STOCHASTIC DISCONTINUOUS GALERKIN METHODS FOR PDE–BASED MODELS WITH RANDOM COEFFICIENTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

PELİN ÇİLOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
SCIENTIFIC COMPUTING

JUNE 2023

Approval of the thesis:

STOCHASTIC DISCONTINUOUS GALERKIN METHODS FOR PDE–BASED MODELS WITH RANDOM COEFFICIENTS

submitted by **PELİN ÇİLOĞLU** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Scientific Computing Department, Middle East Technical University** by,

Prof. Dr. A. Sevtap Selçuk-Kestel _____
Dean, Graduate School of **Applied Mathematics**

Assoc. Prof. Dr. Önder Türk _____
Head of Department, **Scientific Computing**

Assoc. Prof. Dr. Hamdullah Yücel _____
Supervisor, **Scientific Computing, METU**

Examining Committee Members:

Assoc. Prof. Dr. Serdar Göktepe
Department of Civil Engineering, METU _____

Assoc. Prof. Dr. Hamdullah Yücel
Scientific Computing, METU _____

Prof. Dr. Songül Kaya Merdan
Department of Mathematics, METU _____

Prof. Dr. Ayhan Aydın
Department of Mathematics, Atılım University _____

Assist. Prof. Dr. Firdevs Ulus
Department of Industrial Engineering, Bilkent University _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: PELİN ÇİLOĞLU

Signature :

ABSTRACT

STOCHASTIC DISCONTINUOUS GALERKIN METHODS FOR PDE–BASED MODELS WITH RANDOM COEFFICIENTS

Çiloğlu, Pelin

Ph.D., Department of Scientific Computing

Supervisor : Assoc. Prof. Dr. Hamdullah Yücel

June 2023, 157 pages

Uncertainty, such as uncertain parameters, arises from many complex physical systems in engineering and science, e.g., fluid dynamics, heat transfer, chemically reacting systems, underwater pollution, radiation transport, and oil field reservoirs. It is well known that these systems can be modeled by partial differential equations (PDEs) with random input data. However, the information available on the input data is very limited, which causes a high level of uncertainty in approximating the solution to these problems. Therefore, the idea of uncertainty quantification (UQ) has become a powerful tool to model such physical problems in the last decade.

In this thesis, the aim is the development, analysis, and application of stochastic discontinuous Galerkin method for partial differential equation (PDE)–based models with random coefficients. As a model, we first focus on the single convection diffusion equation containing uncertainty. To identify the random coefficients, we use the well–known technique Karhunen Loève (KL) expansion. Stochastic Galerkin (SG) approach, turning the original problem containing uncertainties into a large system of deterministic problems, is applied to discretize the stochastic domain, while a discontinuous Galerkin method is preferred for the spatial discretization due to its better convergence behaviour for convection dominated PDEs. A priori and a posteriori error estimates are also derived. SG method generally results in a large cou-

pled system of linear equations, the solution of which is computationally difficult to compute using standard solvers. Therefore, we provide low-rank iterative solvers for efficient computing of such solutions, which compute low-rank approximations to the solutions of those systems. Moreover, to overcome boundary and/or interior layers, localized regions where the derivative of the solution is large, an efficient adaptive algorithm is presented for the numerical solution of the parametric convection diffusion equations. On the other hand, certain parameters of a model are needed to be optimized in order to reach the desired target, for instance, the location where the oil is inserted into the medium, the temperature of a melting/heating process, or the shape of the aircraft wings. Therefore, we extend our findings to optimization problems and consider optimal control problems governed by convection diffusion equations involving random inputs.

Keywords: PDE–constrained optimization, uncertainty quantification, stochastic discontinuous Galerkin, error estimates, low–rank approximation, convection diffusion equation with random coefficients, adaptive finite elements

ÖZ

RASTGELE KATSAYILI KISMİ DİFERANSİYEL DENKLEM TABANLI MODELLER İÇİN STOKASTİK SÜREKSİZ GALERKİN YÖNTEMLERİ

Çiloğlu, Pelin
Doktora, Bilimsel Hesaplama Bölümü
Tez Yöneticisi : Doç. Dr. Hamdullah Yücel

Haziran 2023, 157 sayfa

Belirsiz parametreler gibi belirsizlik, mühendislik ve bilimdeki birçok karmaşık fiziksel sistemden kaynaklanır; örneğin, akışkanlar dinamiği, ısı transferi, kimyasal olarak reaksiyona giren sistemler, su altı kirliliği, radyasyon taşınımı ve petrol sahası rezervuarları. Bu sistemlerin rasgele girdi verileri ile kısmi diferansiyel denklemler ile modellenebileceği iyi bilinmektedir. Bununla birlikte, girdi verilerinde mevcut olan bilgiler çok sınırlıdır ve bu da bu problemlerin çözümüne yaklaşmada yüksek düzeyde belirsizliğe neden olur. Bu nedenle, belirsizlik ölçümü fikri, son on yılda bu tür fiziksel problemleri modellemek için güçlü bir araç haline geldi.

Bu tezde amaç, rastgele katsayılara sahip kısmi diferansiyel denklem (PDE) tabanlı modeller için stokastik süreksiz Galerkin yönteminin geliştirilmesi, analizi ve uygulanmasıdır. Bir model olarak, önce belirsizlik içeren tek konveksiyon difüzyon denklemine odaklanıyoruz. Rastgele katsayıları belirlemek için iyi bilinen Karhunen Loève (KL) genişletme tekniğini kullanıyoruz. Belirsizlikler içeren orijinal problemi büyük bir deterministik problemler sistemine dönüştüren Stokastik Galerkin (SG) yaklaşımı, stokastik alanı ayrıklaştırmak için uygulanırken, konveksiyon ağırlıklı PDE'ler için daha iyi yakınsama davranışı nedeniyle uzaysal ayrıklaştırma için süreksiz bir Galerkin yöntemi tercih edilir. Priori ve posteriori hata tahminleri de türetilir. SG yöntemi genellikle, çözümünün standart çözücüler kullanılarak hesaplanması zor olan

büyük bir birleşik lineer denklem sistemiyle sonuçlanır. Bu sebeple, bu sistemlerin çözümlerine düşük kerteli yaklaşımlar hesaplayan bu tür çözümlerin verimli bir şekilde hesaplanması için düşük kerteli yinelemeli çözücüler sağlıyoruz. Ayrıca, sınır ve/veya iç katmanların, çözümün türevinin büyük olduğu yerel bölgelerin üstesinden gelmek için, parametrik konveksiyon difüzyon denklemlerinin sayısal çözümü için etkili bir uyarlamalı algoritma sunulmuştur. Öte yandan, istenen hedefe ulaşmak için bir modelin belirli parametrelerinin, örneğin yağın ortama eklendiği konum, bir eritme/ısıtma işleminin sıcaklığı veya uçak kanatlarının şekli gibi bazı parametrelerin optimize edilmesi gerekir. Bu nedenle, bulgularımızı optimizasyon problemlerine genişlettik ve rastgele girdiler içeren konveksiyon difüzyon denklemleri tarafından yönetilen optimal kontrol problemlerini ele aldık.

Anahtar Kelimeler: PDE–kısıtlı eniyileme, belirsizlik ölçümü, stokastik kesintili Galerkin, hata tahminler, düşük kertegeli yaklaşımlar, rastgele katsayılı konveksiyon difüzyon denklemi, adaptif sonlu elemanlar

To my family

ACKNOWLEDGMENTS

I would like to express my very great appreciation to my thesis supervisor Assoc. Prof. Dr. Hamdullah Yücel for his patient guidance, enthusiastic encouragement and valuable advices during the development and preparation of this thesis. His willingness to give his time and to share his experiences has brightened my path.

I would also like to thank members of my thesis defense committee for their insightful comments and discussions.

I gratefully acknowledge the partial financial support of The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the project Numerical Studies for Petrol and Gas Reservoir Problems with project number TUBITAK 1001: 119F022, and Middle East Technical University (METU) under the project Numerical Studies of Korteweg-de Veries Equation with Random Input Data with the project number YÖP:705-2018-2820.

I am very thankful to all members of Institute of Applied Mathematics at Middle East Technical University for the pleasant atmosphere, and everybody who helped me to complete this thesis.

Last but not least, I owe my most profound gratitude to my family, especially my precious cat Şefim, for their continuous support through my entire life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Many physical systems in science and engineering, such as fluid dynamics, heat transfer, chemically reacting systems, underwater pollution, radiation transport, oil field reservoir, climate science, and structural mechanics, are modeled by partial differential equations (PDEs) together with appropriate initial and boundary conditions [48, 128, 162]. To simulate complex kinds of behaviour in the physical systems, one makes predictions and hypotheses about certain outputs of interest with the help of simulation of mathematical models. Basically, the scientific computation paradigm can be represented as in Figure 1.1 [58].

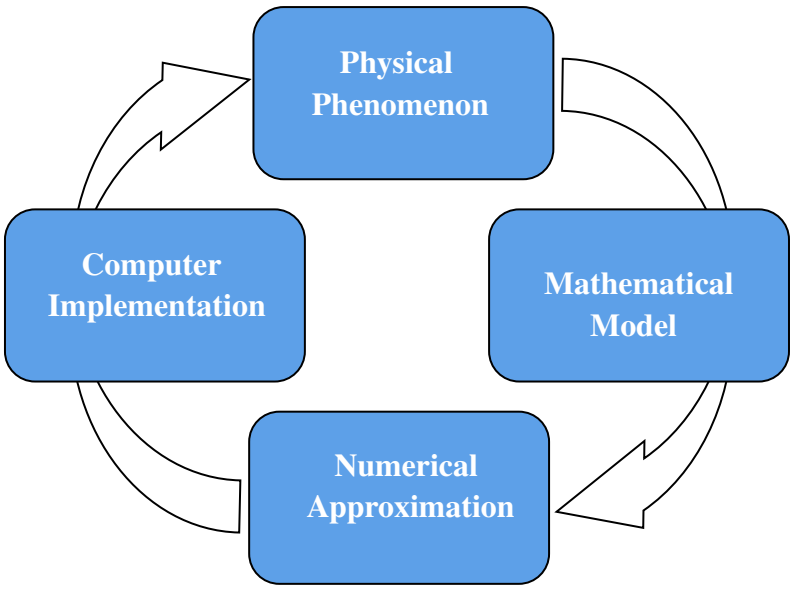

Figure 1.1: Computational science framework.

However, it is not always possible to precisely measure or determine the parameters, source functions or boundary conditions in the model due to lack of knowledge

(called as epistemic uncertainty), or the inherent variability (called as aleatoric uncertainty) of the model parameters. Epistemic uncertainties can be eliminated with additional measurements or improved measuring instruments, but such approaches are costly and also impractical to implement. Uncertainties arising from the nature of the parameters, aleatoric uncertainty, can only be resolved by determining suitable probability distributions for random parameters. Therefore, the idea of uncertainty quantification, i.e., quantifying the effects of uncertainty on the result of a computation, has become a powerful tool for modeling physical phenomena in the last decade [8, 11].

In mathematical models used in the oil and gas reservoir problems or in the management and remediation of groundwater resources, parameters are defined as random data to explain the limited knowledge of geological features and natural heterogeneity. In particular, taking the permeability parameter of the porous medium as random data is more suitable for the nature of the problem since the classical (deterministic) partial differential equations do not completely represent the actual behaviour of the physical phenomena [46, 114]. In such problems, the dissolved concentration corresponds to the solution of a convection diffusion equation, which includes the flow rate as random data. Therefore, in this thesis, the fast and reliable solution of convection diffusion equations containing random data is first emphasized.

In numerous applications, one may also rather be keen on deciding some unknown parameters in the model by comparing the anticipated reaction with actual estimators, or optimizing, certain parameters of the model. The motivation for this setup is uncertainty quantification in the complex PDE-based simulations, where it is crucial to account for imprecise or missing information in the input data. A well–known example can occur in the petroleum engineering, especially in the search for ways to extract more oil and gas from the earth's subsurface. A percent increase in the production may lead to a growth in profit of millions of dollars. A common process of oil recovery, called "water flooding", employs two wells: injection and production wells. While the production wells are used to transport gas and liquid from the reservoir, the injection wells inject water into the oil reservoir to maintain high pressure and adequate flow rate in the oil field. The balance between these wells is very crucial since redundant water is used while the oil production rate decreases. Consequently, there

may be a large amount of oil in the reservoir even the production has stopped. On the other hand, the significant developments in petroleum engineering, computer speed, storage capacity, and scientific computing techniques make it possible to manipulate and control the fluid flow paths through the oil reservoir. This ability provides a control strategy that will result in the maximization of oil recovery. In the literature, one of the well–known control approaches is the usage of an optimal control technique to increase the oil production rate. As a result, such scenario in the real–world can be mathematically expressed as PDE-constrained optimization under uncertainty, i.e., minimization of an objective functional subject to constraint equations in the form of PDEs [45, 113, 131, 142]:

$$\begin{aligned} &\text{minimize} \quad \mathcal{J} = \mathcal{J}\left(y(\omega), u(\omega), z(\omega)\right) \\ &\text{subject to} \quad c\left(y(\omega), u(\omega), z(\omega)\right) = 0, \end{aligned} \tag{1.1}$$

where the constraint c represents a PDE containing an uncertain parameter $z(\omega)$ and $\mathcal{J}$ is the corresponding cost functional. Here, u is the control variable, whereas y is the corresponding state variable. In this context, the numerical solutions of optimization problems constrained by convection diffusion equations containing random data and how to do the mathematical analysis will also be investigated in this thesis.

