{S}_{p,L}$ with normalized Hermite polynomials when $p = 5$ and $L = 2$. The black points are those with even total degrees while the white points are those with odd total degrees. The black points are only adjacent (connected) to white points and vice versa. For example, point #13 (corresponding to $\psi_2(\xi_1)\psi_1(\xi_2)$ with total degree three) is connected to point #2, #3, #6, #7, all with even total degree. Moreover, the entries of $G(13, \cdot)$ are zero except $G(13, 7) = G_1(13, 7) = \sqrt{2}$, $G(13, 6) = G_1(13, 6) = \sqrt{3}$, $G(13, 2) = G_2(13, 2) = 1$

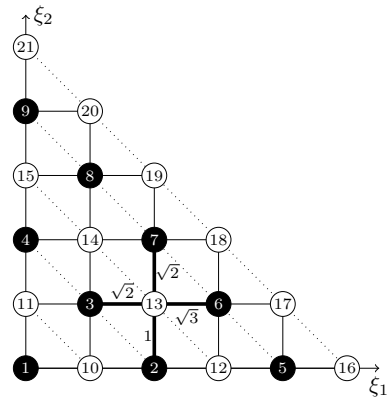


Figure 3.3: The simplex lattice and corresponding graph of stochastic matrix G with parameters $p = 5$ and $L = 2$.

and $G(13, 3) = G_2(13, 3) = \sqrt{2}$.

Due to the lattice structure, the vertices of the simplex graph $\mathfrak{S}_{p,L}$ are well-separated into two groups: odd total degree and even total degree. Inside each group, the vertices are disjointed. More precisely, define

$$\mathcal{I}_{\text{odd}} = \left\{ \boldsymbol{k} \in \mathcal{V}_{p,L}^{\mathfrak{S}} \mid \sum_{s=1}^{L} k_s \text{ is odd} \right\}, \quad \mathcal{I}_{\text{even}} = \left\{ \boldsymbol{k} \in \mathcal{V}_{p,L}^{\mathfrak{S}} \mid \sum_{s=1}^{L} k_s \text{ is even} \right\}.$$

Then $G = \sum_{s=1}^{L} G_s$ admits a block matrix structure (after suitable reordering)

$$G = \begin{pmatrix} O & B^{\intercal} \\ B & O \end{pmatrix}, \text{ with } B = G(\mathcal{I}_{\text{odd}}, \mathcal{I}_{\text{even}}).$$

We will use this property to explore more on the spectrum of the preconditioned system in the next chapter.

Figure 3.4 is an example for the sparsity of matrix $\mathcal{A}$ on the simplex graph $\mathcal{G}_{p,L}$ with normalized Hermite polynomials when $p = 4$ and $L = 4$. We will use this property to design a block triangular preconditioner in Section 4.3.

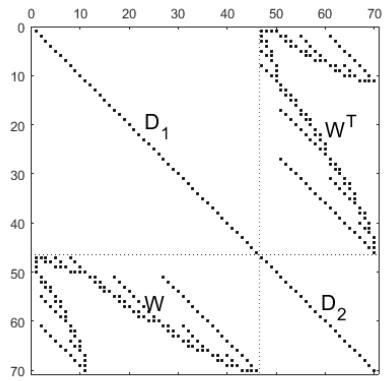


Figure 3.4: Block sparsity structure of the Galerkin matrix $\mathcal{A}$ under the odd-even ordering of their total degree ($L = 4$, $p = 4$).

Given $k_i,\ i \in \{1, 2, \cdots, L\} \setminus \{s\}$ such that $(\sum_{i \neq s} k_i) + r = p$ for some $r \geqslant 0$, we can define a "line" parallel to the s-th coordinate whose multi-indices are

$$
\begin{aligned}
&(k_1, \cdots, k_{s-1}, 0, k_{s+1}, \cdots, k_L),\\
&(k_1, \cdots, k_{s-1}, 1, k_{s+1}, \cdots, k_L),\\
&\qquad\qquad \vdots\\
&(k_1, \cdots, k_{s-1}, r, k_{s+1}, \cdots, k_L).
\end{aligned}
$$

Along this line, the corresponding submatrix in G_s is simply the tri-diagonal matrix T_r defined in (3.14). It implies that a block diagonal matrix can be obtained by reordering G_s according to the lines parallel to the s-th coordinate, i.e. for given $s \in \{1, 2, \cdots, L\}$, there exists a permutation vector J_s (reordering of $\{1, 2, \cdots, N_\xi\}$) such that $G_s(J_s)$ is a block-diagonal matrix where each block is a tri-diagonal matrix $T_r,\ r \in \{0, 1, \cdots, p\}$ in (3.14). Moreover, one can even properly choose the permutation (reordering) vector J_s such that the resulting block-diagonal matrix $G_s(J_s, J_s)$ is the same along different directions, provided the joint density function is isotropic, i.e. $p_1(\xi_1) = p_2(\xi_2) = \cdots = p_L(\xi_L)$. A simple example for $L = 3$ and $p = 3$ is given below in Figure 3.5. Here $N_\xi = \binom{3+3}{3} = 20$. Without loss of generality, we take $s = 1$. The multi-index $\boldsymbol{k}$ and $\boldsymbol{l}$ should be matched in all directions except ξ_1 direction. And the corresponding stochastic

matrix G_1 is

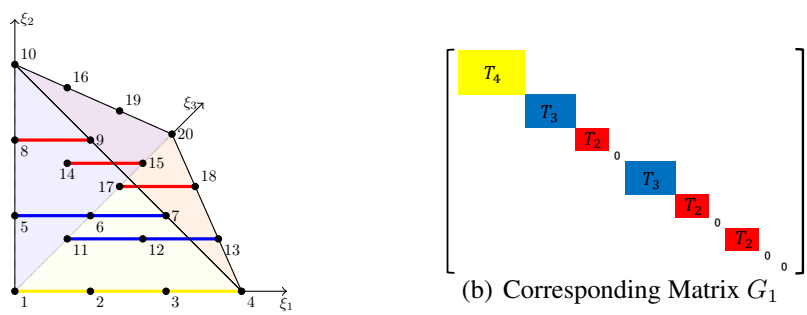
$$G_1 = \text{diag}(T_4, T_3, T_2, 0, T_3, T_2, 0, T_2, 0, 0) \in \mathbb{R}^{20\times 20}.$$

Moreover, if we choose

$$J_2 = (1, 5, 8, 10, 11, 14, 16, 17, 19, 20, 2, 6, 9, 12, 15, 18, 3, 7, 13, 4)^\intercal$$

$$J_3 = (1, 11, 17, 20, 2, 12, 18, 3, 13, 4, 5, 14, 19, 6, 15, 7, 8, 16, 9, 10)^\intercal$$

Then $G_2(J_2, J_2) = G_3(J_3, J_3) = G_1$.



(a) Simplex Representation

(b) Corresponding Matrix G_1

Figure 3.5: A simplex representation with $L = 3$ and $p = 3$ (left) and the corresponding G_1 matrix (right)

Therefore, the sparse matrices G_s need not be formed explicitly. Instead, only the block tridiagonal matrices T_{r+1} for $r \in \{1, 2, \cdots, p+1\}$ and the reordering indices $J_s \in \mathcal{R}^{N_\xi},\ s \in \{1, 2, \cdots, L\}$ are stored in order to save the storage. In the matrix multiplication $UG_s^\intercal$ in (3.5), one only need to compute the matrix multiplication of T_r and the corresponding lines in U with respect to the reordering indices J_s.

Chapter 4

Preconditioning

The stochastic system (3.3) is large and ill-conditioned and hence a preconditioner is required. Pellissetti [30] and Ghanem [22] proposed a mean-based preconditioner $\mathcal{P}_0 = G_0 \otimes K_0$. Powell and Elman [33] then provided detailed spectral analysis on the preconditioned system $\mathcal{P}_0^{-1}\mathcal{A}$. In this chapter we first provide a tighter bound for $\mathcal{P}_0^{-1}\mathcal{A}$ and a symmetric analysis on the eigenvalues that the maximum and minimum eigenvalues of $\mathcal{P}_0^{-1}\mathcal{A}$ add up to 2 [10]. Then we present a new variance-involved block diagonal preconditioner $\mathcal{P}_1 = G_0 \otimes K_0 + G_1 \otimes K_1$ and give corresponding spectral analysis. A block triangular preconditioner $\mathcal{P}_T$ associated with the odd-even indices in Section 3.3.4 is then introduced together with its convergence analysis [9]. At the end of this chapter, a Kronecker product preconditioner $\mathcal{P}_K$ introduced by Ullmann [42] is reviewed, which is a single tensor product minimizer of $\mathcal{A}$ with the Frobenius norm.

4.1 A Block Diagonal Preconditioner using Mean Field

Pellissetti [30] and Ghanem [22] proposed a natural mean-based preconditioner $\mathcal{P}_0 = G_0 \otimes K_0$, which is the diagonal block of the system matrix $\mathcal{A}$. Powell and Elman [33] then provided a

detailed spectral analysis on the preconditioned system $\mathcal{P}_0^{-1}\mathcal{A}$.

We shall first recall results in [33]. We begin with the estimate on the generalized eigenvalues of K_s relative to K_0. Recall that the kernel could be normalized such that the variance σ^2 is separated out with property $\sum_{s=1}^{\infty}\lambda_s = 1$. The eigen-functions $\alpha_s(x)$ may not be positive. Its $L^2(D)$ norm is unit but the L^∞-norm, $\|\alpha_s\|_\infty$ may not be uniformly bounded.

Lemma 4.1.1 (Lemma 3.4 in [33])**.** *Let* $\beta_s = \sqrt{\lambda_s}\|\alpha_s\|_\infty$ *for* $s = 1, 2, \cdots, L$, *then*

$$\rho(K_0^{-1}K_s) \leq \frac{\sigma}{\mu}\beta_s.$$

We then recall the spectral bound obtained for the block-diagonal preconditioner $\mathcal{P}_0 = G_0 \otimes K_0$.

Theorem 4.1.2 (Theorem 3.8 in [33])**.**

$$1 - \tau \leqslant \lambda_{\min}(\mathcal{P}_0^{-1}\mathcal{A}) \leqslant \lambda_{\max}(\mathcal{P}_0^{-1}\mathcal{A}) \leqslant 1 + \tau,$$

where

$$\tau = \frac{\sigma}{\mu}C_{p+1}^{\max}\sum_{s=1}^{L}\beta_s,$$

$$C_{p+1}^{\max} = \begin{cases} H_{p+1}^{\max} = \sqrt{p-1} + \sqrt{p}, & \textit{for Hermite polynomials,} \\ L_{p+1}^{\max} = 1, & \textit{for Legendre polynomials.} \end{cases}$$

4.1.1 A tighter bound for the preconditioned system

Define $\mathcal{N}(\boldsymbol{k}) = \left\{\boldsymbol{l} \in \mathcal{V}_{p,L}^{\mathfrak{S}} \mid \|\boldsymbol{k} - \boldsymbol{l}\|_1 = 1\right\}$ as the neighbors of vertex $\boldsymbol{k}$, and $\mathcal{N}(\boldsymbol{k}, s) = \{\boldsymbol{l} \in \mathcal{N}(\boldsymbol{k}) \mid |k_s - l_s| = 1\}$ as the neighbors of vertex $\boldsymbol{k}$ in s-direction. Further define $s(\boldsymbol{k}, \boldsymbol{l}) = \left\{s \in \{1, 2, \cdots, L\} \mid (\boldsymbol{k}, \boldsymbol{l}) \in \mathcal{E}_{p,L}^{\mathfrak{S}} \text{ and } |k_s - l_s| = 1\right\}$ as the difference direction between indices $\boldsymbol{k}$ and $\boldsymbol{l}$. Based on the simplex structure of the stochastic matrix, the above result can be improved to the following theorem.

Theorem 4.1.3 (New Spectral Bound for $\mathcal{P}_0$)**.**

$$1-\hat{\tau} \leqslant \lambda_{\min}(\mathcal{P}_0^{-1}\mathcal{A}) \leqslant \lambda_{\max}(\mathcal{P}_0^{-1}\mathcal{A}) \leqslant 1+\hat{\tau}, \tag{4.1}$$

where

$$\hat{\tau} = \frac{\sigma}{\mu} \max_{\boldsymbol{k},|\boldsymbol{k}|\leqslant p} \sum_{s=1}^{L} \sum_{\boldsymbol{l}\in\mathcal{N}(\boldsymbol{k},s)} G(\boldsymbol{k},\boldsymbol{l})\beta_s, \quad G = \sum_{s=1}^{L} G_s.$$

In order to prove (4.1), we introduce the Gershgorin's Theorem for block matrices [20, 37, 41]. For completeness, a proof is included here.

Theorem 4.1.4 (Gershgorin's Theorem for block matrices [20, 37, 41])**.** *Consider* $\bar{\mathcal{A}} = (\bar{A}_{ij}) \in \mathbb{R}^{dn\times dn}$ *where* $\bar{A}_{ij} \in \mathbb{R}^{d\times d}$*. Denote by* $\sigma(\cdot)$ *the spectrum of a matrix. Define*

$$\mathcal{G}_i = \sigma(\bar{A}_{ii}) \bigcup \left\{ \lambda \notin \sigma(\bar{A}_{ii}) : \left\|(\bar{A}_{ii}-\lambda I)^{-1}\right\|^{-1} \leqslant \sum_{j\neq i} \|\bar{A}_{ij}\| \right\}, \tag{4.2}$$

for $i = 1, 2, \cdots n$*. Then*

$$\sigma(\bar{\mathcal{A}}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}_i.$$

Proof. Suppose that $\lambda \notin \bigcup_{i=1}^{n} \sigma(\bar{A}_{ii})$ then we have

$$\bar{\mathcal{A}} - \lambda I = \begin{pmatrix} \bar{A}_{11}-\lambda I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \bar{A}_{nn}-\lambda I \end{pmatrix} (I + M(\lambda)), \tag{4.3}$$

where

$$M(\lambda) = \begin{pmatrix} 0 & (\bar{A}_{11}-\lambda I)^{-1}\bar{A}_{12} & \cdots & (\bar{A}_{11}-\lambda I)^{-1}\bar{A}_{1n} \\ (\bar{A}_{22}-\lambda I)^{-1}\bar{A}_{21} & 0 & \cdots & (\bar{A}_{22}-\lambda I)^{-1}\bar{A}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ (\bar{A}_{nn}-\lambda I)^{-1}\bar{A}_{n1} & \cdots & \cdots & 0 \end{pmatrix}.$$

Assume $\|(\bar{A}_{ii} - \lambda I)^{-1}\|^{-1} > \sum_{j\neq i} \|\bar{A}_{ij}\|$ for $i = 1, 2, \cdots, n$. Then $\|M(\lambda)\| < 1$ and hence $I + M(\lambda)$ is not singular. It implies that $\bar{\mathcal{A}} - \lambda I$ is not singular which leads to a contradiction. Therefore λ must be in the given region. $\square$

Now we give a proof of Theorem 4.1.3.