In the following section, we will review some relevant background concepts in the typical computational science framework with uncertainty.

1.1 Literature Review

This thesis combines numerous mathematical topics, including numerical discretization methods, probability theory, optimization theory, approximation theory, and theory of PDE. In the following, we will go over some relevant background concepts of PDEs with random coefficients, such as random field generation, solution representation, discretization of spatial and temporal domains, adaptive finite element methods, and PDE-constrained optimization with random inputs.

1.1.1 Generation of Random Fields

In this thesis, we mainly focus on the phase flow model in porous media with uncertain coefficients, which is a fundamental model in the oil reservoir simulation [46, 74]. It is crucial to describe how uncertain factors, such as permeability parameter, impact the pressure of the fluid [46, 114]. Permeability in fluid mechanics is a property of the porous medium that measures the capacity and ability of the formation to transmit fluids, that is, the capacity of a rock layer to transmit water or other fluids such as oil. However, it is hard to accurately measure the permeability field in the earth due to the large area of the oil reservoir and complicated earth structure. On the other hand, the permeability is a key parameter connecting the flow velocity and the gradient of pressure from Darcy's law. Therefore, the identification of random permeability is crucial for the oil industry.

Norbert Wiener developed homogeneous chaos or polynomial chaos expansion in 1938, which set the groundwork for the contemporary analysis of PDEs with unknown coefficients around the turn of the 20th century [148]. This polynomial chaos expansion, which still holds sway as a popular numerical approach for resolving PDEs with uncertain coefficients [98, 152], and offers a polynomial representation of Gaussian random fields. A special case of polynomial chaos expansion is Karhunen–Loève (KL) expansion [97, 116], which is a Fourier-like serie for representing a stochastic process as a linear combination of orthogonal functions. It is also known as proper orthogonal decomposition (POD), which decomposes the random dimensions and the spatial dimensions of the stochastic process. In the literature, there are also varied approaches to generate random variables, for instance, source point method [77], Kriging based method [132], Markov chain based method [161], and principle component analysis (PCA) [134, 135].

In this thesis, the well-known approach KL expansion will be used. In KL expansion, the random field can be represented as a linear combination of the eigenpair of the corresponding covariance function and the corresponding uncorrelated random variables; see Section 2.5.1 for more details.

1.1.2 Discretization of Probability Domain

PDEs with random input data are one of the most powerful tools in order to model the real-world problems. While solving a forward uncertainty quantification problem, the aim is to determine the effect of uncertainties in the input on the solution of the underlying problem or to investigate the numerical behaviour of the statistical moments of the solution, such as the mean and variance. Engineers and mathematicians have developed several methods to approximate the moments of solutions to such PDEs by means of quantifying uncertainty. Such numerical techniques can be divided into two main classes, which are intrusive and non–intrusive methods [107, 150]. While the intrusive methods are directly based on spectral expansions of the solution on the predefined stochastic subspace, the non-intrusive methods basically generate identically independent samples of the random data according to their probability distribution.

Monte Carlo (MC) method [67] is the most commonly used non–intrusive method for simulating PDEs with random coefficients since it is straightforward to apply as generating samples of the random input and utilizes repetitive deterministic systems for each realization. Although the MC method is very robust and independent of the dimensionality of the random domain, its convergence rate is slow. To obtain small error in the simulations, it requires a large amount of computation in the deterministic systems, i.e., the mean value converges as $1/\sqrt{N}$, where N is the number of realizations. Despite the convergence limitation of the standard Monte Carlo approach, there have been developments that improve its efficiency, such as quasi-Monte Carlo [124] and multi-level Monte Carlo methods [78, 88]. The other common non–intrusive method is stochastic collocation (SC) [9, 150, 151] which consists of a Galerkin approximation in the space and an interpolation of the stochastic domain. It evaluates solutions of a stochastic system at carefully chosen points (collocation points) within the random space. The reason being preferred is to require only a deterministic solver due to its non–intrusive structure as the MC methods and achieves fast convergence rate especially for the small number of realizations. However, the construction of collocation points is crucial because the choice of the collocation points determines the efficiency of the method [118, 126].

In contrast to the Monte Carlo approach and the stochastic collocation method, stochas-

tic Galerkin (SG) method [76], the most popular intrusive method, is a non–sampling approach, which transforms a PDE with random coefficients into a large system of coupled deterministic PDEs. As in the classic (deterministic) Galerkin method, the idea behind the stochastic Galerkin method is to seek a solution for the model equation such that the residue is orthogonal to the space of polynomials. An important feature of this technique is the separation of the spatial and stochastic variables, which allows a reuse of established numerical techniques. The studies in [11, 128] show that the stochastic Galerkin method generally exhibits superior performance compared to the stochastic collocation method for PDE-constrained optimization problems, in the sense that, unlike stochastic Galerkin method, the non-intrusivity property of the stochastic collocation method is lost when the moments of the state variable appear in the cost functional, or when the control is not stochastic. Lastly, the other intrusive method is the perturbation or Neumann series expansion methods [99, 8], which expand the exact random solution in power series of a small parameter about their respective mean values. However, the perturbation methods only give the good results for the problems containing small uncertainties; see, e.g., [84, 8]. Due to the aforementioned, the stochastic Galerkin method will be used in this thesis in order to quantify the uncertainty in the stochastic domain.

A major drawback of the stochastic Galerkin methods is the rapid increase of dimensionality, called as the curse of dimensionality. We address this issue by using low–rank Krylov subspace methods, which reduce both the storage requirements and the computational complexity by exploiting a Kronecker–product structure of system matrices; see, e.g., [13, 102, 139]. Similar approaches have been used to solve steady stochastic diffusion equations [55, 109, 122], unsteady stochastic diffusion equations [20], stochastic Navier–Stokes equations [64, 110], the optimal control problems for unconstrained control problems [17, 19, 21], and for control constraint problems [71]. Further, in the aforementioned studies, randomness is generally defined in the diffusion parameter; however we here consider the randomness both in diffusion or convection parameters.

1.1.3 Discretization of Spatial Domain

In the literature review done so far, the focus is only on quantifying the uncertainty in the stochastic domain. After discretization or approximation of the stochastic domain, it comes to the point that traditional numerical methods such as finite element, finite volume, or finite difference methods are applied to approximate the solution to the parameterized stochastic PDEs. In the literature, several stochastic finite element methods have been proposed and analysed; see, e.g., [11, 12, 58, 66, 68, 121, 153] and the references therein. Standard continuous finite element is the most commonly used method owing to its efficiency and high-order convergence rate by comparison with the finite difference or finite volume [11, 12, 99, 115]. However, they exhibit spurious oscillations while solving convection dominated PDEs, which are the main focus of this study. Therefore, we prefer to use discontinuous Galerkin methods [7, 54, 129, 136] for spatial discretization in this thesis. Compared with the discontinuous Galerkin method, the finite difference method is not able to handle complex geometries, the finite volume method is not capable of achieving high–order accuracy, and the standard continuous finite element method lacks the ability of local mass conservation. Moreover, especially for convection dominated problems, DG methods produce stable discretization without the need for stabilization strategies, and they allow for different orders of approximation to be used on different elements in a very straightforward manner [7, 129].

In the perspective of PDE-constrained optimization problems, optimal control problems governed by convection dominated PDEs have boundary and/or interior layers generated in the state PDE as well as in the adjoint PDE. The standard finite element methods may result in spurious oscillations causing in turn a severe loss of accuracy and stability. To overcome this problem, some effective stabilization techniques are used, i.e., the streamline upwind/Petrov Galerkin (SUPG) finite element method [50], the local projection stabilization [15], the edge stabilization [93, 155]. However, DG methods perform a better convergence behaviour for convection dominated optimal control problems since optimal convergence orders are obtained if the error is computed away from boundary or interior layers, in contrast to the SUPG discretization [111]. To best of our knowledge, there exist few papers related to the application of

the DG with random data [114, 115, 156].

1.1.4 Adaptivity in PDEs with Random Coefficients

With the improvement of computer–processing capacities, the demand for efficient numerical simulation of partial differential equations (PDEs) with uncertainty or parameter dependent inputs, which are widely used in many fields in science and engineering, see, e.g., [1, 53, 119, 163], has begun to grow. Even in the deterministic setting, the reliable and efficient solution of such PDEs, especially convection diffusion equations with random data, is challenging due to two possible issues. First of all, a stable numerical approach is generally not possible to approximate convection dominated problems since such problems may have rapid gradient changes in layers with small widths in solutions. The second obstacle is that to be able to obtain more accurate solutions, we need to include enough number of uncertain parameters in the system, which in turn engenders high computational cost. A natural solution to these issues is the combination of using meshes that are locally refined in the neighborhood of boundary layers and adaptively chosen index sets for the stochastic space.

Design and theoretical analysis of adaptive finite element methods pioneered by the work of Babuška and Rheinbold [10] have become a popular approach for the efficient solution of deterministic PDEs [3, 145] as well as PDEs with random data. Within the SG method setting for PDEs containing uncertainty, several adaptive strategies based on, for instance, implicit error estimators [146], goal–oriented a posteriori error estimates [28, 120], multilevel goal–oriented adaptive approaches [26], local equilibrium error estimates [62], hierarchal error estimates [28, 29, 30, 51], and residual–based error estimators [59, 60, 61, 79], are used to enhance the computed solution and drive the convergence of approximations. In addition to the aforementioned studies, the convergence analysis of adaptive SG methods are discussed in [60] for the residual–based estimators, in [27] for hierarchical a posteriori error indicators for parametric approximations, and in [26] for the multilevel construction of the spatial generalized polynomial chaos. However, according to our best literature review, there exist any study for the convection dominated PDEs containing uncertainty. Here, we intend to fill this gap. In this thesis, the stochastic discontinuous Galerkin method and

a posteriori error estimation will be combined to obtain an adaptive approximation for convection diffusion equations with random inputs. The SG method discretizes a parametric reformulation of the given PDE with random data and searches for approximations in tensor product spaces. Due to the enormous size of the space, computing the SG solution becomes unaffordable if a large number of random variables are used to represent the input data and highly resolved spatial grids are used for finite element approximations on the physical domain. Hence, with the help of judiciously chosen adaptive approaches in both spatial and stochastic domains, one can avoid a fast growth of the dimension of tensor basis consisting of finite element basis functions in the spatial domain and polynomial chaos polynomials in (stochastic) parameter space.

1.1.5 Optimal Control Problems with Random Coefficients

In many applications, optimization of many physical and engineering phenomena can be formulated as optimal control problems governed by partial differential equations which have been a major topic in the applied mathematics and control theory. The methodology of deterministic PDE-constrained optimization problems has been developed and investigated for several decades [31, 92, 113, 142]. However, the research of the optimal control problems governed by PDEs with random coefficients is still in its early stage. We refer to [21, 86, 94, 101, 131] and references therein for optimal control problems with random coefficients.

PDE–constraint optimization problems with uncertainty have been studied in various formulations in the literature, and these formulations can be categorized by assumptions on the cost functional in the minimization problem (1.1) as follows [5]:

1) Mean-based control: Replacing $z(\omega)$ by its expected value $\mathbb{E}[z(\omega)]$, minimize $\mathcal{J}(y, u, \mathbb{E}[z])$ by a deterministic optimal control; see, e.g., [32, 33].

2) Pathwise control: Fixing $z(\omega)$, minimize $\mathcal{J}(y, u, z(\omega))$, and then obtain a realization u^* of the stochastic optimal control $u(\omega)$; see, e.g., [5, 125].

3) Averaged control: Control the averaged state by minimizing $\mathcal{J}(\mathbb{E}[y(\omega)], u, z)$ using a deterministic optimal control; see, e.g., [106, 165].

4) Robust deterministic control: Minimize the expected cost $\mathbb{E}[\mathcal{J}(y,u,z)]$ by a deterministic optimal control; see, e.g., [35, **?**, 86, 94, 101, 108, 131].

5) Robust stochastic control: Minimize the expected cost $\mathbb{E}[\mathcal{J}(y,u,z)]$ by a stochastic optimal control; see, e.g., [18, 21, 45, 104, 105, 141].

Since the source of uncertainty in the state PDE is not well explained by the mean–based control issue, it is impossible to determine whether the optimal control problem is robust or not. When combined with Monte Carlo and stochastic collocation sampling techniques, the pathwise control problem is preferable. The expected value $\mathbb{E}[u^*]$ does not, however, solve an optimum control problem, hence it is not a robust value. In the average control problem, it minimizes the distance between the expected and a certain state, whereas the robust deterministic control seeks to minimize the expected distance between the random state and the desired state. In practice, controllers typically require a deterministic signal, therefore stochastic optimal controls have limited practical usage. Robust deterministic control, which includes an appropriate statistical measure of the objective function to be minimized, is more practical and realistic since randomness cannot be observed during the design of the control. Therefore, the interest in this thesis is the robust deterministic control problem subject to a convection diffusion equation containing uncertain inputs.

Finding an approximate solution for the optimization problems containing uncertainty (1.1) is extremely challenging and requires much more computational resources than the ones in the deterministic setting. In the literature, there exist various competing methods to solve such kinds of problem, for instance, Monte Carlo [5, 14, 87], stochastic collocation method [34, 72, 131, 141], and stochastic Galerkin method [56, 86, 108, 131, 140]. In this thesis, the stochastic Galerkin method is preferred as a stochastic method since it separates the stochastic and spatial domain, and to represent random coefficient, the KL expansion is used; see Section 1.1.2 and Section 1.1.1. On the other hand, for the discretization of the spatial domain, we use a discontinuous Galerkin method due to its better convergence behaviour for the optimization problems governed by convection dominated PDEs; see, e.g., [111, 157, 159]. We also refer to Section 1.1.3 and references therein for more details on the discontinuous Galerkin methods. To overcome the curse of dimensionality problem, we apply

a low–rank variant of generalized minimal residual (GMRES) method [133] with a suitable preconditioner.

1.2 Outline of Thesis

An outline of the thesis is as follows: In Chapter 2, some background information is provided on Kronecker product, low-rank approximation, function spaces, some important inequalities, and stochastic Galerkin approach. Then, in Chapter 3, the formulation of the stochastic discontinuous Galerkin methods is presented and applied to solve a convection diffusion equation with uncertain data. To avoid the curse of dimensionality, low–rank Krylov subspace methods are proposed by examining some numerical examples. Chapter 4 focuses on the development of the adaptive stochastic discontinuous Galerkin method and the derivation of a posteriori error estimation. In Chapter 5, the robust deterministic optimal control problem constrained by a convection diffusion equation containing uncertain coefficients is investigated theoretically and numerically. Finally, the thesis ends with Chapter 6, which includes the concluding remarks and future works.

CHAPTER 2

PRELIMINARIES

In this chapter, some important concepts, that are widely used in the rest of this thesis, are covered, and the fundamental notations are fixed for the reader's convenience. Firstly, important definitions and notations related to the solution of linear system are given in Section 2.1. To be able to analyze and solve partial differential equations (PDEs) containing uncertain terms, the essential function spaces in the physical and probability domains are presented in Section 2.2 and 2.3, respectively. Section 2.4 provides important inequalities, which are commonly utilized in the theoretical parts of this thesis. Finally, the basic procedure of the stochastic Galerkin method, which is a powerful tool for working with PDE with uncertainty, is reviewed in Section 2.5.

2.1 Matrix Computation on Kronecker Product

In this section, the Kronecker product of two matrices and its crucial properties are introduced. Also, the basic notations related to low–rank approach are given.

Definition 2.1.1. *Let $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$. Then, the Kronecker product $A \otimes B \in \mathbb{R}^{mp\times nq}$ is defined by*

$$A \otimes B = \begin{bmatrix} a_{11}B & \ldots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \ldots & a_{mn}B \end{bmatrix}. \tag{2.1}$$

The Kronecker product may alternatively be referred to as the direct product or the tensor product.

Definition 2.1.2. *Let* $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m \times n}$ *and* $Y = \left[\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T\right]^T \in \mathbb{R}^{mn}$. *Then, isomorphic mapping between* $\mathbb{R}^{mn}$ *and* $\mathbb{R}^{m \times n}$ *are defined, respectively, as following*

$$\texttt{vec} : \mathbb{R}^{m \times n} \to \mathbb{R}^{mn}, \qquad \texttt{mat} : \mathbb{R}^{mn} \to \mathbb{R}^{m \times n},$$

such that

$$\texttt{vec}(X) = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}, \qquad \textit{and} \qquad \texttt{mat}(Y) = \begin{bmatrix} \boldsymbol{y}_1 & \ldots & \boldsymbol{y}_n \end{bmatrix}. \tag{2.2}$$

From (2.2), it is clear that the `vec` operator stacks the columns of a matrix into a column vector, whereas the `mat` operator transforms a column vector to a matrix.

Next, a definition of the Kronecker (tensor) rank of a vectorized matrix is given.

Definition 2.1.3. *([83]) Given a matrix* $X \in \mathbb{R}^{m \times m}$ *and its vectorized form* $\mathbf{x} = \texttt{vec}(X) \in \mathbb{R}^{m^2}$. *The smallest* $r \in \mathbb{Z}_+$ *is called as the Kronecker rank of* $\mathbf{x}$ *so that*

$$\mathbf{x} = \sum_{i=1}^{r} u_i \otimes v_i, \qquad u_i, v_i \in \mathbb{R}^m. \tag{2.3}$$

Especially, the rank of the matrix X *corresponds to the Kronecker rank of the its vectorized matrix* $\mathbf{x}$.

2.1.1 Basic Properties

Here, some basic properties of the Kronecker product are listed; see, e.g., [82, 102]. Let A be an $m \times n$ matrix, B be a $p \times q$ matrix, C be an $s \times t$ matrix, D be a $d \times e$ matrix, and X be an $m \times p$ matrix:

- $(A \otimes B) \otimes C = A \otimes (B \otimes C)$,
- $(cA) \otimes B = c(A \otimes B) = A \otimes (cB)$ for all $c \in \mathbb{C}$,
- $(A \otimes B)(C \otimes D) = AC \otimes BD$,
- $\texttt{vec}(AXB) = (B^T \otimes A)\texttt{vec}(X)$,
- $(A \otimes B)^T = A^T \otimes B^T$,

- $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

The proof of all cases is usually simple and is left up to the reader's interpretation; see, e.g., [82].

2.1.2 Low–Rank Approximation

A low–rank approximation of a matrix is applicable to many fields in science and engineering. In particular, the scope of this thesis is taking advantage of the low–rank approximation to the solutions of large linear systems in order to reduce both the storage requirements and the computational complexity by exploiting a Kronecker–product structure of system matrices; see, e.g., [13, 102, 139].

Lemma 2.1.4. *A matrix $A \in \mathbb{R}^{m \times n}$ of rank r can be represented as a factorization of the form*

$$A = BC^T, \qquad B \in \mathbb{R}^{m \times r}, \quad C \in \mathbb{R}^{n \times r}.$$

If rank$(A) \ll m, n$, the matrix A has low-rank.

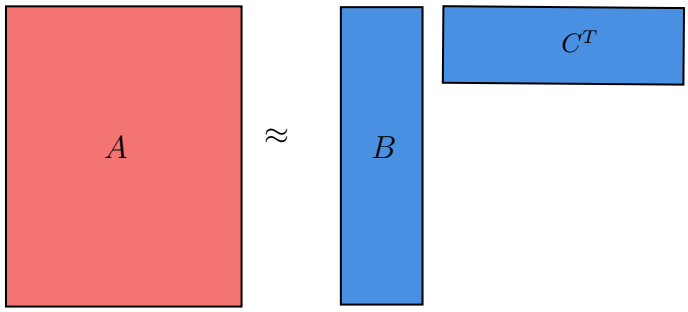

Figure 2.1: Approximation of a matrix A by its low–rank components B and C.

The illustration of a low–rank approximation of a matrix A is given in Figure 2.1. Low–rank approximation is a minimization problem where the difference between the provided matrix A, or array, and the approximating matrix with a lower rank is measured by the cost function. The singular value decomposition (SVD) provides an analytical solution to this minimization issue, which is given by the Eckart-Young-Mirsky theorem:

Theorem 2.1.5. *([57, 80]) Assume that the SVD of a matrix $A \in \mathbb{R}^{m \times n}$ with $r =$*

rank(A) *is given by*

$$A = U\Sigma V^T = \begin{bmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_r^T \end{bmatrix},$$

where $\sigma_i \in \mathbb{R}$ *are the singular values of* A *with* $\sigma_1 > \sigma_2 > \ldots > \sigma_r > 0$*, and* $U \in \mathbb{R}^{m\times m}$ *and* $V \in \mathbb{R}^{n\times n}$ *are orthogonal matrices. If* $A_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \, v_i^T$*, then*

$$\min_{X \text{ s.t. } rank(X)\le k} \|A - X\|_F = \|A - A_k\|_F.$$

The matrix A_k is also called the best approximation of A in the set of all rank–k matrices with respect to the Frobenious norm, that is, $\|A\|_F = \sqrt{\langle A, A\rangle_F}$ with $\langle A, B\rangle_F = \text{trace}(A^T B)$. For a tensor product matrix, or a multi-dimensional array, its low–rank approximation will be discussed in Section 3.3.

2.2 Sobolev Spaces

Throughout this thesis, $\mathcal{D}$ stands for a bounded polygonal domain in $\mathbb{R}^2$. The vector spaces $L^p(\mathcal{D})$ are given by

$$L^p(\mathcal{D}) = \left\{ v \text{ Lebesgue measurable} \,:\, \|v\|_{L^p(\mathcal{D})} < \infty \right\},$$

which is endowed with the norm

$$\|v\|_{L^p(\mathcal{D})} = \left(\int_{\mathcal{D}} |v(x)|^p \right)^{1/p},$$

and

$$\|v\|_{L^\infty(\mathcal{D})} = \operatorname{ess\,sup} \left\{ |v(x)| \,:\, x \in \mathcal{D} \right\}.$$

The following introduces the space of locally integrable functions and the space of infinitely many times differentiable functions with the compact support on a subset of $\mathcal{D}$, respectively:

$$L^p_{loc} = \left\{ v \text{ Lebesgue measurable} \,:\, v \in L^p(K) \text{ for all } K \subset \mathcal{D} \text{ compact} \right\},$$

and

$$C_c^\infty = \left\{ v \in C^\infty(\overline{\mathcal{D}}) \,:\, \text{supp}(v) \subset \mathcal{D} \text{ compact} \right\}.$$

Definition 2.2.1 (Weak Derivative)**.** *Let $u, v \in L^1_{loc}(\mathcal{D})$, and α be a multi-index. $D^\alpha u := v$ is called the α-th weak partial derivative of u if*

$$\int_{\mathcal{D}} u D^\alpha \psi \, dx = (-1)^{|\alpha|} \int_{\mathcal{D}} v\psi \, dx, \quad \forall \psi \in C_c^\infty(\mathcal{D}),$$

where $|\alpha| = \alpha_1 + \ldots + \alpha_n$ for $\alpha = (\alpha_1, \ldots, \alpha_n)$.

Now, the Sobolev space $W^{k,p}$ is defined by

$$W^{k,p}(\mathcal{D}) = \{v \in L^p(\mathcal{D}) \, : \, v \text{ has weak derivatives } D^\alpha v \in L^p(\mathcal{D}), \ \forall |\alpha| \leq k\},$$

with the Sobolev norm and the Sobolev semi–norm, respectively,

$$\|v\|_{W^{k,p}(\mathcal{D})} := \begin{cases} \left(\sum\limits_{|\alpha| \leq k} \int\limits_{\mathcal{D}} |D^\alpha v|^p \right)^{1/p}, & 1 \leq p \leq \infty, \\ \sum\limits_{|\alpha| \leq k} \operatorname*{ess\,sup}\limits_{\mathcal{D}} |D^\alpha v|, & p = \infty, \end{cases}$$

$$|v|_{W^{k,p}(\mathcal{D})} := \begin{cases} \left(\sum\limits_{|\alpha| = k} \int\limits_{\mathcal{D}} |D^\alpha v|^p \right)^{1/p}, & 1 \leq p \leq \infty, \\ \sum\limits_{|\alpha| = k} \operatorname*{ess\,sup}\limits_{\mathcal{D}} |D^\alpha v|, & p = \infty. \end{cases}$$

Remark 2.2.2. *For $p = 2$, it can be written as*

$$H^k(\mathcal{D}) = W^{k,2}, \ k = 0, 1, \ldots,$$

called a Hilbert space, with the following associated norm and semi-norm

$$\|v\|_{H^k(\mathcal{D})} := \left(\sum_{|\alpha| \leq k} \|D^\alpha v\|^2_{L^2(\mathcal{D})} \right)^{1/2},$$

$$|v|_{H^k(\mathcal{D})} := \left(\sum_{|\alpha| = k} \|D^\alpha v\|^2_{L^2(\mathcal{D})} \right)^{1/2},$$

respectively.