Proof. Since $G_0 = I$ and K_0 is positive definite, we can define $\mathcal{P}_0^{\frac{1}{2}} = I \otimes K_0^{\frac{1}{2}}$ and then

$$\sigma(\mathcal{P}_0^{-1}\mathcal{A}) = \sigma(\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}).$$

It suffices to show the following bound for the symmetric matrix $\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}$

$$1 - \hat{\tau} \leqslant \lambda_{\min}(\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}) \leqslant \lambda_{\max}(\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}) \leqslant 1 + \hat{\tau}.$$

Let

$$\bar{\mathcal{A}} = \mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}} = (I \otimes K_0^{-\frac{1}{2}}) \left(\sum_{s=0}^{L} G_s \otimes K_s \right) (I \otimes K_0^{-\frac{1}{2}}) = I + \sum_{s=1}^{L} G_s \otimes \left(K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}} \right).$$

The diagonal blocks $\bar{A}_{ii} = I$. For a scalar index, we shall use the boldface letter to denote its corresponding multi-index, and the structure is easier to explore in the multi-index system, cf. Fig. 3.3. Using these notation, together with $G = \sum_{s=1}^{L} G_s$, we have for $i \neq j$,

$$\bar{A}_{ij} = \sum_{s=1}^{L} (G_s)_{ij} \left(K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}} \right) = G(\boldsymbol{i}, \boldsymbol{j}) \left(K_0^{-\frac{1}{2}} K_{s(\boldsymbol{i},\boldsymbol{j})} K_0^{-\frac{1}{2}} \right).$$

Therefore, (4.2) can be written as

$$
\begin{aligned}
\mathcal{G}_i &= \{1\} \bigcup \left\{ \lambda \neq 1 : |1-\lambda| \leqslant \sum_{j \neq i} \left\| G(\boldsymbol{i},\boldsymbol{j}) \left(K_0^{-\frac{1}{2}} K_{s(\boldsymbol{i},\boldsymbol{j})} K_0^{-\frac{1}{2}} \right) \right\| \right\} \\
&= \left\{ \lambda : |1-\lambda| \leqslant \sum_{\boldsymbol{j} \in \mathcal{N}(\boldsymbol{i})} G(\boldsymbol{i},\boldsymbol{j}) \left\| K_0^{-\frac{1}{2}} K_{s(\boldsymbol{i},\boldsymbol{j})} K_0^{-\frac{1}{2}} \right\| \right\} \\
&= \left\{ \lambda : |1-\lambda| \leqslant \sum_{s=1}^{L} \sum_{\boldsymbol{j} \in \mathcal{N}(\boldsymbol{i},s)} G(\boldsymbol{i},\boldsymbol{j}) \left\| K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}} \right\| \right\} \\
&\subseteq \left\{ \lambda : |1-\lambda| \leqslant \frac{\sigma}{\mu} \sum_{s=1}^{L} \sum_{\boldsymbol{j} \in \mathcal{N}(\boldsymbol{i},s)} G(\boldsymbol{i},\boldsymbol{j}) \beta_s \right\}
\end{aligned}
\tag{4.4}
$$

Since all $\mathcal{G}_i$ are concentric at 1, we can conclude that

$$
\sigma(\mathcal{P}_0^{-\frac{1}{2}} \mathcal{A} \mathcal{P}_0^{-\frac{1}{2}}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}_i \subseteq D\left(1, \hat{\tau}\right),
$$

where $D\left(1, \hat{\tau}\right)$ is the disk centered at 1 with radius $\hat{\tau}$. And so is $\sigma(\mathcal{P}_0^{-1}\mathcal{A})$. □

We claim that $\hat{\tau}$ in Theorem 4.1.3 is indeed an improvement of τ in Theorem 4.1.2, i.e. $\hat{\tau} \leqslant \tau$. In each direction $s \in \{1, 2, \cdots, L\}$, for any $\boldsymbol{k}$ given, there are at most two vertices $\boldsymbol{i}, \boldsymbol{j}$ in $\mathcal{N}(\boldsymbol{k}, s)$, cf. Fig. 3.3. Therefore

$$
\sum_{\boldsymbol{l} \in \mathcal{N}(\boldsymbol{k},s)} G(\boldsymbol{k},\boldsymbol{l}) \beta_s \leq C_{p+1}^{\max} \beta_s,
$$

and the claim holds by simply summing up all the directions. The above equality holds only when $k_s = p-1$ and all other $k_i = 0,\ i \neq s$. This implies that $\hat{\tau}$ in Theorem 4.1.3 is a much tighter bound compared to τ in Theorem 4.1.2, and $\tau = \hat{\tau}$ only when $L = 1$.

Since $\mathcal{P}_0$ is symmetric positive definite, we can write a symmetric version of the preconditioned

system as

$$\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}} = \sum_{s=0}^{L} G_s \otimes (K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}}) = I + \sum_{s=1}^{L} G_s \otimes (K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}}) = \begin{bmatrix} I & B^\intercal \\ B & I \end{bmatrix}$$

where $B = \sum_{s=1}^{L} G_s(\mathcal{I}_2, \mathcal{I}_1) \otimes (K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}})$ is the off-diagonal of the preconditioned system $\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}$ corresponding to the odd-even ordering described before based on the total degrees in the stochastic space. In the proof of Theorem 4.1.3, with the help of Gershgorin's Theorem for block matrices [20, 37], we can get that the off-diagonal blocks of the preconditioned system satisfies

$$\left\| \begin{bmatrix} 0 & B^\intercal \\ B & 0 \end{bmatrix} \right\| = \hat{\tau}. \tag{4.5}$$

We will use this fact in the convergence analysis of the block-triangular preconditioner in Section 4.3.

4.1.2 Symmetry for the Spectrum of the Preconditioned System

We first show the eigenvalues of G is symmetric with respect to 0.

Lemma 4.1.5. *For matrix $G = \sum_{s=1}^{L} G_s$, if $\lambda \in \sigma(G)$, then $-\lambda \in \sigma(G)$. Consequently $0 \in \sigma(G)$ with multiplicity $\mathrm{abs}(|\mathcal{I}_{\mathrm{odd}}| - |\mathcal{I}_{\mathrm{even}}|)$.*

Proof. Since each G_s is symmetric, so is G. Consequently all eigenvalues of G are real. Recall that the block matrix structure

$$G = \begin{pmatrix} O & B^\intercal \\ B & O \end{pmatrix}, \text{ with } B = G(\mathcal{I}_{\mathrm{odd}}, \mathcal{I}_{\mathrm{even}}).$$

Suppose $\langle \lambda, (\boldsymbol{x}_{\text{odd}}, \boldsymbol{x}_{\text{even}}^{\intercal}) \rangle$ is an eigen-pair of matrix G, i.e.

$$B^{\intercal}\boldsymbol{x}_{\text{even}} = \lambda \boldsymbol{x}_{\text{odd}}, \quad B\boldsymbol{x}_{\text{odd}} = \lambda \boldsymbol{x}_{\text{even}},$$

which is equivalent to

$$B^{\intercal}(-\boldsymbol{x}_{\text{even}}) = (-\lambda)\boldsymbol{x}_{\text{odd}}, \quad B\boldsymbol{x}_{\text{odd}} = (-\lambda)(-\boldsymbol{x}_{\text{even}}),$$

namely $\langle -\lambda, (\boldsymbol{x}_{\text{odd}}, -\boldsymbol{x}_{\text{even}})^{\intercal} \rangle$ is also an eigen-pair of G.

As B has at most $\min\{|\mathcal{I}_{\text{odd}}|, |\mathcal{I}_{\text{even}}|\}$ number of non-zero singular values, we conclude $0 \in \sigma(G)$ with multiplicity $\text{abs}(|\mathcal{I}_{\text{odd}}| - |\mathcal{I}_{\text{even}}|)$. □

Next we give the symmetric result for the preconditioned system $\mathcal{P}_0^{-1}\mathcal{A}$.

Theorem 4.1.6. *The spectrum of $\mathcal{P}_0^{-1}\mathcal{A}$ is symmetric with respect to 1. Consequently the maximum and minimum eigenvalues for matrix $\mathcal{P}_0^{-1}\mathcal{A}$ add up to 2 and the eigenvalue 1 has multiplicity $N_x \cdot \text{abs}(|\mathcal{I}_{\text{odd}}| - |\mathcal{I}_{\text{even}}|)$.*

Proof. Again we consider the symmetric matrix $\bar{A} = \mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}$ and show its eigenvalues are symmetric with respect to 1. Since

$$\mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}} = I + \sum_{s=1}^{L} G_s \otimes \left(K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}}\right),$$

by a similar proof in Theorem 4.1.3, we get for $j \neq i$ the block

$$\bar{A}_{ij} = \sum_{s=1}^{L} (G_s)_{ij} \left(K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}}\right) = G(\boldsymbol{i}, \boldsymbol{j}) \left(K_0^{-\frac{1}{2}} K_{s(\boldsymbol{i},\boldsymbol{j})} K_0^{-\frac{1}{2}}\right). \tag{4.6}$$

Therefore, $\bar{A} - I$ is a block version of the matrix G in Lemma 4.1.5 with each block is of size $N_x \times N_x$ and defined in (4.6). By Lemma 4.1.5, we conclude that the spectrum of $\bar{A} - I$ is

symmetric with respect to 0, and the eigenvalue 0 has multiplicity $N_x \cdot \mathrm{abs}(|\mathcal{I}_{\mathrm{odd}}| - |\mathcal{I}_{\mathrm{even}}|)$. Thus the spectrum of $\mathcal{P}_0^{-1}\mathcal{A}$, which is the same as the spectrum of $\bar{A} = \mathcal{P}_0^{-\frac{1}{2}}\mathcal{A}\mathcal{P}_0^{-\frac{1}{2}}$, is symmetric with respect to 1, and the eigenvalue 1 has multiplicity $N_x \cdot \mathrm{abs}(|\mathcal{I}_{\mathrm{odd}}| - |\mathcal{I}_{\mathrm{even}}|)$. $\square$

Remark 4.1.7. Having eigenvalue 1 with multiplicity $N_x \cdot \mathrm{abs}(|\mathcal{I}_{\mathrm{odd}}| - |\mathcal{I}_{\mathrm{even}}|)$ is advantageous when applying the Krylov subspace iterative methods as one iteration is enough to take care of these well centered eigenvalues.

4.2 A New Block Diagonal Preconditioner

The limitation of $\mathcal{P}_0 = G_0 \otimes K_0$ is that it only contains the mean of the stochastic fields while the variance is not involved. Here we provide a new preconditioner that contains the information on both the mean and part of the variance of the random fields, i.e.

$$\mathcal{P}_1 = G_0 \otimes K_0 + G_1 \otimes K_1.$$

Due to the exponential decay on the eigenvalues of KL-expansion, $G_1 \otimes K_1$ is able to capture some of the information in the variance.

In order to compute $\mathcal{P}_1^{-1}$, we have to diagonalize G_0 and G_1 at the same time. Consider the eigen-decomposition of the generalized eigenvalue problem

$$G_1 y = \tilde{\lambda} G_0 y.$$

Note that the eigenvalue $\tilde{\lambda}$ has nothing to do with the eigenvalues $\lambda_s,\ s = 1, 2, \cdots, L$ mentioned in K-L expansion. The eigen-function can simultaneously diagonalize G_0 and G_1. We introduce a

double orthogonal basis [19, 2] $\{\tilde{\psi}_k\}_{k=0}^{p}$ in the ξ_1 direction, i.e.

$$\int_{\Gamma_1} \tilde{\psi}_{k_1}(\xi_1)\tilde{\psi}_{l_1}(\xi_1)p(\xi_1)\,d\xi_1 = \delta_{kl}, \quad \int_{\Gamma_1} \xi_1\tilde{\psi}_{k_1}(\xi_1)\tilde{\psi}_{l_1}(\xi_1)p(\xi_1)\,d\xi_1 = \tilde{\lambda}_k\delta_{kl},$$

where $G_1 = Q^\intercal\tilde{\Lambda}Q$, $\tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_{N_\xi})$ are the eigenvalues of G_1 and Q is the orthonormal matrix consists of the eigenvectors of G_1. We have

$$\begin{aligned}
\mathcal{P}_1 &= I \otimes K_0 + G_1 \otimes K_1 = (Q^\intercal Q) \otimes K_0 + (Q^\intercal\tilde{\Lambda}Q) \otimes K_1 \\
&= (Q^\intercal \otimes I)(I \otimes K_0)(Q \otimes I) + (Q^\intercal \otimes I)(\tilde{\Lambda} \otimes K_1)(Q \otimes I) \\
&= (Q^\intercal \otimes I)(\mathrm{diag}(K_0 + \tilde{\lambda}_k K_1))(Q \otimes I) \\
&\overset{\text{def}}{=} (Q^\intercal \otimes I)\tilde{\mathcal{P}}_1(Q \otimes I)
\end{aligned}$$

Then $\mathcal{P}_1$ can be inverted by

$$\mathcal{P}_1^{-1} = (Q \otimes I)\tilde{\mathcal{P}}_1^{-1}(Q^\intercal \otimes I).$$

Similar to the diagonal precondtitioner $\mathcal{P}_0$, we give a bound for the eigenvalues of $\mathcal{P}_1^{-1}\mathcal{A}$. It suffices to prove the bound for the eigenvalues of $\tilde{\mathcal{P}}_1^{-1}\tilde{\mathcal{A}}$ where $\tilde{\mathcal{P}}_1 = G_0 \otimes K_0 + \tilde{\Lambda} \otimes K_1$ and $\tilde{\mathcal{A}} = G_0 \otimes K_0 + \tilde{\Lambda} \otimes K_1 + \sum_{s=2}^{L} G_s \otimes K_s$.

Theorem 4.2.1. *Assume $K_0 + \tilde{\lambda}_i K_1$ are non-singular for all $\tilde{\lambda}_i \in \tilde{\Lambda}$. Then*

$$1 - \tilde{\tau} \leqslant \lambda_{\min}(\tilde{\mathcal{P}}_1^{-1}\tilde{\mathcal{A}}) \leqslant \lambda_{\max}(\tilde{\mathcal{P}}_1^{-1}\tilde{\mathcal{A}}) \leqslant 1 + \tilde{\tau},$$

where

$$\tilde{\tau} = \max_{\boldsymbol{k},|k|\leqslant p} \sum_{s=2}^{L} \sum_{\boldsymbol{l}\in\mathcal{N}(\boldsymbol{k},s)} G(\boldsymbol{k},\boldsymbol{l})\tilde{\beta}_s, \quad \tilde{\beta}_s = \max_{\tilde{\lambda}_i\in\tilde{\Lambda}} \left\|(K_0 + \tilde{\lambda}_i K_1)^{-1}K_s\right\|.$$

Proof. Let

$$\bar{A} = \tilde{\mathcal{P}}_1^{-1}\mathcal{A} = I + \sum_{s=2}^{L}(G_0 \otimes K_0 + \tilde{\Lambda} \otimes K_1)^{-1}(G_s \otimes K_s).$$

Since $G_0 \otimes K_0 + \tilde{\Lambda} \otimes K_1$ is block diagonal, the diagonal blocks $\bar{A}_{ii} = I$ and for $i \neq j$

$$\bar{A}_{ij} = \sum_{s=2}^{L} (G_s)_{ij} \left((K_0 + \tilde{\lambda}_i K_1)^{-1} K_s \right) = \tilde{G}(\boldsymbol{i}, \boldsymbol{j}) \left((K_0 + \tilde{\lambda}_i K_1)^{-1} K_{s(\boldsymbol{i},\boldsymbol{j})} \right),$$

here $\tilde{G} = \sum_{s=2}^{L} G_s$. Then

$$\begin{aligned}
\mathcal{G}_i &= \{1\} \bigcup \left\{ \lambda \neq 1 : |1-\lambda| \leqslant \sum_{j \neq i} \tilde{G}(\boldsymbol{i}, \boldsymbol{j}) \left\| (K_0 + \tilde{\lambda}_i K_1)^{-1} K_{s(\boldsymbol{i},\boldsymbol{j})} \right\| \right\} \\
&= \left\{ \lambda : |1-\lambda| \leqslant \sum_{\boldsymbol{j} \in \mathcal{N}(\boldsymbol{i})} \tilde{G}(\boldsymbol{i}, \boldsymbol{j}) \left\| (K_0 + \tilde{\lambda}_i K_1)^{-1} K_{s(\boldsymbol{i},\boldsymbol{j})} \right\| \right\} \\
&= \left\{ \lambda : |1-\lambda| \leqslant \sum_{s=2}^{L} \sum_{\boldsymbol{j} \in \mathcal{N}(\boldsymbol{i},s)} G(\boldsymbol{i}, \boldsymbol{j}) \left\| (K_0 + \tilde{\lambda}_i K_1)^{-1} K_s \right\| \right\} \\
&\subseteq \left\{ \lambda : |1-\lambda| \leqslant \sum_{s=2}^{L} \sum_{\boldsymbol{j} \in \mathcal{N}(\boldsymbol{i},s)} G(\boldsymbol{i}, \boldsymbol{j}) \tilde{\beta}_s \right\}
\end{aligned} \tag{4.7}$$

Since all $\mathcal{G}_i$ are concentric at 1, we can conclude that

$$\sigma(\tilde{\mathcal{P}}_1^{-1} \tilde{\mathcal{A}}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}_i \subseteq D \left(1, \max_{\boldsymbol{k}, |k| \leqslant p} \sum_{s=2}^{L} \sum_{\boldsymbol{l} \in \mathcal{N}(\boldsymbol{k},s)} G(\boldsymbol{k}, \boldsymbol{l}) \tilde{\beta}_s \right).$$

□

Theorem 4.2.2. *Assume $K_0 + \tilde{\lambda}_i K_1$ are positive definite for all $\tilde{\lambda}_i \in \tilde{\Lambda}$. Then the spectrum of $\mathcal{P}_1^{-1}\mathcal{A}$ is symmetric with respect to 1. Consequently the maximum and minimum eigenvalues for matrix $\mathcal{P}_1^{-1}\mathcal{A}$ add up to 2 and the eigenvalue 1 has multiplicity $N_x \cdot \mathrm{abs}(|\mathcal{I}_{\mathrm{odd}}| - |\mathcal{I}_{\mathrm{even}}|)$.*

Proof. The proof is exactly the same as Theorem 4.1.6 via substituting $\mathcal{P}_0$ by $\tilde{\mathcal{P}}_1$ and rewriting $\bar{A}$ as $\tilde{\mathcal{P}}_1^{-\frac{1}{2}} \mathcal{A} \tilde{\mathcal{P}}_1^{-\frac{1}{2}}$. □

4.3 A Block Triangular Preconditioner

4.3.1 Block-triangular preconditioner

Due to the block structure discussed in 3.3.4, the matrix $\mathcal{A}$ can be written as the following block structures:

$$\mathcal{A} = \begin{bmatrix} D_1 & W^T \\ W & D_2 \end{bmatrix}, \tag{4.8}$$

where

$$D_1 = I_1 \otimes K_0, \quad D_2 = I_2 \otimes K_0, W = \left(\sum_{s=1}^{L} G_s \otimes K_s \right) (\mathcal{I}_2, \mathcal{I}_1) = \sum_{s=1}^{L} G_s(\mathcal{I}_2, \mathcal{I}_1) \otimes K_s,$$

and $\mathcal{I}_1$ is the one in $\{\mathcal{I}_{odd}, \mathcal{I}_{even}\}$ with larger cardinality, $\mathcal{I}_2$ is the one with smaller cardinality.

We define the block triangular preconditioner [9]

$$\mathcal{P}_T := \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix}. \tag{4.9}$$

The corresponding preconditioner system

$$\begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix} \begin{bmatrix} \tilde{U}_1 \\ \tilde{U}_2 \end{bmatrix} = \begin{bmatrix} \tilde{F}_1 \\ \tilde{F}_2 \end{bmatrix}$$

may be solved inexactly by the standard multigrid V-cycle.

Remark 4.3.1. *The block triangular preconditioner $\mathcal{P}_T$ may also be motivated by considering the*

block LU factorization

$$\mathcal{A} = \begin{bmatrix} D_1 & 0 \\ W & S \end{bmatrix} \begin{bmatrix} I & D_1^{-1}W^\intercal \\ 0 & I \end{bmatrix}, \quad S = D_2 - WD_1^{-1}W^\intercal. \tag{4.10}$$

It is known that with the "ideal" block triangular preconditioner

$$\tilde{\mathcal{P}}_T = \begin{bmatrix} D_1 & 0 \\ W & S \end{bmatrix}$$

the GMRes method converges in at most two iterations. However $\tilde{\mathcal{P}}_T$ is impractical because the Schur complement S is computationally expensive to invert. Replacing S by D_2 in $\tilde{\mathcal{P}}_T$ results in the block triangular preconditioner $\mathcal{P}_T$.

Since $\mathcal{P}_T$ is nonsymmetric, we can use it with the GMRes method [36] or GPCG method [4]. To apply the standard PCG method, we may consider the block symmetric Gauss-Seidel method as the preconditioner, i.e.

$$\mathcal{P}_S := \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}^{-1} \begin{bmatrix} D_1 & W^\intercal \\ 0 & D_2 \end{bmatrix}.$$

It is clear that $\mathcal{P}_S$ is SPD and the standard PCG method is guaranteed to converge. However it might not be efficient as one more matrix-vector multiplication is computed compared with the block-triangular preconditioner while the iteration steps do not decrease much.

4.3.2 Convergence Analysis

In this subsection, we give eigenvalue bounds for the matrix preconditioned by the block triangular preconditioner $\mathcal{P}_T$. We note that the spectral properties of block triangular preconditioner for

saddle point problems have been studied in [27, 38, 1].

Recall that in (4.8) and (4.9), according to whether the total degree on the simplex points are odd or even, the matrix A and $\mathcal{P}_T$ can be reordered as the following:

$$\mathcal{A} = \begin{bmatrix} D_1 & W^\intercal \\ W & D_2 \end{bmatrix}, \quad \mathcal{P}_T = \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix}.$$

By direct computation,

$$\mathcal{P}_T^{-1} = \begin{bmatrix} D_1^{-1} & 0 \\ -D_2^{-1} W D_1^{-1} & D_2^{-1} \end{bmatrix}.$$

The following lemmas are useful. Here $\sigma(A) = \{\lambda \mid |A - \lambda I| = 0\}$ and $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$.

Lemma 4.3.2. *[38] The eigenvalues of $\mathcal{A}\mathcal{P}_T^{-1}$ are positive real numbers, and the spectrum satisfies*

$$\sigma(\mathcal{A}\mathcal{P}_T^{-1}) \subset \{1\} \cup \sigma(S, D_1)$$

where $S = D_1 - W^\intercal D_2^{-1} W$ is the Schur complement of D_2 in $\mathcal{A}$, and $\sigma(S, D_1)$ contains the eigenvalues μ corresponding to the generalized eigenvalue problem

$$Sz = \mu D_1 z. \tag{4.11}$$

Proof. Let $\mu \in \sigma(\mathcal{A}\mathcal{P}_T^{-1})$ and $[v_1; v_2]$ is the corresponding eigenvector. Then we have

$$\mathcal{A}\mathcal{P}_T^{-1} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} I - W^\intercal D_2^{-1} W D_1^{-1} & W^\intercal D_2^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mu \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

It can be written as

$$\begin{cases} (I - W^\intercal D_2^{-1} W D_1^{-1}) v_1 + (W^\intercal D_2^{-1}) v_2 = \mu v_1 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0 = (1-\mu) v_2 \end{cases}$$

If $\mu = 1$, then $[v_1; v_2]$ satisfies $W^\intercal D_2^{-1} v_2 = W^\intercal D_2^{-1} W D_1^{-1} v_1$. If $\mu \neq 1$, from the first equation we get $(I - W^\intercal D_2^{-1} W D_1^{-1}) v_1 = \mu v_1$ which implies that $\mu \in \sigma(S, D_1)$. □

Lemma 4.3.3. *[29] For a rectangular matrix B,*

$$\rho(B^\intercal B) = \left\| \begin{bmatrix} 0 & B^\intercal \\ B & 0 \end{bmatrix} \right\|^2.$$

Next, we give the main result of this subsection.

Theorem 4.3.4. *The spectrum of the preconditioned Galerkin matrix $\mathcal{A}\mathcal{P}_T^{-1}$ satisfies*

$$\sigma(A\mathcal{P}_T^{-1}) \subset (1 - \hat{\tau}^2, 1],$$

where $\hat{\tau}$ is defined in (4.1).

Proof. By Lemma 4.1.3, it suffices to show that $\sigma(S, D_1) \subset (1 - \hat{\tau}^2, 1]$.

For the upper bound, let μ be an eigenvalue of the generalized eigenvalue problem (4.11), and z be the corresponding eigenvector,

$$\begin{aligned} & (D_1 - W^\intercal D_2^{-1} W) z = \mu D_1 z \\ \Rightarrow\ & (1-\mu) D_1 z = W^\intercal D_2^{-1} W z \\ \Rightarrow\ & (1-\mu) z^\intercal D_1 z = (Wz)^\intercal D_2^{-1} (Wz) \\ \Rightarrow\ & 1 - \mu = \frac{(Wz)^\intercal D_2^{-1} (Wz)}{z^\intercal D_1 z}, \end{aligned}$$

by the symmetric positive definiteness of D_1 and D_2^{-1}, we conclude that

$$1 - \mu \geq 0 \quad \Rightarrow \quad \mu \leq 1.$$

For the lower bound, notice $\sigma(\mathcal{A}\mathcal{P}_T^{-1}) = \sigma(\mathcal{P}_T^{-1}\mathcal{A})$, and

$$\mathcal{P}_T^{-1}\mathcal{A} = \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix}^{-1} \begin{bmatrix} D_1 & W^\intercal \\ W & D_2 \end{bmatrix} = I + \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & W^\intercal \\ 0 & 0 \end{bmatrix}$$

Let $\mu \in \sigma(\mathcal{P}_T^{-1}\mathcal{A})$, $v = [v_1; v_2]$ be the corresponding eigenvector. We have

$$\begin{bmatrix} 0 & W^\intercal \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = (\mu - 1) \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

If $\mu = 1$, then v_2 satisfies $W^\intercal v_2 = 0$. If $\mu \neq 1$, it implies

$$(D_1 - W^\intercal D_2^{-1} W) v_1 = \mu D_1 v_1,$$

and it's equivalent to

$$(I - D_1^{-1} W^\intercal D_2^{-1} W) v_1 = \mu v_1.$$

Therefore, we have $\mu \in \sigma(I - D_1^{-1}W^\intercal D_2^{-1}W)$, and it implies

$$
\begin{aligned}
\mu &\geqslant 1 - \rho(D_1^{-1}W^\intercal D_2^{-1}W) \\
&= 1 - \rho\left((D_1^{-1/2}W^\intercal D_2^{-1/2})(D_2^{-1/2}WD_1^{-1/2})\right) \\
&= 1 - \rho(B^\intercal B) \\
&= 1 - \left\| \begin{bmatrix} 0 & B^\intercal \\ B & 0 \end{bmatrix} \right\|^2 \\
&= 1 - \hat{\tau}^2,
\end{aligned}
$$

where $B = \sum_{s=1}^{L} G_s(\mathcal{I}_2, \mathcal{I}_1) \otimes (K_0^{-\frac{1}{2}} K_s K_0^{-\frac{1}{2}}) = D_2^{-1/2} W D_1^{-1/2}$ introduced in section 3.3.4 is the off-diagonal of the preconditioned system $\mathcal{P}_0^{-\frac{1}{2}} \mathcal{A} \mathcal{P}_0^{-\frac{1}{2}}$. The third equation holds by Lemma 4.3.3 and the last equation holds by (4.5).

Therefore, we conclude that $\mu \in (1 - \hat{\tau}^2, 1]$. □

4.4 A Kronecker Product Preconditioner

Another way to develop a good preconditioner of the $N_x N_\xi \times N_x N_\xi$ matrix $\mathcal{A}$ is to approximate it by a single tensor product $L \otimes R$ with certain matrix L of size $N_\xi \times N_\xi$ and R of size $N_x \times N_x$. Van Loan and Pitsianis [43] presented a general framework (for any $N_x N_\xi \times N_x N_\xi$ matrix $\mathcal{A}$) to find such L and R by minimizing the Frobenius norm $\|\mathcal{A} - L \otimes R\|_F$. However, the additional computational cost introduced in solving L and R might be dominant compare to solving the original stochastic Galerkin system with the previous preconditioners, and therefore is not worth it in practice. Hence a simpler approximation with fixed L or R is brought up in Van Loan and Pitsianis [43] and then specified to the stochastic Galerkin system by Ullmann [42]. We first recall a lemma in [43] for fixed L or R.