Moreover, $H^1_0(\mathcal{D})$ represents a Hilbert space with homogeneous Dirichlet boundary condition defined as

$$H^1_0(\mathcal{D}) = \left\{ u \in H^1(\mathcal{D}) \, : \, u|_{\partial\mathcal{D}} = 0 \right\}.$$

Then, the broken Sobolev space is given by

$$H^k(\mathcal{T}_h) = \left\{ v \in L^2(\mathcal{D}) \,:\, \forall K \in \mathcal{T}_h,\ v|_K \in H^k(K) \right\},$$

where $\mathcal{T}_h$ is the subdivision domain obtained by dividing $\mathcal{D}$ into triangle elements K. So, the broken Sobolev norm becomes

$$\|v\|_{H^k(\mathcal{T}_h)} = \left(\sum_{K \in \mathcal{T}_h} \|v\|^2_{H^k(K)} \right)^{1/2}.$$

2.3 Stochastic Sobolev Spaces

Necessary spaces and stochastic notations, which are used in stochastic discretizations, are introduced in this section. The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a complete probability space, where Ω is a sample space of events, $\mathcal{F} \subset 2^\Omega$ denotes a σ–algebra, and $\mathbb{P} : \mathcal{F} \to [0,1]$ is the associated probability measure. The space of all real–value square integrable random variable is defined as

$$L^2(\Omega) = L^2(\Omega, \mathcal{F}, \mathbb{P}) := \left\{ X : \Omega \to \mathbb{R} \,:\, \int_\Omega |X(\omega)|^2 \, d\mathbb{P}(\omega) < \infty \right\},$$

which is a Hilbert space equipped with the inner product

$$\langle X, Y \rangle = \int_\Omega X(\omega) Y(\omega) d\mathbb{P}(\omega), \quad X, Y \in L^2(\Omega).$$

Further, for a given separable Hilbert space H equipped with the norm $\|\cdot\|_H$ and seminorm $|\cdot|_H$, the Bochner–type space $L^p(H;\Omega)$ for a random variable $X : \Omega \to H$ is introduced as

$$L^p(H;\Omega) := \left\{ X : \Omega \to H : \|X\|_{L^p(H;\Omega)} < \infty \right\},$$

where

$$\|X\|_{L^p(H;\Omega)} = \begin{cases} \left(\int_\Omega \|X(\omega)\|_H^p \, d\mathbb{P}(\omega) \right)^{1/p}, & \text{for } 1 \le p < \infty, \\ \operatorname*{ess\,sup}_{\omega \in \Omega} \|X(\omega)\|_H, & \text{for } p = \infty. \end{cases}$$

A generic random field z on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is denoted by $z(\boldsymbol{x}, \omega) : \mathcal{D} \times \Omega \to \mathbb{R}$. For a fixed $\boldsymbol{x} \in \mathcal{D}$, $z(\boldsymbol{x}, \cdot) \in L^2(\Omega)$ is a real–value square integrable

random variable. Then, the mean $\mathbb{E}[z]$, the standard deviation $\mathbb{S}(z)$, and the corresponding variance $\mathbb{V}(z)$ for any random field z, are given, respectively, by

$$\mathbb{E}[z] = \int_{\Omega} z \, d\mathbb{P}(\omega), \tag{2.4a}$$

$$\mathbb{S}(z) = \left[\int_{\Omega} (z - \mathbb{E}[z])^2 \, d\mathbb{P}(\omega) \right]^{1/2}, \tag{2.4b}$$

$$\mathbb{V}(z) = [\mathbb{S}(z)]^2 = \mathbb{E}[z^2] - (\mathbb{E}[z])^2 . \tag{2.4c}$$

Also, the covariance of z is denoted by

$$\mathbb{C}_z(\boldsymbol{x}, \mathbf{y}) := \big\langle \big(z(\boldsymbol{x}, \cdot) - \mathbb{E}[z(\boldsymbol{x})]\big)\big(z(\boldsymbol{y}, \cdot) - \mathbb{E}[z(\boldsymbol{y})]\big) \big\rangle \quad \boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}. \tag{2.5}$$

In the following, the tensor–product space $H^k(\mathcal{D}) \otimes L^2(\Omega)$ of random fields can be stated as

$$H^k(\mathcal{D}) \otimes L^2(\Omega) = \left\{ z : \mathcal{D} \otimes \Omega \to \mathbb{R} \ : \ \int_{\Omega} \|z(\cdot, \omega)\|^2_{H^k(\mathcal{D})} d\mathbb{P}(\omega) < \infty \right\}, \tag{2.6}$$

which is equipped with the norm

$$\|z\|_{H^k(\mathcal{D}) \otimes L^2(\Omega)} := \left(\int_{\Omega} \|z(\cdot, \omega)\|^2_{H^k(\mathcal{D})} \, d\mathbb{P}(\omega) \right)^{1/2}. \tag{2.7}$$

In addition, the following isomorphism relation [11, 12] holds

$$H^k(\mathcal{D}) \otimes L^2(\Omega) \simeq L^2(H^k(\mathcal{D}); \Omega) \simeq H^k(\mathcal{D}; L^2(\Omega))$$

with the definitions

$$L^2(H^k(\mathcal{D}); \Omega) = \left\{ z : \mathcal{D} \times \Omega \to \mathbb{R} \ : \ \int_{\Omega} \|z(\cdot, \xi)\|^2_{H^k(\mathcal{D})} d\mathbb{P}(\omega) < \infty \right\}$$

and

$$\begin{aligned} H^k(\mathcal{D}; L^2(\Omega)) = \big\{ & z : \mathcal{D} \times \Omega \to \mathbb{R} \ : \ \forall |\alpha| \le k, \ \exists \, \partial^\alpha z \in L^2(\mathcal{D}) \otimes L^2(\Omega) \text{ with} \\ & \int_{\Omega} \int_{\mathcal{D}} \partial_\alpha z(x, \omega) \varphi(x, \omega) \, dx \, d\mathbb{P}(\omega) = (-1)^\alpha \int_{\Omega} \int_{\mathcal{D}} z(x, \omega) \partial_\alpha \varphi(x, \omega) \, dx \, d\mathbb{P}(\omega), \\ & \forall \varphi \in C_0^\infty(\mathcal{D} \times \Omega) \big\} . \end{aligned}$$

Thus, for a given complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathcal{D} \subset \mathbb{R}^n$, the stochastic Sobolev space is defined by

$$\begin{aligned} L^p(W^{k,q}(\mathcal{D}); \Omega) = \big\{ & z : \mathcal{D} \times \Omega \to \mathbb{R} \text{ Lebesgue measurable } : \\ & \int_{\Omega} \|z(\cdot, \xi)\|^p_{W^{k,q}(\mathcal{D})} d\mathbb{P}(\omega) < \infty \big\}. \end{aligned}$$

2.4 Important Inequalities

This section presents some significant inequalities that are commonly utilized in this thesis.

2.4.1 Trace Inequalities

Theorem 2.4.1. *[129, Theorem 2.5] Given a bounded domain $\mathcal{D}$ with polygonal boundary $\partial\mathcal{D}$ and outward normal vector* $\mathbf{n}$*, there exists trace operators* $\mathtt{tr}_0 : H^k(\mathcal{D}) \to H^{k-\frac{1}{2}}$ *for* $k > \frac{1}{2}$ *and* $\mathtt{tr}_1 : H^k(\mathcal{D}) \to H^{k-\frac{3}{2}}$ *for* $k > \frac{3}{2}$*, that are, respectively, extensions of the boundary values and boundary normal derivatives. Moreover, if* $v \in C^1(\overline{\mathcal{D}})$*, the following assumptions are satisfied:*

$$\mathtt{tr}_0 v = v|_{\partial D}, \qquad \mathtt{tr}_1 v = \nabla v \cdot \mathbf{n}|_{\partial D}.$$

Moreover, for positive constant c_{tr} independent of diameter of K, h_K, the trace inequalities are given as follow:

$$\|v\|_E^2 \leq c_{tr}\left(\|v\|_{L^2(K)}^2 + h_K|v|_{H^1(K)}^2\right), \quad v \in H^1(K), \tag{2.8a}$$

$$\|\nabla v \cdot \mathbf{n_E}\|_E^2 \leq c_{tr}\left(|v|_{H^1(K)}^2 + h_K|v|_{H^2(K)}^2\right), \quad v \in H^2(K), \tag{2.8b}$$

where E denotes an edge or a face of a triangle K. Let $\mathbb{P}^\ell(K)$ be the space of polynomials on K of degree less than or equal to ℓ. Then, the trace inequalities become

$$\|v\|_E^2 \leq c_{tr} h_K^{-\frac{1}{2}} \|v\|_{L^2(K)}, \quad v \in \mathbb{P}^\ell(K), \tag{2.8c}$$

$$\|\nabla v \cdot \mathbf{n_E}\|_E^2 \leq c_{tr} h_K^{-\frac{1}{2}} \|\nabla v\|_{L^2(K)}, \quad v \in \mathbb{P}^\ell(K). \tag{2.8d}$$

2.4.2 Inverse Inequality

There is a positive constant c_{inv} independent of h_K such that for any polynomial function v defined on K, the inverse inequality is given as follows (see, e.g., [36, Section 4.5]):

$$|v|_{j,K} \leq c_{inv}\, h^{i-j}|v|_{i,K}, \qquad 0 \leq i \leq j \leq 2. \tag{2.9}$$

2.4.3 Well–known Inequalities in Finite Element Analysis

Hölder's, Cauchy-Schwarz's, Young's, Poincaré's, and Jensen's inequalities are the most well-known inequalities used in the analysis of numerical methods. Below, their basic descriptions will be given.

- **Hölder's inequality:**
  Let $1 \leq p, q \leq \infty$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then, it holds

$$\int_{\mathcal{D}} |u\, v| \leq \|u\|_{L^p(\mathcal{D})} \|v\|_{L^q(\mathcal{D})}, \qquad u \in L^p(\mathcal{D}),\ v \in L^q(\mathcal{D}). \tag{2.10}$$

- **Cauchy-Schwarz's inequality:**
  For $p = q = 2$, Hölder's inequality (2.10) becomes

$$|\,(u, v)_{\mathcal{D}}\,| \leq \|u\|_{L^2(\mathcal{D})} \|v\|_{L^2(\mathcal{D})}, \qquad \forall u, v \in L^2(\mathcal{D}). \tag{2.11}$$

- **Young's inequality:**

$$ab \leq \frac{\epsilon}{2} a^2 + \frac{1}{2\epsilon} b^2, \qquad \forall \epsilon > 0,\ a, b \in \mathbb{R}. \tag{2.12}$$

- **Poincaré's inequality:**

$$\|u\|_{H^k(\mathcal{D})} \leq |u|_{H^l(\mathcal{D})}, \qquad k < l, \quad \forall u \in H_0^1(\mathcal{D}). \tag{2.13}$$

- **Jensen's inequality:**
  For any convex function g and every random variable ξ,

$$g(\mathbb{E}[\xi]) \leq \mathbb{E}[g(\xi)]. \tag{2.14}$$

2.4.4 Gronwall's Inequalities

In the analysis of time–dependent problems, Gronwall's inequalities are crucial tools and can be represented both continuous and discrete forms [91, 129]:

- **Continuous Gronwall inequality:**
  Given piecewise continuous nonnegative functions f, g, and h defined on (a, b)

and assume also that g is nondecreasing. If there exists a constant $C > 0$ independent of t such that

$$f(t) + h(t) \leq g(t) + C \int_a^t f(s)\, ds, \qquad \forall t \in (a, b),$$

then

$$f(t) + h(t) \leq e^{C(t-a)} g(t), \qquad \forall t \in (a, b).$$

- **Discrete Gronwall inequality:**
  Assume that $(a_n)_n, (b_n)_n, (c_n)_n$, and $(d_n)_n$ are sequences of nonnegative number satisfying

$$a_n + \Delta t \sum_{i=0}^{n} b_i \leq B + C\Delta t \sum_{i=0}^{n} a_i + \Delta t \sum_{i=0}^{n} c_i, \qquad \forall n \geq 0$$

  for $\Delta t,\ B,\ C > 0$. Then, if $C\Delta t < 1$,

$$a_n + \Delta t \sum_{i=0}^{n} b_i \leq e^{C(n-1)\Delta t} \left(B + \Delta t \sum_{i=0}^{n} c_i \right), \qquad \forall n \geq 0.$$

2.5 Stochastic Galerkin Method

In this section, the fundamental procedure of the stochastic Galerkin method, which is one of the well-known discretization methods for PDE with uncertainty, is discussed.

2.5.1 Karhunen–Loève Expansion

This thesis focuses on the PDE with random inputs derived from some physical characteristics connected to the models represented by the PDEs. To solve the model problem numerically, it is necessary to reduce the stochastic process into a finite number of mutually uncorrelated random variables.

The major emphasis is on a certain class of random processes in $L^2(\Omega)$ and their representations in Fourier-like expansions that are convergent with respect to the norm connected with the appropriate inner product in $L^2(\Omega)$ space. Using a Fourier-like series called the Karhunen-Loève (KL) expansion [97, 116], a stochastic process can

be represented as a linear combination of orthogonal functions. It also goes by the name proper orthogonal decomposition (POD), which divides the stochastic process' spatial and random dimensions.

Following the Karhunen–Loève (KL) expansion [97, 116], a random field $z(\boldsymbol{x}, \omega) : \mathcal{D} \times \Omega \to \mathbb{R}$ with a continuous covariance function $\mathbb{C}_z(\boldsymbol{x}, \boldsymbol{y})$ defined in (2.5) admits a proper orthogonal decomposition

$$z(\boldsymbol{x}, \omega) = \overline{z}(\boldsymbol{x}) + \kappa_z \sum_{k=1}^{\infty} \sqrt{\lambda_k} \phi_k(\boldsymbol{x}) \xi_k(\omega), \tag{2.15}$$

where $\overline{z}(\boldsymbol{x})$ is the mean of the random variable $z(x, \omega)$, κ_z is the standard deviation, and $\xi := \{\xi_1, \xi_2, \ldots\}$ are uncorrelated random variables. The pair $\{\lambda_k, \phi_k\}$ is a set of the eigenvalues and eigenfunctions of the corresponding covariance operator $\mathbb{C}_z$. In order to obtain eigenpairs $\{\lambda_k, \phi_k\}$, one needs to solve the following eigenvalue problem

$$\int_{\mathcal{D}} \mathbb{C}_z(\boldsymbol{x}, \boldsymbol{y}) \phi_i(\boldsymbol{y}) \, d\boldsymbol{y} = \lambda_i \phi_i(\boldsymbol{x}).$$

It is noted that as long as the covariance function $\mathbb{C}_z(\cdot, \cdot)$ is nonnegative definite, the eigenvalues $\{\lambda_k\}$ form a sequence of nonnegative real numbers decreasing to zero; see, e.g., [128]. Moreover, the eigenfunctions $\{\phi_k\}$ form a complete orthogonal basis in $L^2(\Omega)$.