Lemma 4.4.1 ([43], Theorem 3)**.** *If $R \in \mathbb{R}^{N_x \times N_x}$ is fixed, then the matrix $L \in \mathbb{R}^{N_\xi \times N_\xi}$ defined by*

$$L_{kl} = \frac{\mathrm{tr}(\mathcal{A}_{kl}^\intercal R)}{\mathrm{tr}(R^\intercal R)}, \quad k, l = 1, 2, \cdots, N_\xi, \tag{4.12}$$

minimizes $\|\mathcal{A} - L \otimes R\|_F$, where $\mathcal{A}_{kl}$ is the k, l-th $N_x \times N_x$ block of $\mathcal{A}$. Similarly, if $L \in \mathbb{R}^{N_\xi \times N_\xi}$ is fixed, then the matrix $R \in \mathbb{R}^{N_x \times N_x}$ defined by

$$R_{ij} = \frac{\mathrm{tr}(\bar{\mathcal{A}}_{ij}^\intercal L)}{\mathrm{tr}(L^\intercal L)}, \quad i, j = 1, 2, \cdots, N_x, \tag{4.13}$$

minimizes $\|\mathcal{A} - L \otimes R\|_F$, where $\bar{\mathcal{A}}_{ij}$ is the i, j-th $N_\xi \times N_\xi$ block of $\mathcal{A}$.

In the stochastic Galerkin system, if $L = I$ is fixed, R is simply equal to K_0 according to (4.13), which results in the mean-based preconditioner $\mathcal{P}_0 = I \otimes K_0$. If $R = K_0$ is fixed, then based on (4.12),

$$\begin{aligned} L_{kl} &= \frac{\mathrm{tr}(\mathcal{A}_{kl}^\intercal K_0)}{\mathrm{tr}(K_0^\intercal K_0)}, \\ &= \frac{1}{\mathrm{tr}(K_0^\intercal K_0)} \left[\mathrm{tr}(K_0^\intercal K_0)\delta_{kl} - \sum_{s=1}^{L} (G_s)_{kl} \mathrm{tr}(K_s^\intercal K_0) \right] \end{aligned} \tag{4.14}$$

which implies that

$$L = I - \sum_{s=1}^{L} \frac{\mathrm{tr}(K_s^\intercal K_0)}{\mathrm{tr}(K_0^\intercal K_0)} G_s.$$

The corresponding preconditioner is the Kronecker product preconditioner $\mathcal{P}_K = L \otimes K_0$ presented in [42]. Instead of using only the mean information for the mean-based preconditioner, the Kronecker product preconditioner utilizes all the information in the stochastic Galerkin matrices.

Chapter 5

Numerical Experiments

In this chapter, we present the computational results (iteration steps, CPU time, eigenvalues, etc.) for four preconditioners we discussed in Chapter 4, including the mean-based block diagonal preconditioner $\mathcal{P}_0$, the two-term variance-involved preconditioner $\mathcal{P}_1$, the Kronecker product preconditioner $\mathcal{P}_K$ and the block triangular preconditioner $\mathcal{P}_T$ on two 2-D benchmark problems in [14, 33]. The block triangular preconditioner $\mathcal{P}_T$ outperforms the other three preconditioners in both problems, and it converges at least twice faster compared to the mean-based preconditioner $\mathcal{P}_0$, especially when the variance σ and total degree p are large. The variance-involved preconditioner $\mathcal{P}_1$ outperforms $\mathcal{P}_0$ when the total degree p and the correlation length c are large. We also compare the performances of $\mathcal{P}_0$ and $\mathcal{P}_T$ on a 3-D problem and $\mathcal{P}_T$ again outperforms $\mathcal{P}_0$ by at least twice faster.

The covariance function has a 2-D exponential kernel

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

with $c_1 = c_2 = 1$, and mean $\mu = \mathbf{E}[a(\boldsymbol{x}, \cdot)] = 1$. We use linear finite element space as X_h and assemble the stiffness matrices $K_s, s = 0, 1, 2, \ldots, L$ using the software package iFEM [6].

We apply MINRES method with tolerance $= 10^{-8}$ for $\mathcal{P}_0$ and $\mathcal{P}_1$, and apply GMRES method with same tolerance and $\mathrm{restart} = 20$ for $\mathcal{P}_K$ and $\mathcal{P}_T$. For $\mathcal{P}_1$ diagonal matrices are inverted by the direct solver backslash "\" in MATLAB. While for the other three preconditioners, diagonal matrices are inverted by the geometric multigrid method with $V(2,2)$. All computations are performed using MATLAB 2017b on a Lenove-Y40 with Intel Core i7-4510U CPU and 8GB of RAM.

5.1 Homogeneous Dirichlet Boundary Condition

Let domain $D = [-0.5, 0.5]^2$ with homogeneous boundary condition and $f(x_1, x_2) = 2(0.5 - x_1^2 - x_2^2)$. Figure 5.1 shows the mean and variance of the pressure on a 32×32 grid with $L = 4$, $p = 4$ and Gaussian random variables.

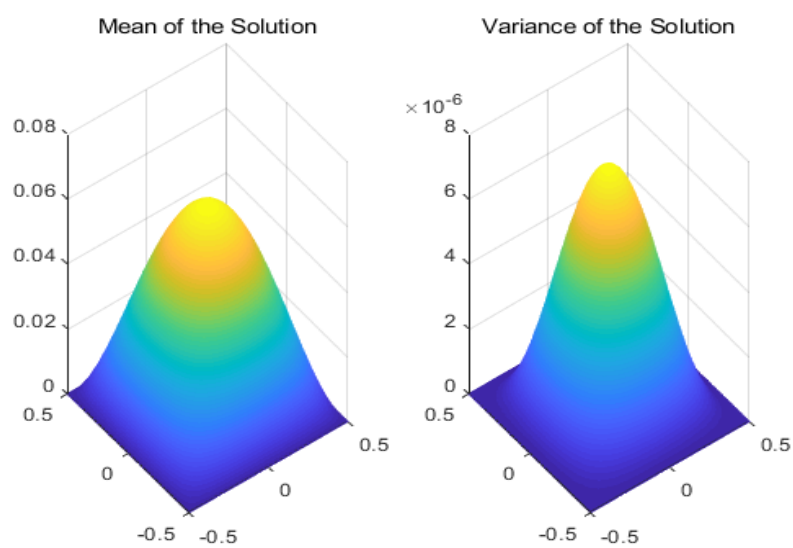


Figure 5.1: Mean (left) and variance (right) of pressure on a 32×32 grid with $L = 4$ and $p = 4$

5.1.1 Gaussian Random Variables and Hermite Polynomial Chaos

We start with Gaussian random variables in the truncated Karhunen-Loève expansion and the corresponding Hermite polynomials as the gPC basis. We compare the performances of four different preconditioners.

Table 5.1: Iteration steps with different mesh size h for all preconditioners (Hermite polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.3$)

h	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
1/16	23	18	17	10
1/32	24	18	17	10
1/64	25	18	17	10
1/128	26	18	17	10

Table 5.2: Time costs (in seconds) with different mesh size h for all preconditioners (Hermite polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.3$)

h	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
1/16	2.8	1.4	2.6	1.7
1/32	16.6	5.3	10.6	6.8
1/64	59.5	26.2	48.3	28.7
1/128	248.4	255.7	195.6	118.2

Table 5.1 and 5.2 show the results of iteration steps and time costs with various mesh sizes h, Hermite polynomials and $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.3$. We can see that the iteration steps for different preconditioners are uniform to various mesh sizes h while the time costs scale linearly to h^2. Therefore, we can fix the mesh size h and variance σ, and discuss how the performances of different preconditioners vary with different L and p.

From Table 5.3 and 5.4 we can see that under the same $p,\ L$ values, the performances of all four preconditioners are ranked as $\mathcal{P}_T > \mathcal{P}_K > \mathcal{P}_1 \approx \mathcal{P}_0$. Here $\mathcal{P}_1$ and $\mathcal{P}_0$ have similar performances, however, if restricted to a for loop to go through all diagonal blocks, $\mathcal{P}_1$ still outperforms $\mathcal{P}_0$. The

Table 5.3: Iteration steps with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	13	11	10	6
4	3	18	14	12	8
	4	25	18	15	11
	2	13	12	10	6
6	3	18	15	13	8
	4	26	20	17	10

Table 5.4: Time costs with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	8.8	9.7	7.9	4.8
4	3	29.4	31.0	23.5	14.5
	4	80.4	79.5	57.5	43.1
	2	16.7	21.7	15.9	9.4
6	3	69.8	74.4	63.6	36.8
	4	248.4	255.7	195.6	118.2

advantage for $\mathcal{P}_0$ is that it only needs one V-cycle to go through all the diagonal blocks while $\mathcal{P}_1$ has to do one V-cycle per diagonal block. All preconditioners are robust to h and iteration steps are positive correlated to p and L. When p and L are large, the best-performance GMRES preconditioner $\mathcal{P}_T$ converges almost twice faster than MINRES with the mean-based preconditioner $\mathcal{P}_0$.

Table 5.5 and 5.6 show the iteration steps and time costs for different preconditioners with correlation lengths $c_1 = c_2 = 10$. Compared to the previous case where $c_1 = c_2 = 1$, the variance-involved preconditioner $\mathcal{P}_1$ and Kronecker product preconditioner $\mathcal{P}_K$ perform much better, while the mean-based preconditioner $\mathcal{P}_0$ and block triangular preconditioner $\mathcal{P}_T$ remain almost the same. When the correlation lengths are large, the eigenvalues in the K-L expansion decay faster and the two-term variance-involved preconditioner $\mathcal{P}_1$ contains more information about the whole system

Table 5.5: Iteration steps with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3,\ c_1 = c_2 = 10$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	10	8	7	5
4	3	15	10	9	8
	4	23	12	11	10
	2	10	8	7	4
6	3	15	10	9	7
	4	22	13	11	10

Table 5.6: Time costs with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3,\ c_1 = c_2 = 10$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	7.3	7.9	6.5	3.6
4	3	26.3	21.4	18.9	13.5
	4	77.7	49.2	46.6	39.7
	2	14.2	14.7	12.1	7.0
6	3	61.9	50.9	46.3	33.8
	4	249.8	157.1	145.8	122.2

matrix $\mathcal{A}$, which leads to a better performance. We will discuss more on this in the next subsection when the condition number of the preconditioned system is computed.

Next we test the convergence for various variance by fixing $\mu = 1,\ h = 1/16,\ p = 4$ and $L = 6$.

Table 5.7: Iteration steps with various σ (Hermite polynomials with $\mu = 1,\ h = 1/16,\ p = 4$ and $L = 6$)

σ	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_D$	$\mathcal{P}_T$
0.1	9	8	7	4
0.2	14	12	11	7
0.3	23	20	17	10
0.37	74	60	55	32

Table 5.7 shows the iteration steps for different preconditioners with variance σ. We can find that the iteration steps increase as σ increases for all preconditioners. Again the performances of all

four preconditioners are ranked as $\mathcal{P}_T > \mathcal{P}_K > \mathcal{P}_1 > \mathcal{P}_0$. It is worth noticing that when $\sigma = 0.4$ and $p = 4$, the original system may not be positive definite therefore none of the preconditioners converges efficiently. That is why we choose $\sigma = 0.37$ instead of 0.4 in the last row for comparison.

5.1.2 Uniform Random Variables and Legendre Polynomial Chaos

We consider a different choice of random variable and gPC basis. Here we use uniform random variable in the truncated Karhunen-Loève expansion and corresponding Legendre polynomials as the gPC basis.

Table 5.8: Iteration steps with different mesh size h for all preconditioners (Legendre polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.7$)

h	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
1/16	37	29	26	16
1/32	39	30	26	16
1/64	40	30	26	16
1/128	40	30	26	16

Table 5.9: Time costs with different mesh size h for all preconditioners (Legendre polynomials with $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.7$)

h	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
1/16	4.7	2.0	3.8	2.5
1/32	26.3	8.7	19.2	12.4
1/64	112.0	44.3	87.2	54.6
1/128	480.7	395.9	351.2	191.9

Similar to the Hermite example, Table 5.8 and 5.9 show the results of iteration steps and time costs with various mesh sizes h with Legendre polynomials and $p = 4,\ L = 6,\ \mu = 1,\ \sigma = 0.3$. We can see that the iteration steps for different preconditioners are still uniform to various mesh sizes h while the time costs scale linearly to h^2. Therefore, we can fix the mesh size h and discuss how the performances of different preconditioners vary with different L and p.

Table 5.10: Iteration steps with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	16	14	12	7
4	3	23	18	15	10
	4	30	23	18	12
	2	17	15	13	8
6	3	26	21	17	11
	4	40	33	26	16

Table 5.11: Time costs with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	10.8	13.0	9.5	5.5
4	3	37.4	37.5	29.5	18.1
	4	98.0	95.2	68.7	46.1
	2	28.5	27.3	24.5	15.0
6	3	126.2	103.6	98.5	66.2
	4	480.7	395.9	351.2	191.9

From Table 5.10 and 5.11 we can see that under the same p and L values, the performances of all four preconditioners are ranked as $\mathcal{P}_T > \mathcal{P}_K > \mathcal{P}_1 > \mathcal{P}_0$. All preconditioners are robust to h and iteration steps are positively correlated to p and L. When p and L are large, the best-performance GMRES preconditioner $\mathcal{P}_T$ converges more than twice faster compared to MINRES with the mean-based preconditioner $\mathcal{P}_0$.

5.1.3 Maximal and Minimal Eigenvalues of the Preconditioned System

We now check the spectral approximation of different preconditioners to the original matrix $\mathcal{A}$. We use Matlab function `eigs` to calculate the extremal eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ with $\mathcal{P} = \mathcal{P}_0$ and $\mathcal{P} = \mathcal{P}_1$. The closer the extremal eigenvalues are compared to 1, the better the approximation of the preconditioner. Here we use Hermite polynomials in Gaussian random variables.

Example 5.1.1. We consider first the case where the covariance function is exponential kernel with $\sigma = 0.1,\ \mu = 1,\ c_1 = c_2 = 1$ and $h = 1/8$. Computed extremal eigenvalues ν and the corresponding condition number κ are listed in Table 5.12.