The random variable $z(\boldsymbol{x}, \omega)$ is then approximated by truncating its KL expansion of the form

$$z(\boldsymbol{x}, \omega) \approx z_N(\boldsymbol{x}, \omega) := \overline{z}(\boldsymbol{x}) + \kappa_z \sum_{k=1}^{N} \sqrt{\lambda_k} \phi_k(\boldsymbol{x}) \xi_k(\omega). \tag{2.16}$$

Here, the choice of the truncated number N is usually based on the speed of decay on the eigenvalues since

$$\sum_{i=1}^{\infty} \lambda_i = \int_{\mathcal{D}} \mathbb{V}_z(\boldsymbol{x}) \, d\boldsymbol{x},$$

see, e.g., [65]. The truncated KL expansion (2.16) is a finite representation of the random field $z(\boldsymbol{x}, \omega)$ in the sense that the mean-square error of approximation is minimized; see, e.g., [8]. Then, the truncation error resulting from the KL–expansion is equivalent to

$$\| z - z_N \|_{L^p(L^\infty(\mathcal{D});\Omega)} \leq C \left(\sum_{i=N+1}^{\infty} \lambda_i \right)^{1/2}, \tag{2.17}$$

where the constant C is independent of the truncation number N. For example, the random field $z(\boldsymbol{x}, \omega)$ is characterized by the following exponential covariance function

$$\mathbb{C}(x, y) = e^{-|x-y|/\ell}, \tag{2.18}$$

where ℓ is correlation length and $x, y \in D = [-a, a]$. To find the truncated KL expansion, the following integral equations is needed to solve:

$$\int_{-a}^{a} e^{-|x-y|/\ell} \phi_k(y) dy = \lambda_k \phi_k(x), \quad x \in [-a, a]. \tag{2.19}$$

By differentiating twice, it is obtained that

$$\frac{d^2\phi}{dx^2} + \omega^2 \phi = 0 \quad \text{with} \quad \omega^2 := \frac{2\ell^{-1} - \ell^{-2}\lambda}{\lambda}$$

and the boundary conditions are

$$\ell^{-1}\phi(-a) - \frac{d\phi}{dx}(-a) = 0, \quad \ell^{-1}\phi(a) + \frac{d\phi}{dx}(a) = 0.$$

By solving this differential equations, the eigenfunctions and the eigenvalues are of the form

$$\phi(x) = A\cos(\omega x) + B\sin(\omega x), \quad \lambda = \frac{2\ell^{-1}}{\omega^2 + \ell^{-2}}.$$

For this special case of the covariance function, one can have the explicit expressions for its eigenfunctions ϕ_k and eigenvalues λ_k. Let $\omega_{k_{\text{odd}}}$ and $\omega_{k_{\text{even}}}$ solve the equations

$$\ell^{-1} - \omega_{k_{\text{odd}}} \tan(a\omega_{k_{\text{odd}}}) = 0,$$
$$\omega_{k_{\text{even}}} + \ell^{-1} \tan(a\omega_{k_{\text{even}}}) = 0.$$

Also, $\omega_{k_{\text{even}}}$ and $\omega_{k_{\text{odd}}}$ are positive roots of above equations, respectively. It has been shown in [117] that the even and odd indexed eigenfunctions are given, respectively, by

$$\phi_{k_{\text{odd}}}(x) = \frac{\cos(\omega_{k_{\text{odd}}} x)}{\sqrt{a + \frac{\sin(2a\omega_{k_{\text{odd}}})}{2\omega_{k_{\text{odd}}}}}}, \qquad \phi_{k_{\text{even}}}(x) = \frac{\sin(\omega_{k_{\text{even}}} x)}{\sqrt{a - \frac{\sin(2a\omega_{k_{\text{even}}})}{2\omega_{k_{\text{even}}}}}}$$

with corresponding indexed eigenvalues

$$\lambda_{k_{\text{even}}} = \frac{2\ell^{-1}}{\omega_{k_{\text{even}}}^2 + \ell^{-2}}, \qquad \lambda_{k_{\text{odd}}} = \frac{2\ell^{-1}}{\omega_{k_{\text{odd}}}^2 + \ell^{-2}}.$$

Now, consider a random field $z(x,\omega)$ characterized by its mean $\bar{z}$ and covariance function

$$\mathbb{C}(x,y) = \prod_{m=1}^{2} e^{-|x_m - y_m|/\ell_m}, \quad \text{on} \quad D = [-a_1, a_1] \times [-a_2, a_2].$$

Since $\mathbb{C}$ is separable, the eigenfunctions can be written as $\phi_k(x) = \phi_i^1(x_1)\phi_j^2(x_2)$ and the eigenvalues are $\lambda_k = \lambda_i^1 \lambda_j^2$, where the eigen–pairs $\{\lambda_i^1, \phi_i^1\}$ and $\{\lambda_j^2, \phi_j^2\}$ are solutions to the one-dimensional problem, which is defined in (2.19)

$$\int_{-a_m}^{a_m} e^{-|x-y|/\ell_m} \phi^m(y) dy = \lambda^m \phi^m(x), \quad m = 1, 2.$$

For the n–dimensional case, the eigenfunctions and the eigenvalues can be written as

$$\phi_k(x) = \prod_{m=1}^{N} \phi_{k_1}^1(x_1)\phi_{k_2}^2(x_2)...\phi_{k_m}^m(x_m), \qquad \lambda_k = \prod_{m=1}^{N} \lambda_{k_1}^1 \lambda_{k_2}^2 ... \lambda_{k_m}^m.$$

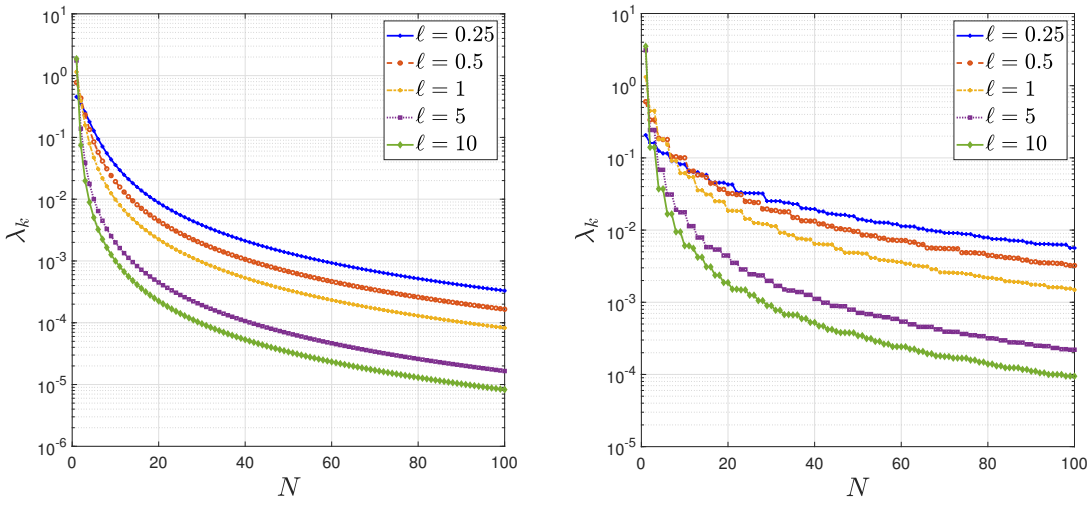

Figure 2.2: Decay of eigenvalues of the KL expansion in (2.16) for one-dimensional (left) and two-dimensional problem (right) with varying correlation length ℓ and $a_i = 1$.

For the exponential covariance in (2.18), if the correlation length ℓ is larger, the decay of eigenvalues will be faster; see Figure 2.2, and so a smaller number of terms are needed in the KL expansion (2.16) to obtain an efficient approximation. Conversely, in applications where the eigenvalues decay slowly, due to small correlation lengths, the truncation number N might be very large.

By the assumption based on finite dimensional noise and Doob–Dynkin lemma [127], the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is replaced with $(\Gamma, \mathcal{B}(\Gamma), \rho(\xi)d\xi)$, where $\mathcal{B}(\Gamma)$ denotes Borel σ–algebra and $\rho(\xi)d\xi$ is the distribution measure of the vector ξ. Then,

the random input in (2.16) is parameterized with a finite-dimensional vector $\xi : \Omega \to \Gamma \in \mathbb{R}^N$ and can be given by $z(\boldsymbol{x}, \omega) = z(\boldsymbol{x}, \xi_1(\omega), \xi_2(\omega), \ldots, \xi_N(\omega))$. Hence, we can state the tensor–product space $H^k(\mathcal{D}) \otimes L^2(\Gamma)$, which is endowed with the norm

$$\|z\|_{H^k(\mathcal{D})\otimes L^2(\Gamma)} := \left(\int_\Gamma \|z(\cdot, \xi)\|^2_{H^k(\mathcal{D})} \rho(\xi)\, d\xi \right)^{1/2} < \infty.$$

2.5.2 Generalized Polynomial Chaos Expansion

The generalized polynomial chaos (gPC) expansion [150, 152] is a way to represent a second order stochastic process $y(x, \omega) \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ with orthogonal multi–dimensional polynomials of independent random variables. The expansion converges to the actual process in the sense of the mean square as both the number of random variables and the order of the polynomials approach infinity; see, e.g., [150]. Thus, by a parameterized random vector $\xi : \Omega \to \Gamma \subset \mathbb{R}^N$, the gPC expansion of the process $y(x, \omega)$ is given by the following form

$$y(x, \omega) \approx y(x, \xi) = \sum_{i=0}^{\infty} y_i(x) \Psi_i(\xi(\omega)), \tag{2.20}$$

where y_i are the deterministic modes of the expansion defined as

$$y_i(x) = \frac{\big\langle y(x, \omega)\Psi_i(\xi) \big\rangle}{\big\langle \Psi_i^2(\xi) \big\rangle}. \tag{2.21}$$

The functions Ψ_i are multivariate orthogonal polynomials satisfying following properties:

- $\big\langle \Psi_0(\xi) \big\rangle = 1$,
- $\big\langle \Psi_i(\xi) \big\rangle = 0, \quad i > 0$,
- $\big\langle \Psi_i(\xi)\Psi_j(\xi) \big\rangle = \big\langle \Psi_i^2(\xi) \big\rangle \delta_{ij}$,

with

$$\big\langle \Psi_i(\xi) \big\rangle = \int_{\omega\in\Omega} \Psi_i(\xi(\omega))\, d\mathbb{P}(\omega) = \int_{\xi\in\Gamma} \Psi_i(\xi)\rho(\xi)\, d\xi, \tag{2.22}$$

where Γ and ρ are the support and probability density function of ξ, respectively. The probability density functions of random distributions are corresponding to the

weight functions of some particular types of orthogonal polynomials. Therefore, the orthogonal polynomials, i.e., Ψ_i, are chosen according to the type of the distribution of random input, for instance, Hermite polynomials and Gaussian random variables, Legendre polynomials and uniform random variables [100, 152]; see Askey–scheme in Table 2.1.

Table 2.1: Correspondence between polynomial basis in Askey–scheme.

Distribution	**Basis**	**Support**
Gaussian	Hermite	$(-\infty, \infty)$
Gamma	Laguerre	$[0, \infty)$
Beta	Jacobi	$[a, b]$
Uniform	Legendre	$[a, b]$

The Cameron–Martin theorem [40] states that the series (2.20) converges in the Hilbert space $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, as done in the case of KL expansion (2.16), truncating the expansion (2.20) gives

$$y(\boldsymbol{x}, \omega) \approx y_J(\boldsymbol{x}, \xi) = \sum_{i=0}^{J-1} y_i(\boldsymbol{x}) \Psi_i(\xi(\omega)), \tag{2.23}$$

where the total number of PC basis functions is determined by the dimension N of the random vector ξ and the highest order Q of the basis polynomials Ψ_i

$$J = 1 + \sum_{s=1}^{Q} \frac{1}{s!} \prod_{j=0}^{s-1} (N + j) = \frac{(N+Q)!}{N!Q!}. \tag{2.24}$$

Specifically, a partition of the support of probability density in finite dimensional space Γ consists of disjoint $\mathbf{R}^N$–boxes, $\gamma = \prod_{n=1}^{N} (r_n^\gamma, s_n^\gamma)$, with $(r_n^\gamma, s_n^\gamma) \subset \Gamma_n$ for $n = 1, \ldots, N$ so that the mesh size k_n becomes $k_n = \max_\gamma |s_n^\gamma - r_n^\gamma|$, $n = 1 \ldots N$. By following [66, 128], the corresponding stochastic space with degree at most q_n on each direction ξ_n is denoted by

$$\mathcal{S}_k^q := \text{span}\{\Psi_q(\xi) : \ q = 0, 1, \ldots, J-1\} \subset L^2(\Gamma) \tag{2.25}$$

for the multi–index $q = (q_1, \ldots, q_N)$.

Next, to exemplify the construction of the stochastic space $\mathcal{S}_k^q$ in the case of Gaussian random variables [128], let $N = 2$ and $Q = 3$ and the multi–index $q = (q_1, q_2)$ denote the degrees of the polynomials of the two random variables ξ_1 and ξ_2. Then, $\mathcal{S}_k^q$ is the

collection of two–dimensional Hermite polynomials from Table 2.1. All possible values of q becomes $(0,0)$, $(1,0)\,(0,1)$, $(2,0)$, $(1,1)$, $(0,2)$, $(3,0)$, $(2,1)$, $(1,2)$, and $(0,3)$. Considering that the univarite Hermite polynomials are $\psi_0(x) = 1$, $\psi_1(x) = x$, $\psi_2(x) = x^2 - 1$, and $\psi_3(x) = x^3 - 3x$, it can be obtained that

$$\begin{aligned}\mathcal{S}_k^q &= \text{span}\{\Psi_q(\xi) :\ q = 0, 1, \ldots, 9\} \\ &= \{1,\ \xi_1,\ \xi_1^2 - 1,\ \xi_1\xi_2,\ \xi_2^2 - 1,\ \xi_1^3 - 3\xi_1,\ (\xi_1^2 - 1)\xi_2,\ \xi_1(\xi_2^2 - 1),\ \xi_2^3 - 3\xi_2\}.\end{aligned}$$

Now, consider the case of uniform random variables with $N = 2$ and $Q = 3$. From Table 2.1, the stochastic space $\mathcal{S}_k^q$ is a set of two–dimensional Legendre polynomials. Since the univarite Legendre polynomials of degrees 0, 1, 2, 3 are $\psi_0(x) = 1$, $\psi_1(x) = x$, $\psi_2(x) = \frac{1}{2}(3x^2 - 1)$, and $\psi_3(x) = \frac{1}{2}(5x^3 - 3x)$, the corresponding space is defined as

$$\begin{aligned}\mathcal{S}_k^q =&\ \text{span}\{\Psi_q(\xi) :\ q = 0, 1, \ldots, 9\} \\ =&\ \{1,\ \xi_1,\ \frac{1}{2}(3\xi_1^2 - 1),\ \xi_1\xi_2,\ \frac{1}{2}(3\xi_2^2 - 1),\ \frac{1}{2}(5\xi_1^3 - 3\xi_1),\ \frac{1}{2}(3\xi_1^2 - 1)\xi_2, \\ &\ \frac{1}{2}\xi_1(3\xi_2^2 - 1),\ \frac{1}{2}(5\xi_2^3 - 3\xi_2)\}.\end{aligned}$$

CHAPTER 3

CONVECTION DIFFUSION EQUATIONS WITH RANDOM COEFFICIENTS

In this chapter, the main focus is the numerical investigation of a convection diffusion equation with random coefficients by using the stochastic Galerkin approach. Corresponding PDE can be considered as a basic model for transport phenomena in a random media. Single phase flow model in porous media with uncertain coefficients is a fundamental model, and it is widely used to describe how uncertain factors, such as permeability, impact the pressure of the fluid [46, 114]. Due to the lack of knowledge about permeability, the deterministic PDEs do not completely represent the actual behaviour of such physical phenomena. Therefore, it is reasonable to model the permeability parameter as a random field, which corresponds to the solution of a convection diffusion equation; see, e.g., [73, 149].

The rest of the chapter is organized as follows: In the next section, the stationary model problem, that is, a convection diffusion equation with random coefficients, is introduced, and an overview of its discretization, obtained by Karhunen–Loève (KL) expansion, stochastic Galerkin method, and discontinuous Galerkin method, is provided. In Section 3.2, a priori error estimates for the stationary problem is derived in the energy norm. Section 3.3 discusses the implementation of low–rank iterative solvers. In Section 3.4, we extend our findings into unsteady convection diffusion with random coefficients and provide some error estimates for the stability and convergence. Numerical results are given in Section 3.5 to show the efficiency of the proposed approaches. Finally, some conclusions and discussions are given in Section 3.6 based on the findings in this chapter.

3.1 Stationary Model Problem with Random Coefficients

This section presents the model problem, which is a stationary convection diffusion equation with random coefficients: find a random function $y : \overline{\mathcal{D}} \times \Omega \to \mathbb{R}$ such that $\mathbb{P}$-almost surely in Ω

$$-\nabla \cdot (a(\boldsymbol{x},\omega)\nabla y(\boldsymbol{x},\omega)) + \mathbf{b}(\boldsymbol{x},\omega)\cdot\nabla y(\boldsymbol{x},\omega) = f(\boldsymbol{x}) \qquad \text{in } \mathcal{D}\times\Omega, \tag{3.1a}$$

$$y(\boldsymbol{x},\omega) = y_{DB}(\boldsymbol{x}) \quad \text{on } \partial\mathcal{D}\times\Omega, \tag{3.1b}$$

where $a : (\mathcal{D}\times\Omega) \to \mathbb{R}$ and $\mathbf{b} : (\mathcal{D}\times\Omega) \to \mathbb{R}^2$ are random diffusivity and velocity coefficients, respectively, which are assumed to have continuous and bounded covariance functions. The functions $f(\boldsymbol{x}) \in L^2(\mathcal{D})$ and $y_{DB}(\boldsymbol{x}) \in H^{1/2}(\partial\mathcal{D})$ correspond to the deterministic source term and Dirichlet boundary condition, respectively. To be ensured the regularity of the solution y, the following assumptions are needed:

i) The diffusivity coefficient $a(\boldsymbol{x},\omega)$ is $\mathbb{P}$–almost surely uniformly positive, that is, there exist constants $a_{\min}, a_{\max}$ such that $0 < a_{\min} \le a_{\max} < \infty$, with

$$a_{\min} \le a(\boldsymbol{x},\omega) \le a_{\max} \qquad \text{a. e. in } \mathcal{D}\times\Omega. \tag{3.2}$$

In addition, $a(\boldsymbol{x},\omega)$ has a uniformly bounded and continuous first derivative.

ii) The velocity coefficient $\mathbf{b}$ satisfies $\mathbf{b} \in \left(L^\infty(\overline{\mathcal{D}})\right)^2$ for a.e. $\omega \in \Omega$ and is incompressible, i.e., $\nabla\cdot\mathbf{b}(\boldsymbol{x},\omega) = 0$.

Under the assumptions on the coefficients provided above, the well–posedness of the model equation (3.1) follows from the classical Lax–Milgram lemma; see, e.g., [11, 117]. It is noted that the truncated KL expansion of the diffusivity coefficient $a(\boldsymbol{x},\omega)$ given in (2.16) should satisfy the positivity condition (3.2) to ensure the well-posedness of the problem (3.1). Throughout this chapter, it is crucial to assume a stronger dominance of the mean of the random input $a(\boldsymbol{x},\omega)$ as discussed in [103, Section 2.3] and [128, Theorem 3.8], i.e.,

$$\overline{a}(\boldsymbol{x}) > \kappa_a \sum_{k=1}^{N} \sqrt{\lambda_k}\phi_k(\boldsymbol{x})\xi_k(\omega).$$

The solution of the model problem (3.1), $y(\boldsymbol{x},\omega) \in L^2(\Omega,\mathcal{F},\mathbb{P})$, is represented by a generalized polynomial chaos (PC) approximation as discussed in Section 2.5.2

$$y(\boldsymbol{x},\omega) = \sum_{i=0}^{J-1} y_i(\boldsymbol{x})\Psi_i(\xi(\omega)), \tag{3.3}$$

where $y_i(\boldsymbol{x})$ are the deterministic modes of the expansion and Ψ_i are multivariate orthogonal polynomials.

Following, if inserting KL expansions (2.16) of the diffusion coefficient $a(\boldsymbol{x},\omega)$ and the convection coefficient $\mathbf{b}(\boldsymbol{x},\omega)$, and the solution expression (3.3) into (3.1), one can obtain

$$\begin{aligned} - \quad \sum_{i=0}^{J-1} \nabla\cdot\left(\left(\overline{a}(\boldsymbol{x}) + \kappa_a \sum_{k=1}^{N} \sqrt{\lambda_k^a}\phi_k^a(\boldsymbol{x})\xi_k\right)\nabla y_i(\boldsymbol{x})\Psi_i\right) \\ + \sum_{i=0}^{J-1}\left(\overline{\mathbf{b}}(\boldsymbol{x}) + \kappa_{\mathbf{b}} \sum_{k=1}^{N} \sqrt{\lambda_k^b}\phi_k^b(\boldsymbol{x})\xi_k\right)\cdot\nabla y_i(\boldsymbol{x})\Psi_i = f(\boldsymbol{x}). \end{aligned} \tag{3.4}$$

After projecting (3.4) onto the space spanned by the PC basis functions, it is obtained the following linear system, consisting of J deterministic convection diffusion equations for $j = 0, ..., J-1$

$$-\sum_{i=0}^{J-1}\left(\nabla\cdot(a_{ij}\nabla y_i(\boldsymbol{x})) + \mathbf{b}_{ij}\cdot\nabla y_i(\boldsymbol{x})\right) = \langle\Psi_j\rangle\, f(\boldsymbol{x}), \tag{3.5}$$

where

$$\begin{aligned} a_{ij} &= \overline{a}(\boldsymbol{x})\left\langle\Psi_i^2(\xi)\right\rangle\delta_{ij} + \kappa_a\sum_{k=1}^{N}\sqrt{\lambda_k^a}\phi_k^a(\boldsymbol{x})\left\langle\xi_k\Psi_i(\xi)\Psi_j(\xi)\right\rangle, \\ \mathbf{b}_{ij} &= \overline{\mathbf{b}}(\boldsymbol{x})\left\langle\Psi_i^2(\xi)\right\rangle\delta_{ij} + \kappa_{\mathbf{b}}\sum_{k=1}^{N}\sqrt{\lambda_k^{\mathbf{b}}}\phi_k^{\mathbf{b}}(\boldsymbol{x})\left\langle\xi_k\Psi_i(\xi)\Psi_j(\xi)\right\rangle, \end{aligned}$$

where the inner product $\langle\cdot\rangle$ is defined in (2.22). Instead of the solution $y(\boldsymbol{x},\omega)$, the quantity of interest is the statistical moments of the solution $y(\boldsymbol{x},\omega)$ in (3.1). It is simple to determine the statistical moments and the probability density of the solution after the modes y_i, $i = 0, 1, \ldots, J-1$, have been determined. For instance, the mean and the variance of the solution are given, respectively, by

$$\begin{aligned} \left\langle y(\boldsymbol{x},\xi)\right\rangle &= \left\langle\sum_{i=0}^{J-1} y_i(\boldsymbol{x})\Psi_i(\xi)\right\rangle \\ &= \sum_{i=0}^{J-1} y_i(\boldsymbol{x})\left\langle\Psi_i(\xi)\right\rangle = \sum_{i=0}^{J-1} y_i(\boldsymbol{x})\delta_{0,i} = y_0(\boldsymbol{x}) \end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(y(\boldsymbol{x},\xi)) &= \big\langle y(\boldsymbol{x},\xi)^2\big\rangle - \big\langle y(\boldsymbol{x},\xi)\big\rangle^2 \\
&= \left\langle \sum_{i=0}^{J-1}\sum_{j=0}^{J-1} y_i(\boldsymbol{x})y_j(\boldsymbol{x})\Psi_i(\xi)\Psi_j(\xi) \right\rangle - y_0^2(\boldsymbol{x}) \\
&= \sum_{i=0}^{J-1}\sum_{j=0}^{J-1} y_i(\boldsymbol{x})y_j(\boldsymbol{x})\big\langle \Psi_i(\xi)\Psi_j(\xi)\big\rangle - y_0^2(\boldsymbol{x}) \\
&= \sum_{i=0}^{J-1}\sum_{j=0}^{J-1} y_i(\boldsymbol{x})y_j(\boldsymbol{x})\big\langle \Psi_i^2(\xi)\delta_{ij}\big\rangle - y_0^2(\boldsymbol{x}) \\
&= \sum_{i=0}^{J-1} y_i^2(\boldsymbol{x})\big\langle \Psi_i^2(\xi)\big\rangle - y_0^2(\boldsymbol{x}) = \sum_{i=1}^{J-1} y_i^2(\boldsymbol{x})\big\langle \Psi_i^2(\xi)\big\rangle.
\end{aligned}$$

3.1.1 Symmetric Interior Penalty Galerkin Method

When the stochastic domain has been discretized, the parameterized stochastic PDEs may now be approximated using classical numerical techniques like finite element (FEM), finite volume (FVM), or finite difference (FDM) approaches. Since it is more effective than the finite difference approach and has a higher rate of high-order convergence, the standard continuous finite element is the technique that is utilized the most frequently in order to solve PDEs [11, 12, 99, 115]. Yet, while solving convection-dominated PDEs, the primary objective of this thesis, the discrete solutions obtained from finite element simulations display spurious oscillations. Consequently, for spatial discretization in this thesis, it is opted to employ discontinuous Galerkin techniques [7, 54, 129, 136]. In comparison to the discontinuous Galerkin approach, the finite volume method, ordinary continuous finite element method, and finite difference method are all unable to handle complicated geometries or to achieve high-order precision. Moreover, especially for convection-dominated problems, DG approaches generate stable discretization without the requirement for stabilization techniques, such as streamline upwind Petrov–Galerkin (SUPG) [37], edge stabilization Galerkin method [38], and they make it relatively simple to apply different degrees of approximation to different elements in the mesh; see Table 3.1 for a detailed discussion [90]. Several types of discontinuous Galerkin schemes, such as nonsymmetric interior penalty Galerkin (NIPG) [130], symmetric interior penalty Galerkin (SIPG)

[6, 147], incomplete interior penalty Galerkin (IIPG) [52], and local discontinuous Galerkin (LDG) [49], have been introduced in the literature. In this thesis, due to the symmetricity and adjoint consistency, SIPG method is chosen as DG method to discretize the spatial domain [6, 147]. Now, the SIPG discretization is briefly recalled following studies in [41, 42, 157].