Table 5.12: Example 5.1.1: Extremal Eigenvalues with $\sigma = 0.1,\ \mu = 1,\ c = 1,\ h = 1/8$

L	p	$\nu_{\min}(\mathcal{P}_0^{-1}A)$	$\nu_{\max}(\mathcal{P}_0^{-1}A)$	$\kappa(\mathcal{P}_0^{-1}A)$	$\nu_{\min}(\mathcal{P}_1^{-1}A)$	$\nu_{\max}(\mathcal{P}_1^{-1}A)$	$\kappa(\mathcal{P}_1^{-1}A)$
1	1	0.9155	1.0845	1.1847	1.0000	1.0000	1.0000
	2	0.8536	1.1464	1.3431	1.0000	1.0000	1.0000
	3	0.8027	1.1973	1.4917	1.0000	1.0000	1.0000
	4	0.7585	1.2415	1.6369	1.0000	1.0000	1.0000
2	1	0.9124	1.0876	1.1919	0.9563	1.0437	1.0914
	2	0.8484	1.1516	1.3575	0.9239	1.0761	1.1649
	3	0.7956	1.2044	1.5137	0.8967	1.1033	1.2305
	4	0.7499	1.2501	1.6671	0.8725	1.1275	1.2922
3	1	0.9106	1.0894	1.1964	0.9455	1.0545	1.1152
	2	0.8451	1.1549	1.3666	0.9052	1.0948	1.2094
	3	0.7912	1.2088	1.5277	0.8716	1.1284	1.2947
	4	0.7445	1.2555	1.6863	0.8419	1.1581	1.3756

Example 5.1.2. We consider the case when $\sigma = 0.3$ and everything else is the same. The purpose of this example is to compare the change of extreme eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ with a bigger variance σ. Computed extremal eigenvalues ν and the corresponding condition number κ are listed in Table 5.13.

Example 5.1.3. We consider the case when $\sigma = 0.1$, $c_1 = c_2 = 10$, and everything else is the same. The purpose of this example is to compare the change of extreme eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ with various correlation length c_1 and c_2. Computed extremal eigenvalues ν and the corresponding condition number κ are listed in Table 5.14.

Example 5.1.4. Finally we consider the case when $\sigma = 0.3$, $c_1 = c_2 = 10$, and everything else is the same. The purpose of this example is to compare the extreme eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$ with

Table 5.13: Example 5.1.2: Extremal Eigenvalues with $\sigma = 0.3,\ \mu = 1,\ c = 1,\ h = 1/8$

L	p	$\nu_{\min}(\mathcal{P}_0^{-1}A)$	$\nu_{\max}(\mathcal{P}_0^{-1}A)$	$\kappa(\mathcal{P}_0^{-1}A)$	$\nu_{\min}(\mathcal{P}_1^{-1}A)$	$\nu_{\max}(\mathcal{P}_1^{-1}A)$	$\kappa(\mathcal{P}_1^{-1}A)$
1	1	0.7464	1.2536	1.6796	1.0000	1.0000	1.0000
	2	0.5607	1.4393	2.5668	1.0000	1.0000	1.0000
	3	0.4080	1.5920	3.9025	1.0000	1.0000	1.0000
	4	0.2754	1.7246	6.2615	1.0000	1.0000	1.0000
2	1	0.7373	1.2627	1.7124	0.8658	1.1342	1.3100
	2	0.5451	1.4549	2.6692	0.7543	1.2457	1.6514
	3	0.3869	1.6131	4.1694	0.6421	1.3579	2.1146
	4	0.2497	1.7503	7.0105	0.5145	1.4855	2.8875
3	1	0.7317	1.2683	1.7334	0.8337	1.1663	1.3991
	2	0.5353	1.4647	2.7363	0.6994	1.3006	1.8597
	3	0.3737	1.6263	4.3519	0.5707	1.4293	2.5044
	4	0.2335	1.7665	7.5644	0.4339	1.5661	3.6091

Table 5.14: Example 5.1.3: Extremal Eigenvalues with $\sigma = 0.1,\ \mu = 1,\ c = 10,\ h = 1/8$

L	p	$\nu_{\min}(\mathcal{P}_0^{-1}A)$	$\nu_{\max}(\mathcal{P}_0^{-1}A)$	$\kappa(\mathcal{P}_0^{-1}A)$	$\nu_{\min}(\mathcal{P}_1^{-1}A)$	$\nu_{\max}(\mathcal{P}_1^{-1}A)$	$\kappa(\mathcal{P}_1^{-1}A)$
1	1	0.9017	1.0983	1.2180	1.0000	1.0000	1.0000
	2	0.8298	1.1702	1.4104	1.0000	1.0000	1.0000
	3	0.7705	1.2295	1.5956	1.0000	1.0000	1.0000
	4	0.7192	1.2808	1.7809	1.0000	1.0000	1.0000
2	1	0.9013	1.0987	1.2189	0.9821	1.0179	1.0365
	2	0.8291	1.1709	1.4122	0.9686	1.0314	1.0647
	3	0.7697	1.2303	1.5984	0.9572	1.0428	1.0894
	4	0.7181	1.2819	1.7850	0.9469	1.0531	1.1121
3	1	0.9010	1.0990	1.2198	0.9750	1.0250	1.0512
	2	0.8285	1.1715	1.4139	0.9563	1.0437	1.0913
	3	0.7689	1.2311	1.6011	0.9405	1.0595	1.1266
	4	0.7172	1.2828	1.7888	0.9262	1.0738	1.1595

large correlation length c_1 and c_2 as well as large σ. Computed extremal eigenvalues ν and the corresponding condition number κ are listed in Table 5.14.

Table 5.15: Example 5.1.4: Extremal Eigenvalues with $\sigma = 0.3,\ \mu = 1,\ c = 10,\ h = 1/8$

L	p	$\nu_{\min}(\mathcal{P}_0^{-1}A)$	$\nu_{\max}(\mathcal{P}_0^{-1}A)$	$\kappa(\mathcal{P}_0^{-1}A)$	$\nu_{\min}(\mathcal{P}_1^{-1}A)$	$\nu_{\max}(\mathcal{P}_1^{-1}A)$	$\kappa(\mathcal{P}_1^{-1}A)$
1	1	0.7051	1.2949	1.8364	1.0000	1.0000	1.0000
	2	0.4893	1.5107	3.0878	1.0000	1.0000	1.0000
	3	0.3116	1.6884	5.4176	1.0000	1.0000	1.0000
	4	0.1576	1.8424	11.6940	1.0000	1.0000	1.0000
2	1	0.7040	1.2960	1.8408	0.9440	1.0560	1.1186
	2	0.4874	1.5126	3.1037	0.8927	1.1073	1.2404
	3	0.3091	1.6909	5.4707	0.8299	1.1701	1.4100
	4	0.1544	1.8456	11.9511	0.7254	1.2746	1.7571
3	1	0.7030	1.2970	1.8449	0.9221	1.0779	1.1689
	2	0.4856	1.5144	3.1187	0.8512	1.1488	1.3497
	3	0.3067	1.6933	5.5213	0.7657	1.2343	1.6121
	4	0.1515	1.8485	12.2020	0.6286	1.3714	2.1818

All examples show that the extremal eigenvalues with both preconditioners are symmetric on 1, which supports our theoretical results in Theorem 4.1.6 and Theorem 4.2.2. The extremal eigenvalues with preconditioner $\mathcal{P}_1$ are much better (closer to 1) than the extremal eigenvalues with $\mathcal{P}_0$, especially when either the variance σ or the correlation length c is large or both are large. The condition number associated with $\mathcal{P}_1$ could be six times less than condition number associated with $\mathcal{P}_0$ in Example 5.1.4.

5.2 Mixed Boundary Condition

Now we consider a second benchmark problem [33] with steady flow from left to right on the domain $D = [0,1]^2$ with $f \equiv 0$, $\partial D_D = \{0,1\} \times [0,1]$ and $\partial D_N = \partial D \setminus \partial D_D$. Let $\vec{q} \cdot \vec{n} = 0$ at the two horizontal walls so that the flow is tangent to those boundaries. $u_D = 1$ on $\{0\} \times [0,1]$ and

$u_D = 0$ on $\{1\} \times [0, 1]$. Still we consider the same covariance function with $L_{c,1} = L_{c,2} = 1$, and mean $\mu = 1$.

5.2.1 Gaussian Random Variables and Hermite Polynomial Chaos

We still start with Gaussian random variables in the truncated Karhunen-Loève expansion and the corresponding Hermite polynomials as the stochastic basis. Figure 5.2 shows the mean and variance of the pressure on a 32×32 grid with $L = 4$ and $p = 4$.

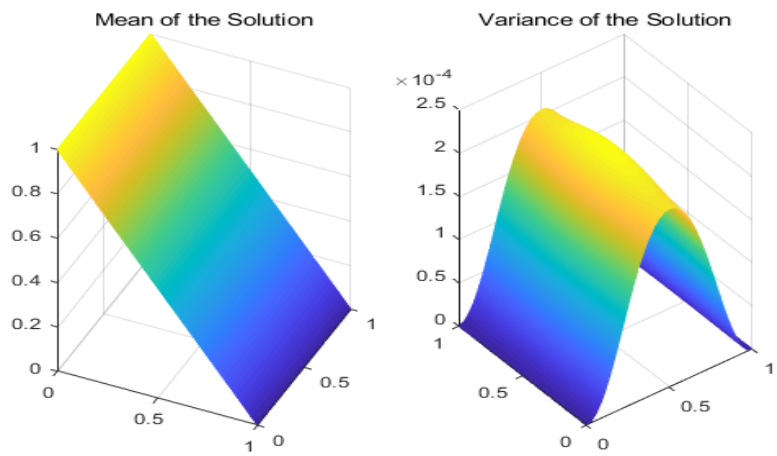


Figure 5.2: Mean (left) and variance (right) of pressure on a 32×32 grid with $L = 4$ and $p = 4$

Next we compare the performances of four different preconditioners.

Table 5.16: Iteration steps with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	11	10	10	5
4	3	15	12	13	9
	4	20	15	16	9
	2	11	11	11	5
6	3	15	14	14	9
	4	20	17	18	10

Table 5.17: Time costs with different L and p for all preconditioners (Hermite polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.3$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	8.6	10.3	9.0	4.5
4	3	27.0	25.0	23.8	16.9
	4	65.2	56.0	62.6	35.8
	2	13.1	17.7	13.8	6.7
6	3	48.2	62.7	50.4	33.1
	4	169.1	191.3	174.8	95.6

From Table 5.16 and 5.17 we can see that under the same p and L values, $\mathcal{P}_T$ and $\mathcal{P}_1$ outperform the other two preconditioners. All preconditioners are robust to h and iteration steps are positively correlated to p and L. When p and L are large, the best-performance GMRES preconditioner $\mathcal{P}_T$ still converges almost twice faster than MINRES with the mean-based preconditioner $\mathcal{P}_0$.

5.2.2 Uniform Random Variables and Legendre Polynomial Chaos

We consider a different choice of random variables and stochastic basis. Here we use uniform random variable in the truncated Karhunen-Loève expansion and corresponding Legendre polynomials as the stochastic basis.

Table 5.18: Iteration Steps with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	14	12	12	6
4	3	19	15	16	11
	4	24	19	19	11
	2	15	13	14	7
6	3	21	18	18	12
	4	31	26	26	14

From Table 5.18 and 5.19 we can see that under the same p and L values, $\mathcal{P}_T$ and $\mathcal{P}_1$ outperform the other two preconditioners. All preconditioners are robust to h and iteration steps are positive

Table 5.19: Time costs with different L and p for all preconditioners (Legendre polynomials with $h = 1/128,\ \mu = 1,\ \sigma = 0.7$)

L	p	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_K$	$\mathcal{P}_T$
	2	8.8	11.6	8.7	4.0
4	3	29.9	32.9	27.0	18.3
	4	72.6	81.3	67.8	41.1
	2	21.6	24.0	20.7	11.7
6	3	86.5	93.7	86.2	57.9
	4	322.8	327.3	331.0	176.4

correlated to p and L. When p and L are large, the best-performance GMRES preconditioner $\mathcal{P}_T$ still converges almost twice faster than MINRES with the mean-based preconditioner $\mathcal{P}_0$.

Next we test the convergence for various variance by fixing $\mu = 1,\ h = 1/16,\ p = 4$ and $L = 6$.

Table 5.20: Iteration steps with various σ (Hermite polynomials with $\mu = 1,\ h = 1/16,\ p = 4$ and $L = 6$)

σ	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_D$	$\mathcal{P}_T$
0.1	9	8	8	4
0.2	13	12	11	6
0.3	22	18	18	10
0.37	66	52	53	29

Table 5.20 shows the iteration steps for different preconditioners with various variance σ. We can find that the iteration steps increase as σ increases for all preconditioners. Similar to the first benchmark problem, the performances of all four preconditioners are ranked as $\mathcal{P}_T > \mathcal{P}_K \approx \mathcal{P}_1 > \mathcal{P}_0$. It is worth noticing that when $\sigma = 0.4$ and $p = 4$, the original system may not be positive definite therefore none of the preconditioners converges efficiently. That is why we choose $\sigma = 0.37$ instead of 0.4 in the last row for comparison.

5.3 A 3-D Homogeneous Dirichlet Boundary Condition Problem

We extend the first benchmark problem to 3-D. Let domain $D = [-0.5, 0.5]^3$ with homogeneous boundary condition and $f(x_1, x_2) = 2(0.5 - x_1^2 - x_2^2 - x_3^2)$. Figure 5.3 shows the mean and variance of the pressure on a $16 \times 16 \times 16$ grid with $L = 10$, $p = 4$ and uniform random variables.

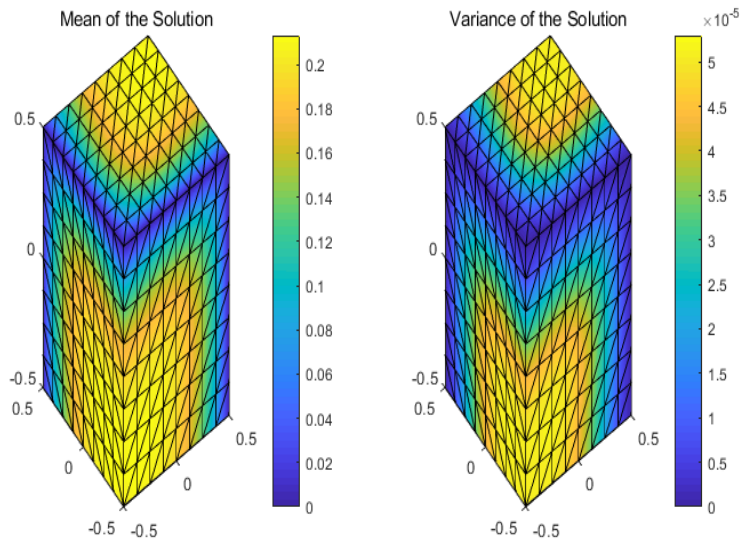


Figure 5.3: Mean (left) and variance (right) of pressure on a $16 \times 16 \times 16$ grid with $L = 10$ and $p = 4$

We use uniform random variable in the truncated Karhunen-Loève expansion and corresponding Legendre polynomials as the gPC basis. Now we compare the performances of two preconditioners: the mean-based block preconditioner $\mathcal{P}_0$ and the block triangular preconditioner $\mathcal{P}_T$.

From Table 5.21 we can see that similar to the 2-D example, the block triangular preconditioner $\mathcal{P}_T$ outperforms the mean-based preconditioner $\mathcal{P}_0$. When p and σ are large, $\mathcal{P}_T$ converges more than twice faster compared to $\mathcal{P}_0$.

Table 5.21: Iteration steps and time costs with different L and p for both preconditioners (Legendre polynomials with $h = 1/16,\ \mu = 1,\ \sigma = 0.7$)

	$\mathcal{P}_0$		$\mathcal{P}_T$	
p	iter	time	iter	time
2	17	25.7	7	12.4
3	25	160.0	10	71.2
4	39	859.3	14	370.9

Chapter 6

Stochastic Galerkin Methods for Mixed Formulation

In this chapter, we extend the stochastic Galerkin methods from the primal formulation in (2.1) to the mixed formulation of stochastic diffusion problems [18, 16, 7]. The motivation of the mixed formulation is that a second variable of interest is introduced in the formulation due to the physical application [15]. The second variable is usually some kind of derivatives to the original variable, for example the stress and the displacement in the elasticity equations. The second variable can also be independent of the original one, for example the pressure and velocity in the Stokes equations.

6.1 Stochastic Galerkin Mixed Formulation

In this section, we introduce the mixed formulation of stochastic diffusion problem, and convert the stochastic problem into a deterministic one by approximating the random coefficient with truncated Karhunen-Loève expansion.

6.1.1 Model Problem

In a similar setting in (2.1), let $D \subset \mathbb{R}^2$ be a bounded domain. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space where Ω is the set of outcomes, $\mathcal{F}$ is the σ-algebra of the subsets of Ω, and P is a probability measure on $\mathcal{F}$. Consider the first-order system of the diffusion problem with random coefficient is given as follows: find two random fields $\boldsymbol{q}(\boldsymbol{x}, \omega)$ and $u(\boldsymbol{x}, \omega)$ such that for P-almost surely,

$$\begin{cases} T(\boldsymbol{x}, \omega)\boldsymbol{q}(\boldsymbol{x}, \omega) + \nabla u(\boldsymbol{x}, \omega) = 0, & \text{in } D \times \Omega, \\ -\text{div}\boldsymbol{q}(\boldsymbol{x}, \omega) = f(\boldsymbol{x}), & \text{in } D \times \Omega, \\ u(\boldsymbol{x}, \omega) = g(\boldsymbol{x}), & \text{on } \partial D_D \times \Omega, \\ \boldsymbol{n} \cdot \boldsymbol{q}(\boldsymbol{x}, \omega) = 0, & \text{on } \partial D_N \times \Omega, \end{cases} \tag{6.1}$$

where $\partial D_D \cap \partial D_N = \emptyset$ and $\partial D_D \cup \partial D_N = \partial D$. Here we still assume the randomness comes from the diffusive coefficient $a(\boldsymbol{x}, \omega) : \ D \times \Omega \to \mathbb{R}$, while the source term f and boundary conditions are deterministic.

To obtain the variational form of problem (6.1) we introduce the space $L^2_P(\Omega)$ which includes all random variables in the probability space $(\Omega, \mathcal{F}, P)$ with finite variances. Let $H_0(\text{div}, D) \triangleq \{\boldsymbol{r} \in L^2(D)^2 : \text{div}\boldsymbol{r} \in L^2(D), \ \boldsymbol{n} \cdot \boldsymbol{r}|_{\partial D_N} = 0\}$ be the Sobolev space of square-integrable symmetric vector fields with square-integrable divergence and zero normal component on the Neumann boundary. The mixed formulation of problem (6.1) is to find $\boldsymbol{q}(\boldsymbol{x}, \omega) \in H_0(\text{div}, D) \otimes L^2_P(\Omega)$ and $u(\boldsymbol{x}, \omega) \in L^2(D) \otimes L^2_P(\Omega)$ such that

$$\begin{aligned} \int_{D\times\Omega} T\boldsymbol{q} \cdot \boldsymbol{r} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\omega - \int_{D\times\Omega} u \, \text{div}\boldsymbol{r} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\omega &= -\int_{\partial D\times\Omega} g\boldsymbol{n} \cdot \boldsymbol{r} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\omega, \quad \forall \boldsymbol{r} \in H_0(\text{div}, D) \otimes L^2_P(\Omega), \\ \int_{D\times\Omega} v \, \text{div}\boldsymbol{q} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\omega &= -\int_{D\times\Omega} f \, v \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\omega, \qquad \forall v \in L^2(D) \otimes L^2_P(\Omega). \end{aligned} \tag{6.2}$$

Assume that the diffusion coefficient $T(\boldsymbol{x}, \omega)$ is bounded and strictly positive, i.e. there exist constants $T_{\min} < T_{\max}$ such that

$$0 < T_{\min} < T(\boldsymbol{x}, \omega) < T_{\max} \quad a.e. \text{ in } D \times \Omega,$$

then the mixed formulation (6.2) is well-posed. A detailed proof of well-posedness is given in [16, 18].

6.1.2 Karhunen-Loève Expansion

To convert the stochastic problem into a deterministic one, the infinite dimensional diffusion coefficient $T(\boldsymbol{x}, \omega)$ is approximated by finitely many random variables. Similar to problem (2.2), we apply truncated Karhunen-Loève expansion to $T(\boldsymbol{x}, \omega)$

$$T(\boldsymbol{x}, \omega) \approx T_L(\boldsymbol{x}, \omega) \triangleq \mu(\boldsymbol{x}) + \sum_{s=1}^{L} \sqrt{\lambda_s} \alpha_s(\boldsymbol{x}) \xi_s(\omega), \tag{6.3}$$

where $\mu(\boldsymbol{x}) = \mathbf{E}[T(\boldsymbol{x}, \omega)]$ is the mean of random field and random variables ξ_s given by

$$\xi_s(\omega) = \frac{1}{\sqrt{\lambda_s}} \int_D T(\boldsymbol{x}, \omega) \alpha_s(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

are uncorrelated with zero mean $\mathbf{E}[\xi_s] = 0$ and unit variance $\mathrm{Var}(\xi_s) = 1$.

Let $\Gamma_s \subset \mathbb{R}$ be the range of the random variable $\xi_s(\omega)$, and each ξ_s possesses a density function $p_s : \Gamma_s \to \mathbb{R}_0^+$. Define $\Gamma \triangleq \Gamma_1 \times \cdots \times \Gamma_L$, $\boldsymbol{\xi} \triangleq (\xi_1, \cdots, \xi_L)$ and the joint density function can be written as $p(\boldsymbol{\xi}) = p(\xi_1) \cdots p(\xi_L)$. Substituting $L_P^2(\Omega)$ by $L_p^2(\Gamma)$, a weighted L^2 space of all measurable functions on Γ with finite variances, the mixed formulation becomes to the following:

find $\boldsymbol{q}(\boldsymbol{x},\omega) \in H_0(\mathrm{div}, D) \otimes L_p^2(\Gamma)$ and $u(\boldsymbol{x},\omega) \in L^2(D) \otimes L_p^2(\Gamma)$ such that

$$\begin{aligned} \int_{D\times\Gamma} \hat{T}_L \boldsymbol{q} \cdot \boldsymbol{r} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi} - \int_{D\times\Gamma} u\,\mathrm{div}\boldsymbol{r}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi} &= -\int_{\partial D\times\Gamma} g\boldsymbol{n}\cdot\boldsymbol{r}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}, \quad \forall \boldsymbol{r} \in H_0(\mathrm{div}, D) \otimes L_p^2(\Gamma), \\ \int_{D\times\Gamma} v\,\mathrm{div}\boldsymbol{q}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi} &= -\int_{D\times\Gamma} f\,v\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}, \qquad \forall v \in L^2(D) \otimes L_p^2(\Gamma), \end{aligned} \tag{6.4}$$

where $\hat{T}_L(\boldsymbol{x},\boldsymbol{\xi}) = \mu(\boldsymbol{x}) + \sum_{s=1}^{L} \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\xi_s$.

6.2 Matrix Form of Stochastic Galerkin Methods

In this section, we present the matrix form of stochastic Galerkin methods for the mixed formulation (6.4).

6.2.1 Mixed Finite Element Method

We start with the discretization of the mixed formulation (6.4). Let $\{\mathcal{T}_h\}_{h>0}$ be a regular family [5, 11] of partitions of D where $h = \max_{K\in\mathcal{T}_h} h_K$ and $h_K = \mathrm{diag(K)}$. For any element $K \subset D$ and non-negative integer k, $P_k(K)$ stands for the set of all polynomials in K with the total degree less than or equal to k, and $\boldsymbol{P}_k(K, \mathbb{R}^2)$ represents the vector version of $P_k(K)$. Define

$$W_h = \left\{ v_h \in L^2(D) : v_h|_K \in P_k(K) \text{ for each } K \in \mathcal{T}_h \right\},$$

and let $\boldsymbol{V}_h$ be the corresponding Raviart-Thomas (RT) element space

$$\boldsymbol{V}_h = \left\{ \boldsymbol{r}_h \in H_0(\mathrm{div}, D) : r_h|_K \in \boldsymbol{P}_k(K, \mathbb{R}^2) + P_k(K)\boldsymbol{x} \text{ for each } K \in \mathcal{T}_h \right\}$$

or Brezzi-Douglas-Marini (BDM) element space

$$\boldsymbol{V}_h = \left\{\boldsymbol{r}_h \in H_0(\mathrm{div}, D) : r_h|_K \in \boldsymbol{P}_{k+1}(K, \mathbb{R}^2) \text{ for each } K \in \mathcal{T}_h\right\}$$

with $k \geqslant 0$. Let $\{\boldsymbol{\varphi}_j(\boldsymbol{x})\}_{j=1}^{n_q}$ and $\{\phi_j(\boldsymbol{x})\}_{j=1}^{n_u}$ be the bases of the finite element spaces $\boldsymbol{V}_h$ and W_h respectively. Recall that $\mathfrak{C}_{p,L}$ is the complete polynomial space for the stochastic space introduced in (3.2). Define the tensor product spaces

$$\begin{aligned} \boldsymbol{V}_h \otimes \mathfrak{C}_{p,L} &= \mathrm{span}\{\psi_{\boldsymbol{k}}(\boldsymbol{\xi})\boldsymbol{\varphi}_j(\boldsymbol{x}) \text{ for } |\boldsymbol{k}| \leqslant p \text{ and } j = 1, \cdots, n_q\}, \\ W_h \otimes \mathfrak{C}_{p,L} &= \mathrm{span}\{\psi_{\boldsymbol{k}}(\boldsymbol{\xi})\phi_j(\boldsymbol{x}) \text{ for } |\boldsymbol{k}| \leqslant p \text{ and } j = 1, \cdots, n_u\}. \end{aligned} \tag{6.5}$$

Therefore, the stochastic Galerkin method for the mixed formulation (6.4) is as follows: find $\boldsymbol{q}_{hp}(\boldsymbol{x}, \omega) \in V_h \otimes \mathfrak{C}_{p,L}$ and $u_{hp}(\boldsymbol{x}, \omega) \in \boldsymbol{W}_h \otimes \mathfrak{C}_{p,L}$ such that

$$\begin{aligned} \int_{D\times\Gamma} \hat{T}_L \boldsymbol{q}_{hp} \cdot \boldsymbol{r}_{hp} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi} - \int_{D\times\Gamma} u_{hp}\, \mathrm{div}\boldsymbol{r}_{hp} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi} &= -\int_{\partial D\times\Gamma} g\boldsymbol{n} \cdot \boldsymbol{r}_{hp} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}, \\ \int_{D\times\Gamma} v_{hp}\, \mathrm{div}\boldsymbol{q}_{hp} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi} &= -\int_{D\times\Gamma} f\, v_{hp} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}, \end{aligned} \tag{6.6}$$

$\forall \boldsymbol{r}_{hp} \in \boldsymbol{V}_h \otimes \mathfrak{C}_{p,L}$ and $\forall v_{hp} \in W_h \otimes \mathfrak{C}_{p,L}$. The well-posedness of (6.6) and its priori error estimate is given in [16].