Table 3.1: Comparison of the well-known numerical approaches to discretize the spatial domain.

	Complex geometries	Higher-order accuracy and hp-adaptivity	Local mass Conservation
FDM	$\times$	✓	✓
FVM	✓	$\times$	✓
FEM	✓	✓	$\times$
DG	✓	✓	✓

A shape-regular simplicial triangulations of $\mathcal{D}$ stands for $\{\mathcal{T}_h\}_h$ providing $\overline{\mathcal{D}} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$ for each mesh $\mathcal{T}_h$. It is assumed for the regularity of the mesh that the intersection $K_i \cap K_j$ is either empty or a vertex or an edge, i.e., there are no hanging nodes, for different triangles $K_i, K_j \in \mathcal{T}_h$, $i \neq j$. The diameter of an element K and the length of an edge E are denoted by h_K and h_E, respectively. In addition, the maximum value of the element diameter is denoted by $h = \max_{K\in\mathcal{T}_h} h_K$.

The set of all edges $\mathcal{E}_h$ is divided into the interior edges $\mathcal{E}_h^0$ and the boundary edges $\mathcal{E}_h^\partial$ such that $\mathcal{E}_h = \mathcal{E}_h^0 \cup \mathcal{E}_h^\partial$. For a fixed realization ω and the unit outward normal $\mathbf{n}$ to $\partial\mathcal{D}$, the inflow and outflow parts of $\partial\mathcal{D}$ are denoted by $\partial\mathcal{D}^-$ and $\partial\mathcal{D}^+$, respectively,

$$\begin{aligned}\partial\mathcal{D}^- &= \{\boldsymbol{x} \in \partial\mathcal{D} \,:\, \mathbf{b}(\boldsymbol{x},\omega)\cdot\mathbf{n}(\boldsymbol{x}) < 0\}\,,\\ \partial\mathcal{D}^+ &= \{\boldsymbol{x} \in \partial\mathcal{D} \,:\, \mathbf{b}(\boldsymbol{x},\omega)\cdot\mathbf{n}(\boldsymbol{x}) \geq 0\}\,.\end{aligned}$$

Also, denoting $\mathbf{n}_K$ the unit normal vector on the boundary ∂K of an element K, the inflow and outflow boundaries of an element K are defined by

$$\begin{aligned}\partial K^- &= \{\boldsymbol{x} \in \partial K \,:\mathbf{b}(\boldsymbol{x},\omega)\cdot\mathbf{n}_K(\boldsymbol{x}) < 0\}\,,\\ \partial K^+ &= \{\boldsymbol{x} \in \partial K \,:\mathbf{b}(\boldsymbol{x},\omega)\cdot\mathbf{n}_K(\boldsymbol{x}) \geq 0\}\,,\end{aligned}$$

respectively. Let the edge E be a common edge for two elements K and K^e. For a piecewise continuous scalar function y, there are two traces of y along E, denoted by $y|_E$ from inside K and $y^e|_E$ from inside K^e. The jump and average of y across the

edge E are defined by:

$$[\![y]\!] = y|_E \mathbf{n}_K + y^e|_E \mathbf{n}_{K^e}, \quad \{\!\{y\}\!\} = \frac{1}{2}\big(y|_E + y^e|_E\big). \tag{3.6}$$

Similarly, for a piecewise continuous vector field ∇y, the jump and average across an edge E are given by

$$[\![\nabla y]\!] = \nabla y|_E \cdot \mathbf{n}_K + \nabla y^e|_E \cdot \mathbf{n}_{K^e}, \quad \{\!\{\nabla y\}\!\} = \frac{1}{2}\big(\nabla y|_E + \nabla y^e|_E\big). \tag{3.7}$$

For a boundary edge $E \in K \cap \partial\mathcal{D}$, the operators are defined by $\{\!\{\nabla y\}\!\} = \nabla y$ and $[\![y]\!] = y\mathbf{n}$, where $\mathbf{n}$ is the outward normal unit vector on $\partial\mathcal{D}$.

Then, the discrete state and test spaces are defined as follows

$$V_h = \left\{ y \in L^2(\mathcal{D}) \,:\, y\,|_K \in \mathbb{P}^\ell(K) \quad \forall K \in \mathcal{T}_h \right\}, \tag{3.8}$$

where $\mathbb{P}^\ell(K)$ be the set of all polynomials on K of degree at most ℓ for an integer ℓ and $K \in \mathcal{T}_h$. It should be noted that the space of discrete state and test functions are identical since it is possible to impose boundary conditions weakly in the discontinuous Galerkin discretization.

By following the standard discontinuous Galerkin structure discussed in [7, 129], the (bi)–linear forms of the SIPG discretization for a finite dimensional vector ξ can be expressed as indicated below:

$$\begin{aligned} a_h(y,v,\xi) \;=\; & \sum_{K\in\mathcal{T}_h} \int_K a(.,\xi)\nabla y \cdot \nabla v \, dx - \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \int_E \{\!\{a(.,\xi)\nabla y\}\!\}[\![v]\!]\, ds \\ & - \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \int_E \{\!\{a(.,\xi)\nabla v\}\!\}[\![y]\!]\, ds + \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E [\![y]\!]\cdot[\![v]\!]\, ds \\ & + \sum_{K\in\mathcal{T}_h} \int_K \mathbf{b}(.,\xi)\cdot\nabla y v \, dx + \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\mathcal{D}} \mathbf{b}(.,\xi)\cdot\mathbf{n}_E (y^e - y) v \, ds \\ & - \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\cap\partial\mathcal{D}^-} \mathbf{b}(.,\xi)\cdot\mathbf{n}_E y v \, ds \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} l_h(v,\xi) \;=\; & \sum_{K\in\mathcal{T}_h} \int_K f v \, dx + \sum_{E\in\mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E y_{DB}[\![v]\!]\, ds - \sum_{E\in\mathcal{E}_h^\partial} \int_E y_{DB}\{\!\{a(.,\xi)\nabla v\}\!\}\, ds \\ & - \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\cap\partial\mathcal{D}^-} \mathbf{b}(.,\xi)\cdot\mathbf{n}_E y_{DB} v \, ds, \end{aligned} \tag{3.10}$$

where the parameter $\sigma \in \mathbb{R}_0^+$, called as the penalty parameter, should be sufficiently large to ensure the stability of the SIPG scheme; independent of the mesh size h. However, as discussed in [129, Section 2.7.1], it depends on the degree of polynomials used in the DG discretization and whether the edge E is the interior or boundary edge.

Then, (bi)–linear forms of the stochastic discontinuous Galerkin (SDG) correspond to

$$a_\xi(y,v) = \int_\Gamma a_h(y,v,\xi)\rho(\xi)\,d\xi, \quad l_\xi(v) = \int_\Gamma l_h(v,\xi)\rho(\xi)\,d\xi. \tag{3.11}$$

Now, the associated energy norm on $\mathcal{D}\times\Gamma$ is defined as

$$\|y\|_\xi = \left(\int_\Gamma \|y(.,\xi)\|_e^2 \rho(\xi)\,d\xi\right)^{\frac{1}{2}}, \tag{3.12}$$

where $\|y(.,\xi)\|_e$ is the energy norm on $\mathcal{D}$, given as

$$\begin{aligned}\|y(.,\xi)\|_e \;=\; &\left(\sum_{K\in\mathcal{T}_h}\int_K a(.,\xi)(\nabla y)^2\,dx + \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{\sigma}{h_E}\int_E [\![y]\!]^2\,ds\right.\\ &\left.+\frac{1}{2}\sum_{E\in\mathcal{E}_h^\partial}\int_E \mathbf{b}(.,\xi)\cdot\mathbf{n}_E y^2 ds + \frac{1}{2}\sum_{E\in\mathcal{E}_h^0}\int_E \mathbf{b}(.,\xi)\cdot\mathbf{n}_E(y^e-y)^2\,ds\right)^{\frac{1}{2}}.\end{aligned}$$

Following standard arguments as done in the deterministic case, one can easily show the coercivity and continuity of $a_\xi(\cdot,\cdot)$.

Lemma 3.1.1. *For $y, v \in V_h \otimes \mathcal{S}_k^q$, it holds that*

$$a_\xi(y,y) \;\geq\; c_{cv}\,\|y\|_\xi^2, \tag{3.13a}$$

$$a_\xi(y,v) \;\leq\; c_{ct}\,\|y\|_\xi\|v\|_\xi, \tag{3.13b}$$

where the coercivity constant c_{cv} depends on $a_{\min}$, whereas the continuity constant c_{ct} depends on $a_{\max}$.

Proof. Following the works [11, 12], and by the definition of the energy norm in (3.12) and the coercivity and continuity of $a_h(y,y,\xi)$, the bounds (3.13a) and (3.13b) are obtained. □

Thus, the SDG variational formulation of (3.1) is as follows: Find $y \in V_h \otimes \mathcal{S}_k^q$ such that

$$a_\xi(y,v) = l_\xi(v), \qquad \forall v \in V_h \otimes \mathcal{S}_k^q. \tag{3.14}$$

3.1.2 Linear System

After applying of the KL expansion (2.16), the gPC expansion (2.23), and SIPG scheme (3.9)-(3.10), one can get the following linear system:

$$\underbrace{\left(\sum_{i=0}^{N} \mathcal{G}_i \otimes \mathcal{K}_i\right)}_{\mathcal{A}} \mathbf{y} = \underbrace{\left(\sum_{i=0}^{N} \mathbf{g}_i \otimes \mathbf{f}_i\right)}_{\mathcal{F}}, \tag{3.15}$$

where $\mathbf{y} = (y_0, \ldots, y_{J-1})^T$ with $y_i \in \mathbb{R}^{N_d}$, $i = 0, 1, \ldots, J-1$ and N_d corresponds to the degree of freedoms for the spatial discretization. The stiffness matrices $\mathcal{K}_i \in \mathbb{R}^{N_d \times N_d}$ and the right–hand side vectors $\mathbf{f}_i \in \mathbb{R}^{N_d}$ in (3.15) are given, respectively, by

$$\begin{aligned}
\mathcal{K}_0(r,s) = & \sum_{K\in\mathcal{T}_h} \int_K \left(\overline{a}\,\nabla\varphi_r\cdot\nabla\varphi_s + \overline{\mathbf{b}}\cdot\nabla\varphi_r\varphi_s\right) d\boldsymbol{x} \\
& - \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \int_E \left(\{\!\{\overline{a}\,\nabla\varphi_r\}\!\}[\![\varphi_s]\!] + \{\!\{\overline{a}\,\nabla\varphi_s\}\!\}[\![\varphi_r]\!]\right) ds \\
& + \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E [\![\varphi_r]\!]\cdot[\![\varphi_s]\!]\, ds + \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\mathcal{D}} \overline{\mathbf{b}}\cdot\mathbf{n}_E(\varphi_r^e-\varphi_r)\varphi_s\, ds \\
& - \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\cap\partial\mathcal{D}^-} \overline{\mathbf{b}}\cdot\mathbf{n}_E\varphi_r\varphi_s\, ds,
\end{aligned} \tag{3.16}$$

$$\begin{aligned}
\mathcal{K}_i(r,s) = & \sum_{K\in\mathcal{T}_h} \int_K \left(\left(\kappa_a\sqrt{\lambda_i^a}\phi_i^a\right)\nabla\varphi_r\cdot\nabla\varphi_s + \left(\kappa_{\mathbf{b}}\sqrt{\lambda_i^{\mathbf{b}}}\phi_i^{\mathbf{b}}\right)\cdot\nabla\varphi_r\varphi_s\right) d\boldsymbol{x} \\
& - \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \int_E \left(\left\{\!\!\left\{\left(\kappa_a\sqrt{\lambda_i^a}\phi_i^a\right)\nabla\varphi_r\right\}\!\!\right\}[\![\varphi_s]\!] + \left\{\!\!\left\{\left(\kappa_a\sqrt{\lambda_i^a}\phi_i^a\right)\nabla\varphi_s\right\}\!\!\right\}[\![\varphi_r]\!]\right) ds \\
& + \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E [\![\varphi_r]\!]\cdot[\![\varphi_s]\!]\, ds \\
& + \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\mathcal{D}} \left(\kappa_{\mathbf{b}}\sqrt{\lambda_i^{\mathbf{b}}}\phi_i^{\mathbf{b}}\right)\cdot\mathbf{n}_E(\varphi_r^e-\varphi_r)\varphi_s\, ds \\
& - \sum_{T\in\mathcal{T}_h} \int_{\partial K^-\cap\partial\mathcal{D}^-} \left(\kappa_{\mathbf{b}}\sqrt{\lambda_i^{\mathbf{b}}}\phi_i^{\mathbf{b}}\right)\cdot\mathbf{n}_E\varphi_r\varphi_s\, ds,
\end{aligned} \tag{3.17}$$

$$\begin{aligned}
f_0(s) = & \sum_{K\in\mathcal{T}_h} \int_K f\varphi_s\, d\boldsymbol{x} + \sum_{E\in\mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E y_{DB}[\![\varphi_s]\!]\, ds - \sum_{E\in\mathcal{E}_h^\partial} \int_E y_{DB}\{\!\{\overline{a}\nabla\varphi_s\}\!\}\, ds \\
& - \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\cap\partial\mathcal{D}^-} \overline{\mathbf{b}}\cdot\mathbf{n}_E y_{DB}\varphi_s\, ds,
\end{aligned} \tag{3.18}$$

$$f_i(s) = \sum_{E\in\mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E y_{DB}[\![\varphi_s]\!]\, ds - \sum_{E\in\mathcal{E}_h^\partial} \int_E y_{DB}\left\{\!\!\left\{ \left(\kappa_a \sqrt{\lambda_i^a}\phi_i^a\right)\nabla\varphi_s \right\}\!\!\right\} ds$$
$$- \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\cap\partial\mathcal{D}^-} \left(\kappa_{\mathbf{b}} \sqrt{\lambda_i^{\mathbf{b}}}\phi_i^{\mathbf{b}}\right)\cdot \mathbf{n}_E y_{DB}\varphi_s\, ds, \tag{3.19}$$

where $\{\varphi_i(\boldsymbol{x})\}$ is the set of basis functions for the spatial discretization, i.e., $V_h = \text{span}\{\varphi_i(\boldsymbol{x})\}$.