6.2.2 Matrix Form

With the tensor product bases defined in (6.5), the solutions $\boldsymbol{q}_{hp}$ and u_{hp} can be written as

$$\boldsymbol{q}(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{j=1}^{N_q} \sum_{l=1}^{N_\xi} q_{jl} \boldsymbol{\varphi}_j(\boldsymbol{x}) \psi_k(\boldsymbol{\xi}), \quad u(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{j=1}^{N_u} \sum_{l=1}^{N_\xi} u_{jl} \phi_i(\boldsymbol{x}) \psi_k(\boldsymbol{\xi}).$$

Plug them into (6.6) together with the truncated KL expansion (6.3) and test functions

$$\boldsymbol{r}(\boldsymbol{x},\boldsymbol{\xi}) = \boldsymbol{\varphi}_i(\boldsymbol{x})\psi_k(\boldsymbol{\xi}),\ i = 1,\cdots,N_q,\ k = 1,\cdots,N_\xi,$$

$$v(\boldsymbol{x},\boldsymbol{\xi}) = \boldsymbol{\phi}_i(\boldsymbol{x})\psi_k(\boldsymbol{\xi}),\ i = 1,\cdots,N_u,\ k = 1,\cdots,N_\xi,$$

we get the matrix form of the saddle-point problem

$$\begin{pmatrix} \mathcal{A} & \mathcal{B}^\intercal \\ \mathcal{B} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{g} \\ \boldsymbol{f} \end{pmatrix} \tag{6.7}$$

The matrices $\mathcal{A}$ and $\mathcal{B}$ have the tensor product structure as

$$\mathcal{A} = I \otimes A_0 + \sum_{s=1}^{L} G_s \otimes A_s, \quad \mathcal{B} = I \otimes B_0,$$

where $B_0 \in \mathbb{R}^{N_u \times N_q}$, $A_s \in \mathbb{R}^{N_q \times N_q}$ are the stiffness matrices using coefficients $\sqrt{\lambda_s}\alpha_s(\boldsymbol{x})$, and $G_s \in \mathbb{R}^{N_\xi \times N_\xi}$ are the stochastic matrices. More precisely, for $s = 1,\cdots,L$,

$$\begin{aligned}
A_0(i,j) &= \int_D \mu\, \boldsymbol{\varphi}_j(\boldsymbol{x}) \cdot \boldsymbol{\varphi}_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, && i,j = 1,\cdots,N_q, \\
A_s(i,j) &= \int_D \sqrt{\lambda_s}\alpha_s(\boldsymbol{x})\boldsymbol{\varphi}_j(\boldsymbol{x}) \cdot \boldsymbol{\varphi}_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, && i,j = 1,\cdots,N_q, \\
B_0(i,j) &= -\int_D \phi_i(\boldsymbol{x})\, \mathrm{div}\boldsymbol{\varphi}_j(\boldsymbol{x})\, \mathbf{d}\boldsymbol{x}, && i = 1,\cdots,N_u,\ j = 1,\cdots,N_q, \\
G_s(k,l) &= \int_\Gamma \xi_s \psi_l(\boldsymbol{\xi})\psi_k(\boldsymbol{\xi})p(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi}, && k,l = 1,\cdots,N_\xi.
\end{aligned} \tag{6.8}$$

The solution vectors can be written into blocks

$$\boldsymbol{q} = \begin{pmatrix} \boldsymbol{q}_1 \\ \vdots \\ \boldsymbol{q}_{N_\xi} \end{pmatrix} \in \mathbb{R}^{N_q N_\xi}, \qquad \boldsymbol{u} = \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_{N_\xi} \end{pmatrix} \in \mathbb{R}^{N_u N_\xi}, \tag{6.9}$$

with each block has $(\boldsymbol{q}_l)_j = q_{jl},\ j = 1, \cdots, N_q,\ (\boldsymbol{u}_l)_j = u_{jl},\ j = 1, \cdots, N_u$, for $l = 1, \cdots, N_\xi$. Similarly, the right-hand side can also be written into blocks

$$\boldsymbol{g} = \begin{pmatrix} \boldsymbol{g}_1 \\ \vdots \\ \boldsymbol{g}_{N_\xi} \end{pmatrix} \in \mathbb{R}^{N_q N_\xi}, \qquad \boldsymbol{f} = \begin{pmatrix} \boldsymbol{f}_1 \\ \vdots \\ \boldsymbol{f}_{N_\xi} \end{pmatrix} \in \mathbb{R}^{N_u N_\xi}, \tag{6.10}$$

with each block has

$$\begin{aligned} (\boldsymbol{g}_l)_j &= -\int_{\partial D \times \Gamma} \psi_l(\boldsymbol{\xi}) g(\boldsymbol{x}) \boldsymbol{n} \cdot \boldsymbol{\varphi}_j(\boldsymbol{x})\, \mathbf{d}\boldsymbol{x}\, \mathbf{d}\boldsymbol{\xi}, \qquad j = 1, \cdots, N_q, \\ (\boldsymbol{f}_l)_j &= -\int_{D \times \Gamma} \psi_l(\boldsymbol{\xi}) f(\boldsymbol{x}) \phi_j(\boldsymbol{x})\, \mathbf{d}\boldsymbol{x}\, \mathbf{d}\boldsymbol{\xi}, \qquad j = 1, \cdots, N_u, \end{aligned} \tag{6.11}$$

for $l = 1, \cdots, N_\xi$.

An alternative way to write equation (6.7) is to rearrange the order of its solution vectors such that

$$\mathcal{C}\, \tilde{\boldsymbol{u}} = \tilde{\boldsymbol{f}}, \qquad \text{with} \quad \mathcal{C} = I \otimes C_0 + \sum_{s=1}^{L} G_s \otimes C_s, \tag{6.12}$$

with

$$C_0 = \begin{pmatrix} A_0 & B_0^\intercal \\ B_0 & 0 \end{pmatrix}, \qquad C_s = \begin{pmatrix} A_s & 0 \\ 0 & 0 \end{pmatrix}, \quad s = 1, \cdots, L. \tag{6.13}$$

The solution vector $\tilde{\boldsymbol{u}}$ and the right-hand side $\tilde{\boldsymbol{f}}$ can be written into blocks

$$\tilde{\boldsymbol{u}} = \begin{pmatrix} \tilde{\boldsymbol{u}}_1 \\ \vdots \\ \tilde{\boldsymbol{u}}_{N_\xi} \end{pmatrix} \in \mathbb{R}^{(N_q+N_u)N_\xi}, \quad \tilde{\boldsymbol{f}} = \begin{pmatrix} \tilde{\boldsymbol{f}}_1 \\ \vdots \\ \tilde{\boldsymbol{f}}_{N_\xi} \end{pmatrix} \in \mathbb{R}^{(N_q+N_u)N_\xi}, \tag{6.14}$$

with each block has

$$\tilde{\boldsymbol{u}}_l = \begin{pmatrix} \boldsymbol{q}_l \\ \boldsymbol{u}_l \end{pmatrix}, \qquad \tilde{\boldsymbol{f}}_l = \begin{pmatrix} \boldsymbol{g}_l \\ \boldsymbol{f}_l \end{pmatrix}, \qquad l = 1, \cdots, N_\xi. \tag{6.15}$$

In addition, we now explore the tensor product form by reshaping the solution vector $\tilde{\boldsymbol{u}} \in \mathbb{R}^{(N_q+N_u)N_\xi}$ and the right-hand side $\tilde{\boldsymbol{f}} \in \mathbb{R}^{(N_q+N_u)N_\xi}$ in (6.12) into $(N_q + N_u) \times N_\xi$ matrices

$$\tilde{U} = \begin{pmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_{N_\xi} \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_{N_\xi} \end{pmatrix}_{(N_q+N_u)\times N_\xi}, \quad \tilde{F} = \begin{pmatrix} \boldsymbol{g}_1 & \boldsymbol{g}_2 & \cdots & \boldsymbol{g}_{N_\xi} \\ \boldsymbol{f}_1 & \boldsymbol{f}_2 & \cdots & \boldsymbol{f}_{N_\xi} \end{pmatrix}_{(N_q+N_u)\times N_\xi}, \tag{6.16}$$

then the matrix-vector equation (6.12) is equivalent to a matrix-matrix equation:

$$\mathcal{C}(\tilde{U}) = \tilde{F}, \qquad \text{with} \quad \mathcal{C}(\tilde{U}) = \sum_{s=0}^{L} C_s \tilde{U} G_s^{\intercal}. \tag{6.17}$$

6.3 Preconditioning

In this section we will discuss the iterative solvers of the linear system (6.17) or (6.12) with different preconditioners. From now on unless specified otherwise, we will use the matrix representation (6.12) since it is simple and coincides with the matrix form in (3.3).

6.3.1 A Mean-Based Diagonal Preconditioner

Recall that in Section 4.1 the mean-based preconditioner $I \otimes K_0$ is the first term of the summation of $\mathcal{A}$, and it's also the only term that contains the mean information. Similarly the diagonal information $I \otimes C_0$ in the matrix $\mathcal{C}$ would be an efficient preconditioner to the linear system with mixed

formulation as well, especially when σ is small. Based on the deterministic preconditioner to the traditional saddle-point problem [34], a mean-based diagonal preconditioner $\mathcal{P}_D$ is presented in [19] with the following formula:

$$\mathcal{P}_D = I \otimes \begin{pmatrix} D_0 & 0 \\ 0 & V \end{pmatrix}. \tag{6.18}$$

D_0 is the diagonal of A_0, and V is a symmetric positive definite approximation to the Schur complement $S_{D_0} = B_0 D_0^{-1} B_0^\intercal$, which is a sparse approximation to the true Schur complement $S = B_0 A_0^{-1} B_0^\intercal$ of the deterministic saddle-point system. In practice, the action V^{-1} on a vector is defined as one V-cycle of multigrid method applied to a linear system with coefficient matrix S_{D_0}.

6.3.2 A Mean-Based Triangular Preconditioner

We present a new mean-based triangular preconditioner $\mathcal{P}_0$ in this subsection. Instead of using block diagonal preconditioning for C_0 in Subsection 6.3.1, we shall derive a preconditioning scheme for the stochastic Galerkin matrix with a triangular preconditioner for the deterministic saddle-point system C_0 [8].

Since $D_0 = \mathrm{diag}(A_0)$ is spectrally equivalent to A_0, we can consider the following block decomposition

$$\begin{pmatrix} D_0 & B_0^\intercal \\ B_0 & 0 \end{pmatrix} \begin{pmatrix} I & D_0^{-1} B_0^\intercal \\ 0 & -I \end{pmatrix} = \begin{pmatrix} D_0 & 0 \\ B_0 & S_{D_0} \end{pmatrix} \tag{6.19}$$

to obtain an approximate triangular preconditioner where

$$S_{D_0} = B_0 D_0^{-1} B_0^\intercal.$$

It is easy to verify that S_{D_0} is spectrally equivalent to $S = B_0 A_0^{-1} B_0^{\intercal}$.

Therefore we can define the triangular preconditioner for the deterministic saddle-point system C_0 as

$$\begin{pmatrix} I & D_0^{-1} B_0^{\intercal} \\ 0 & -I \end{pmatrix} \begin{pmatrix} D_0 & 0 \\ B_0 & S_{D_0} \end{pmatrix}^{-1}. \tag{6.20}$$

And the mean-based triangular preconditioner to the stochastic Galerkin system (6.12) can be define as

$$\mathcal{P}_0 = I \otimes \begin{pmatrix} I & D_0^{-1} B_0^{\intercal} \\ 0 & -I \end{pmatrix} \begin{pmatrix} D_0 & 0 \\ B_0 & S_{D_0} \end{pmatrix}^{-1}. \tag{6.21}$$

6.3.3 A Block Triangular Preconditioner

Similar to the discussion in Section 4.3, once we have a mean-based preconditioner $\mathcal{P}_0$, we can derive a block triangular preconditioner to the stochastic Galerkin matrix (6.12) based on the odd-even indices simplex representation in Subsection 3.3.4.

Due to the block structure discussed in 3.3.4, the stochastic Galerkin matrix $\mathcal{C}$ can be written as the following block structures:

$$\mathcal{C} = \begin{bmatrix} D_1 & W^T \\ W & D_2 \end{bmatrix}, \tag{6.22}$$

where

$$D_1 = I_1 \otimes C_0, \quad D_2 = I_2 \otimes C_0, W = \left(\sum_{s=1}^{L} G_s \otimes C_s \right) (\mathcal{I}_2, \mathcal{I}_1) = \sum_{s=1}^{L} G_s(\mathcal{I}_2, \mathcal{I}_1) \otimes C_s,$$

and $\mathcal{I}_1$ is the one in $\{\mathcal{I}_{odd}, \mathcal{I}_{even}\}$ with larger cardinality, $\mathcal{I}_2$ is the one with smaller cardinality.

We define the block triangular preconditioner for the mixed formulation system as

$$\mathcal{P}_T := \begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix}. \tag{6.23}$$

The corresponding preconditioner system

$$\begin{bmatrix} D_1 & 0 \\ W & D_2 \end{bmatrix} \begin{bmatrix} \tilde{U}_1 \\ \tilde{U}_2 \end{bmatrix} = \begin{bmatrix} \tilde{F}_1 \\ \tilde{F}_2 \end{bmatrix}$$

may be solved inexactly by the standard multigrid V-cycle with the mean-based triangular preconditioner $\mathcal{D}$ acting on $\mathcal{I}_{odd},\ \mathcal{I}_{even}$ separately.

6.4 Numerical Examples

In this section, we compare the performances of different preconditioners (the mean-based diagonal preconditioner $\mathcal{P}_D$, the mean-based triangular preconditioner $\mathcal{P}_0$ and the block triangular preconditioner $\mathcal{P}_T$) on a benchmark problem in [18].

The steady flow comes from left to right on the domain $D = [0,1]^2$ with $f \equiv 0$, $\partial D_D = \{0,1\} \times [0,1]$ and $\partial D_N = \partial D \setminus \partial D_D$. Let $\vec{q} \cdot \vec{n} = 0$ at the two horizontal walls so that the flow is tangent to those boundaries. $u_D = 1$ on $\{0\} \times [0,1]$ and $u_D = 0$ on $\{1\} \times [0,1]$. Still we consider the same covariance function with $L_{c,1} = L_{c,2} = 1$, and mean $\mu = 1$.

We use Raviart-Thomas element space with $k = 0$ (RT0 element) as $\boldsymbol{V}_h$ and assemble the stiffness matrices $A_s, s = 0, 1, 2, \ldots, L$ using the software package iFEM [6]. We apply MINRES method with tolerance $= 10^{-8}$ for $\mathcal{P}_D$, and apply GMRES method with same tolerance and $\mathrm{restart} = 20$

for $\mathcal{P}_0$ and $\mathcal{P}_T$. For all three preconditioners, the diagonal matrices are inverted by the geometric multigrid method with $V(2,2)$. All computations are performed using MATLAB 2017b on a Lenove-Y40 with Intel Core i7-4510U CPU and 8GB of RAM.

Figure 6.1 shows the mean and variance of $\boldsymbol{u}$ on a 64×64 grid with $\sigma = 0.3$, $L = 4$, $p = 4$ and Gaussian random variables.

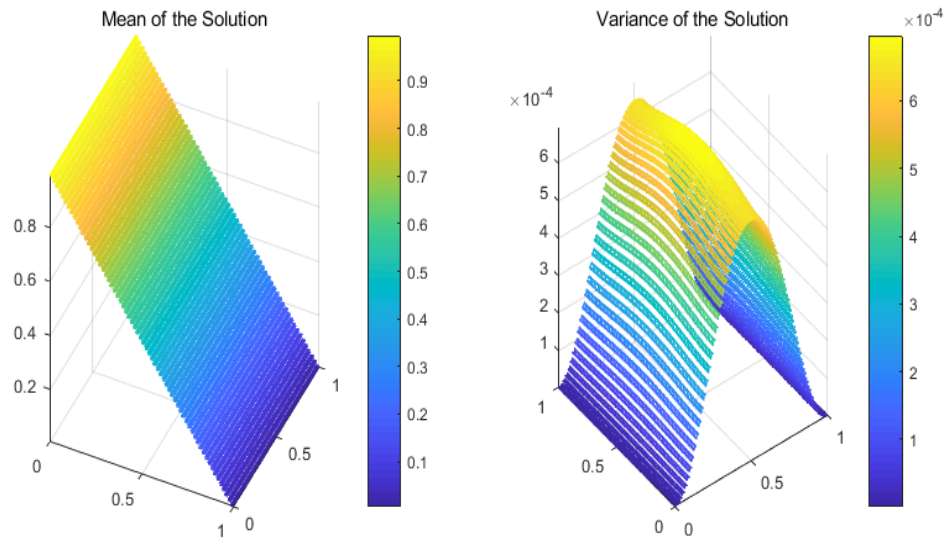


Figure 6.1: Mean (left) and variance (right) of $\boldsymbol{u}$ on a 64 $\times$ 64 grid with $L = 4$ and $p = 4$

6.4.1 Hermite Polynomials with Gaussian Random Variables

We start with Hermite polynomials with Gaussian random variables. Table 6.1 and 6.2 show the iteration steps and time costs for different preconditioners with various h, L and p values. For all preconditioners, the iteration steps are uniform against various mesh size h while the time costs scale linearly to h^2. All preconditioners are robust to h and iteration steps are positive correlated to p and L. The two new preconditioners we presented $\mathcal{P}_0$ and $\mathcal{P}_T$ outperform the original mean-based diagonal preconditioner $\mathcal{P}_D$. When L, p and σ are large, the best-performance GMRES with block triangular preconditioner $\mathcal{P}_T$ converges more than twice faster compared to MINRES with

the mean-based preconditioner $\mathcal{P}_D$.

Table 6.1: Iteration steps with various $h,\ L$ and p for all preconditioners (Hermite polynomials with $\mu = 1,\ \sigma = 0.3$)

		$\mathcal{P}_D$			$\mathcal{P}_0$			$\mathcal{P}_T$		
L	h	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
4	1/16	70	88	113	20	26	32	17	18	22
	1/32	71	89	116	20	25	32	16	18	22
	1/64	71	89	117	20	25	31	16	18	22
	1/128	71	89	116	20	25	31	17	18	21
6	1/16	71	91	119	20	26	34	17	19	23
	1/32	73	92	121	20	26	34	17	18	23
	1/64	74	92	123	20	26	33	17	18	23
	1/128	74	92	121	20	26	33	17	19	22

Table 6.2: Time costs with various $h,\ L$ and p for all preconditioners (Hermite polynomials with $\mu = 1,\ \sigma = 0.3$)

		$\mathcal{P}_D$			$\mathcal{P}_0$			$\mathcal{P}_T$		
L	h	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
4	1/16	0.6	1.2	3.0	0.3	0.6	1.1	0.2	0.5	1.0
	1/32	1.6	5.7	17.6	0.6	2.3	7.1	0.6	2.1	5.4
	1/64	8.9	23.8	55.4	4.2	12.1	27.5	3.5	9.7	23.1
	1/128	30.0	78.8	191.4	18.0	44.2	102.3	15.0	37.6	86.1
6	1/16	1.0	2.9	13.8	0.3	1.4	5.4	0.6	1.3	4.1
	1/32	4.4	17.5	47.0	1.8	7.4	24.7	1.6	5.8	20.1
	1/64	17.7	53.8	164.4	8.3	29.0	87.5	7.5	24.3	72.4
	1/128	59.8	202.3	653.7	32.4	112.6	333.0	29.6	104.3	268.2

6.4.2 Legendre Polynomials with Uniform Random Variables

We then switch to Legendre polynomials with Uniform random variables. Table 6.3 and 6.4 show the iteration steps and time costs for different preconditioners with various $h,\ L$ and p values. Similar to the Hermite case, for all preconditioners, the iteration steps are uniform against various mesh size h while the time costs scale linearly to h^2. All preconditioners are robust to h and

iteration steps are positive correlated to p and L. The block triangular preconditioner $\mathcal{P}_T$ is more robust to the degree p compared to the other two preconditioners. The two new preconditioners we presented $\mathcal{P}_0$ and $\mathcal{P}_T$ outperform the original mean-based diagonal preconditioner $\mathcal{P}_D$. When $L,\ p$ and σ are large, the best-performance GMRES with block triangular preconditioner $\mathcal{P}_T$ converges more than twice faster compared to MINRES with the mean-based preconditioner $\mathcal{P}_D$.

Table 6.3: Iteration steps with various $h,\ L$ and p for all preconditioners (Legendre polynomials with $\mu = 1,\ \sigma = 0.3$)

		$\mathcal{P}_D$			$\mathcal{P}_0$			$\mathcal{P}_T$		
L	h	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
4	1/16	81	101	122	24	29	35	17	20	24
	1/32	84	105	129	24	29	35	17	20	24
	1/64	84	106	130	24	29	35	17	20	24
	1/128	84	106	130	23	29	34	18	20	24
6	1/16	84	109	150	25	32	43	18	21	28
	1/32	87	115	161	25	32	46	18	22	29
	1/64	88	117	167	25	32	46	18	22	29
	1/128	88	117	168	25	32	44	18	22	29

Table 6.4: Time costs with various $h,\ L$ and p for all preconditioners (Legendre polynomials with $\mu = 1,\ \sigma = 0.3$)

		$\mathcal{P}_D$			$\mathcal{P}_0$			$\mathcal{P}_T$		
L	h	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$	$p=2$	$p=3$	$p=4$
4	1/16	0.7	1.5	3.2	0.2	0.6	1.5	0.2	0.6	1.0
	1/32	1.8	6.9	19.6	0.8	2.8	8.1	0.8	2.2	5.8
	1/64	10.6	29.0	58.5	4.9	14.0	31.9	3.9	11.2	24.3
	1/128	35.4	92.2	212.9	19.3	49.9	116.6	16.6	43.5	95.4
6	1/16	1.4	3.1	12.5	0.4	1.3	5.9	0.4	0.9	4.1
	1/32	3.5	15.5	50.4	1.5	8.0	30.1	1.2	6.5	20.8
	1/64	14.9	59.2	207.0	8.8	31.4	108.2	7.0	27.4	78.1
	1/128	62.8	253.4	885.9	33.4	126.5	459.2	28.4	114.0	346.6

Chapter 7

Conclusion

In this thesis we proposed a simplex lattice representation of the stochastic space with an odd-even ordering, and presented two new block preconditioners for the stochastic steady-state diffusion problem (primal form), including the two-term variance-involved preconditioner $\mathcal{P}_1$ and the block triangular preconditioner $\mathcal{P}_T$. Instead of containing only the mean information, the variance-involved block diagonal preconditioner captures most of the variance information in the random field. The simplex structure in the stochastic space $S_{p,L}$ is illustrated with an odd-even ordering, and an economic way is introduced to store the stochastic matrices as well as do the matrix multiplication. We compared the performances of all preconditioners with the mean-based preconditioner $\mathcal{P}_0$ proposed by [33] in two benchmark problems. The block triangular preconditioner $\mathcal{P}_T$ outperformed the mean-based preconditioner $\mathcal{P}_0$ by two times, especially when the KL-expansion term L and the total polynomial degree p is large, which coincides with the theoretical spectral analysis for different preconditioners. We then extended the stochastic Galerkin methods framework to the mixed formulation together with the triangular preconditioner $\mathcal{P}_T$. We also developed a new mean-based triangular preconditioner $\mathcal{P}_0$. Both $\mathcal{P}_0$ and $\mathcal{P}_T$ outperformed the original mean-based diagonal preconditioner $\mathcal{P}_D$. When $L,\ p$ and σ are large, both preconditioners converge at least twice faster than the mean-based diagonal preconditioner $\mathcal{P}_D$.

Bibliography

[1] O. Axelsson and R. Blaheta. Preconditioning of Matrices Partitioned in 2 by 2 Block Form: Eigenvalue Estimates and Schwarz DD for Mixed FEM. *Numerical Linear Algebra with Applications*, 17(5):787–810, 2010.

[2] I. Babuška, F. Nobile, and R. Tempone. A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data. *SIAM Review*, 52(2):317–355, 2010.

[3] A. Barth, C. Schwab, and N. Zollinger. Multi-level Monte Carlo Finite Element Method for Elliptic PDEs with Stochastic Coefficients. *Numerische Mathematik*, 119(1):123–161, 2011.

[4] R. Blaheta. GPCG-Generalized Preconditioned CG Method and Its Use with Non-linear and Non-symmetric Displacement Decomposition Preconditioners. *Numerical Linear Algebra with Applications*, 9(6-7):527–550, 2002.

[5] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2008.

[6] L. Chen. *i*FEM: an Integrated Finite Element Methods Package in MATLAB. *Technical Report, University of California at Irvine*, pages 1–35, 2009.

[7] L. Chen, L. Guo, X. Huang, and D. Wang. Approximate Block Factorization Preconditioner for Stochastic Galerkin Mixed Finite Element Methods (In Preparation).

[8] L. Chen, J. Hu, and X. Huang. Fast Auxiliary Space Preconditioners for Linear Elasticity in Mixed Form. *Mathematics of Computation*, 87(312):1601–1633, 2017.

[9] L. Chen, G. Lin, D. Wang, J. Xu, and B. Zheng. Block Triangular Preconditioning for Stochastic Galerkin Method (To be Submited).

[10] L. Chen and D. Wang. A New Block Diagonal Preconditioner for Spectral Stochastic Finite Element Methods (Submitted). *Numerical Linear Algebra with Applications*, 2018.

[11] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Society for Industrial and Applied Mathematics, 2002.

[12] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3–15, 2011.

[13] R. Courant and D. Hilbert. Methods of Mathematical Physics, 1989.

[14] M. K. Deb, I. M. Babuška, and J. T. Oden. Solution of Stochastic Partial Differential Equations Using Galerkin Finite Element Techniques. *Computer Methods in Applied Mechanics and Engineering*, 190(48):6359–6372, 2001.

[15] R. G. Durán. Mixed Finite Element Methods. pages 1–44. 2008.

[16] H. C. Elman, D. G. Furnival, and C. E. Powell. H(div) Preconditioning for a Mixed Finite Element Formulation of the Diffusion Problem with Random Data. *Mathematics of Computation*, 79(270):733–760, 2009.

[17] H. C. Elman and T. Su. A Low-Rank Multigrid Method for the Stochastic Steady-State Diffusion Problem. 20742:1–17, 2016.

[18] O. Ernst, C. Powell, D. Silvester, and E. Ullmann. Efficient Solvers for a Linear Stochastic Galerkin Mixed Formulation of Diffusion Problems with Random Data. *SIAM Journal on Scientific Computing*, 31(2):1424–1447, 2009.

[19] O. G. Ernst and E. Ullmann. Stochastic Galerkin Matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1848–1872, 2010.

[20] D. G. Feingold and R. Varga. Block Diagonally Dominant Matrices and Generalizations of the Gerschgorin Circle Theorem. *Pacific Journal of Mathematics*, 12(4):1241–1250, 1962.

[21] P. Frauenfelder, C. Schwab, and R. A. Todor. Finite Elements for Elliptic Problems with Stochastic Coefficients. *Computer Methods in Applied Mechanics and Engineering*, 194(2-5 SPEC. ISS.):205–228, 2005.

[22] R. Ghanem and S. Dham. Stochastic Finite Element Analysis for Multiphase Flow in Heterogeneous Porous Media. *Transport in Porous Media*, 32(3):239–262, 1998.

[23] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer New York, New York, NY, 1991.

[24] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan. Quasi-Monte Carlo Methods for Elliptic PDEs with Random Coefficients and Applications. *Journal of Computational Physics*, 230(10):3668–3694, 2011.

[25] D. H. Griffel and T. S. Chihara. An Introduction to Orthogonal Polynomials. *The Mathematical Gazette*, 63(425):222, 1979.

[26] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, 53(2):217–288, 2011.

[27] A. Klawonn. Block-Triangular Preconditioners for Saddle Point Problems with a Penalty Term. *SIAM Journal on Scientific Computing*, 19(1):172–184, 1998.

[28] M. Loève. *Probability Theory II*. Springer-Verlag New York, New York, 4 edition, 1978.

[29] V. Paulsen. *Completely Bounded Maps and Operator Algebras*. Cambridge University Press, Cambridge, 2003.

[30] M. F. Pellissetti and R. G. Ghanem. Iterative Solution of Systems of Linear Equations Arising in the Context of Stochastic Finite Elements. *Advances in Engineering Software*, 31(8):607–616, 2000.

[31] K. K. Phoon, S. P. Huang, and S. T. Quek. Implementation of Karhunen-Loeve expansion for simulation using a wavelet-Galerkin scheme. *Probabilistic Engineering Mechanics*, 17(3):293–303, 2002.

[32] R. Pogorzelski. Orthogonal Polynomials for Engineers and Physicists. *Proceedings of the IEEE*, 63(8):1264–1264, 1975.

[33] C. E. Powell and H. C. Elman. Block-diagonal Preconditioning for Spectral Stochastic Finite-Element Systems. *IMA Journal of Numerical Analysis*, 29(2):350–375, 2009.

[34] C. E. Powell and D. Silvester. Optimal Preconditioning for Raviart–Thomas Mixed Formulation of Second-Order Elliptic Problems. *SIAM Journal on Matrix Analysis and Applications*, 25(3):718–738, 2003.

[35] E. Rosseel and S. Vandewalle. Iterative Solvers for the Stochastic Finite Element Method. *SIAM Journal on Scientific Computing*, 32(1):372–397, 2010.

[36] Y. Saad and M. H. Schultz. GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[37] H. N. Salas. Gershgorin's Theorem for Matrices of Operators. *Linear Algebra and its Applications*, 291(1-3):15–36, 1999.

[38] V. Simoncini. Block Triangular Preconditioners for Symmetric Saddle-Point Problems. *Applied Numerical Mathematics*, 49(1):63–80, 2004.

[39] B. Sousedík, R. G. Ghanem, and E. T. Phipps. Hierarchical Schur Complement Preconditioner for the Stochastic Galerkin Finite Element Methods. *Numerical Linear Algebra with Applications*, 21(1):136–151, 2014.

[40] G. Stefanou. The Stochastic Finite Element Method: Past, Present and Future. *Computer Methods in Applied Mechanics and Engineering*, 198(9-12):1031–1051, 2009.

[41] C. Tretter. *Spectral Theory of Block Operator Matrices and Applications*. IMPERIAL COLLEGE PRESS, 2008.

[42] E. Ullmann. A Kronecker Product Preconditioner for Stochastic Galerkin Finite Element Discretizations. *SIAM Journal on Scientific Computing*, 32(2):923–946, 2010.

[43] C. F. Van Loan and N. Pitsianis. Approximation with Kronecker Products. In *Linear Algebra for Large Scale and Real-Time Applications*, pages 293–314. Springer Netherlands, Dordrecht, 1993.

[44] L. Wang. *Karhunen-Loeve Expansions and their Applications*. Phd thesis, London School of Economics and Political Science, 2008.

[45] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.