For $i = 0, \ldots, N$, the stochastic matrices $\mathcal{G}_i \in \mathbb{R}^{J\times J}$ in (3.15) are given by

$$\mathcal{G}_0(r,s) = \langle \Psi_r \Psi_s \rangle\,, \qquad \mathcal{G}_i(r,s) = \langle \xi_i \Psi_r \Psi_s \rangle\,, \tag{3.20}$$

whereas the stochastic vectors $\mathbf{g}_i \in \mathbb{R}^J$ in (3.15) are defined as

$$\mathbf{g}_0(r) = \langle \Psi_r \rangle\,, \qquad \mathbf{g}_i(r) = \langle \xi_i \Psi_r \rangle\,. \tag{3.21}$$

In (3.20), each stochastic basis function $\Psi_q(\xi)$ is corresponding to a product of N univariate orthogonal polynomials $\psi_{q_n(\xi_n)}$, i.e.,

$$\Psi_q(\xi) = \prod_{n=1}^{N} \psi_{q_n}(\xi_n),$$

where $\xi = \{\xi_1, \ldots, \xi_N\}$ are the uncorrelated random variables and the multi–index q is defined by $q = (q_1, q_2, \ldots, q_N)$ with $\sum_{n=1}^{N} q_n \le Q$. In this chapter, Legendre polynomials are chosen as stochastic basis functions because the underlying random variables are considered as having a uniform distribution.

Now, suppose Legendre polynomials in uniform random variables on $(-\sqrt{3}, \sqrt{3})$ employed. Then, recalling the following three–term recurrence for the Legendre polynomials [117]

$$\psi_{k+1}(\boldsymbol{x}) = \frac{\sqrt{2k+1}\sqrt{2k+3}}{(k+1)\sqrt{3}}\boldsymbol{x}\psi_k(\boldsymbol{x}) - \frac{k\sqrt{2k+3}}{(k+1)\sqrt{2k-1}}\psi_{k-1}$$

with $\psi_0 = 1,\ \psi_{-1} = 0$, we obtain

$$\begin{aligned}\mathcal{G}_0(i,j) &= \int_\Gamma \Psi_i(\vec{\xi})\Psi_j(\vec{\xi})\rho(\vec{\xi})\, d\vec{\xi} \\ &= \prod_{s=1}^{N}\left(\int_{\Gamma_s} \psi_{i_s}(\xi_s)\psi_{j_s}(\xi_s)\rho(\xi_s)\, d\xi_s\right)\end{aligned}$$

$$= \prod_{s=1}^{N} \langle \psi_{i_s}(\xi_s)\psi_{j_s}(\xi_s)\rangle = \prod_{s=1}^{N} \langle \psi_{i_s}^2(\xi_s)\rangle \delta_{i_s j_s}$$
$$= \prod_{s=1}^{N} \delta_{i_s j_s} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

and for $k = 1 : N$

$$\begin{aligned}
\mathcal{G}_k(i,j) &= \int_{\Gamma} \xi_k \Psi_i(\vec{\xi})\Psi_j(\vec{\xi})\rho(\vec{\xi})\, d\vec{\xi} \\
&= \int_{-\sqrt{3}}^{\sqrt{3}} \cdots \int_{-\sqrt{3}}^{\sqrt{3}} \xi_k \Psi_i(\vec{\xi})\Psi_j(\vec{\xi})\rho(\vec{\xi})\, d\vec{\xi} \\
&= \left(\prod_{s=1, s\neq k}^{N} \langle \psi_{i_s}(\xi_s)\psi_{j_s}(\xi_s)\rangle \right) \langle \xi_k \psi_{i_k}(\xi_k)\psi_{j_k}(\xi_k)\rangle \\
&= \left(\prod_{s=1, s\neq k}^{N} \langle \psi_{i_s}(\xi_s)\psi_{j_s}(\xi_s)\rangle \right) \\
&\quad \times \left(\frac{(i_k+1)\sqrt{3}}{\sqrt{(2i_k+1)(2i_k+3)}} \langle \psi_{i_k+1}\psi_{j_k}\rangle + \frac{i_k\sqrt{3}}{\sqrt{(2i_k+1)(2i_k-1)}} \langle \psi_{i_k-1}\psi_{j_k}\rangle \right) \\
&= \begin{cases} \left(\prod_{s=1, s\neq k}^{N} \delta_{i_s j_s} \right) \dfrac{(i_k+1)\sqrt{3}}{\sqrt{(2i_k+1)(2i_k+3)}}, & \text{if } i_k + 1 = j_k, \\ \left(\prod_{s=1, s\neq k}^{N} \delta_{i_s j_s} \right) \dfrac{i_k\sqrt{3}}{\sqrt{(2i_k+1)(2i_k-1)}}, & \text{if } i_k - 1 = j_k, \\ 0, & \text{otherwise,} \end{cases} \\
&= \begin{cases} \dfrac{(i_k+1)\sqrt{3}}{\sqrt{(2i_k+1)(2i_k+3)}}, & \text{if } i_k + 1 = j_k \text{ and } i_s = j_s,\ s = \{1:N\} \setminus \{k\}, \\ \dfrac{i_k\sqrt{3}}{\sqrt{(2i_k+1)(2i_k-1)}}, & \text{if } i_k - 1 = j_k \text{ and } i_s = j_s,\ s = \{1:N\} \setminus \{k\}, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

Hence, $\mathcal{G}_0$ is an identity matrix, whereas $\mathcal{G}_k,\ k > 0$, contains at most two nonzero entries per row; see, e.g., [66, 128]. On the other hand, $\mathbf{g}_i$ is the first column of $\mathcal{G}_i,\ i = 0, 1, \ldots, N$.

3.2 Error Estimates

In this section, a priori error estimates for stationary convection diffusion equations with random coefficients (3.1), discretized by stochastic discontinuous Galerkin method is presented.

For $v \in H^{q+1}(\Gamma), \Phi \in \mathcal{S}_k^q$, the following estimate is obtained in [11, Section 3.2]

$$\min_{\Phi \in \mathcal{S}_k^q} \|v - \Phi\|_{L^2(\Gamma)} \leq \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{\xi_n}^{q_n+1} v\|_{L^2(\Gamma)}}{(q_n+1)!}, \tag{3.22}$$

where q_n is the degree of polynomial space for the spatial domain and the degree of (discontinuous) finite element approximation space $\mathcal{S}_k^q$ on each direction ξ_n. Also, $k_n = \max_{\gamma} |s_n^\gamma - r_n^\gamma|$ is the mesh size for a partition of the support of probability density in finite dimensional space Γ consists of disjoint $\mathbf{R}^N$–boxes, $\gamma = \prod_{n=1}^{N} (r_n^\gamma, s_n^\gamma)$, with $(r_n^\gamma, s_n^\gamma) \subset \Gamma_n$ for $n = 1, \ldots, N$.

To later use, the L^2–projection operator $\Pi_q : L^2(\Gamma) \to \mathcal{S}_k^q$ is introduced by

$$(\Pi_q(\xi) - \xi, \zeta)_{L^2(\Gamma)} = 0 \qquad \forall \zeta \in \mathcal{S}_k^q, \qquad \forall \xi \in L^2(\Gamma), \tag{3.23}$$

and the H^1–projection operator $\mathcal{R}_h : H^1(\mathcal{D}) \to V_h \cap H^1(\mathcal{D})$ satisfies

$$(\mathcal{R}_h(\nu) - \nu, \chi)_{L^2(\mathcal{D})} = 0 \qquad \forall \chi \in V_h, \qquad \forall \nu \in H^1(\mathcal{D}), \tag{3.24a}$$

$$(\nabla(\mathcal{R}_h(\nu) - \nu), \nabla\chi)_{L^2(\mathcal{D})} = 0 \qquad \forall \chi \in V_h, \qquad \forall \nu \in H^1(\mathcal{D}). \tag{3.24b}$$

Lastly, discontinuous Galerkin approximation estimates are given for all $v \in H^2(K)$ with $K \in \mathcal{T}_h$ in the following.

Theorem 3.2.1. *([129, Theorem 2.6]) Assume that $v \in H^2(K)$ for $K \in \mathcal{T}_h$ and $\widetilde{v} \in \mathbb{P}^\ell$. Then, there exists a constant C independent of v and h such that*

$$\|v - \widetilde{v}\|_{H^q(K)} \leq C\, h^{\min(\ell+1,2)-q} |v|_{H^2(K)} \qquad 0 \leq q \leq 2. \tag{3.25}$$

Let $\widetilde{y} \in V_h \otimes \mathcal{S}_k^q$ be an approximation of the solution y. Following [11, 115], we derive an approximation for the tensor product $V_h \otimes \mathcal{S}_k^q$, which is a direct application of the results for V_h and $\mathcal{S}_k^q$.

Theorem 3.2.2. *Assume that* $v \in L^2(H^2(\mathcal{D});\Gamma) \cap H^{q+1}(H^1(\mathcal{D});\Gamma)$ *and* $\widetilde{v} \in V_h \otimes \mathcal{S}_k^q$. *Then, there exist the following bounds*

$$\|\nabla(v-\widetilde{v})\|_{L^2(L^2(\mathcal{D});\Gamma)} \leq Ch^{\min(\ell+1,2)-1}\|v\|_{L^2(H^2(\mathcal{D});\Gamma)} + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}v\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}, \tag{3.26a}$$

$$\|\nabla^2(v-\widetilde{v})\|_{L^2(L^2(\mathcal{D});\Gamma)} \leq Ch^{\min(\ell+1,2)-2}\|v\|_{L^2(H^2(\mathcal{D});\Gamma)} + Ch^{-1}\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}v\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}, \tag{3.26b}$$

where the constant C *independent of* v, *the mesh sizes* h, *and* k_n.

Proof. By choosing $\widetilde{v} = \Pi_q(\mathcal{R}_h(v))$, it is obtained that

$$\begin{aligned}\|v-\widetilde{v}\|_{L^2(H^1(\mathcal{D});\Gamma)} &= \|v-\Pi_q(\mathcal{R}_h(v))\|_{L^2(H^1(\mathcal{D});\Gamma)}\\ &\leq \|v-\mathcal{R}_h(v)\|_{L^2(H^1(\mathcal{D});\Gamma)} + \|\mathcal{R}_h(v)-\Pi_q(\mathcal{R}_h(v))\|_{L^2(H^1(\mathcal{D});\Gamma)}\\ &= \|v-\mathcal{R}_h(v)\|_{L^2(H^1(\mathcal{D});\Gamma))} + \|\mathcal{R}_h(v)-\mathcal{R}_h(\Pi_q(v))\|_{L^2(H^1(\mathcal{D});\Gamma)}\\ &= \|v-\mathcal{R}_h(v)\|_{L^2(H^1(\mathcal{D});\Gamma)} + \|\mathcal{R}_h(v-\Pi_q(v))\|_{L^2(H^1(\mathcal{D});\Gamma)}\end{aligned} \tag{3.27}$$

for a fixed $v \in L^2(H^2(\mathcal{D});\Gamma) \cap H^{q+1}(H^1(\mathcal{D});\Gamma)$. In the light of the estimate in (3.25), one can easily obtain the following estimate

$$\|v-\mathcal{R}_h(v)\|_{L^2(H^1(\mathcal{D});\Gamma)} \leq Ch^{\min(\ell+1,2)-1}\|v\|_{L^2(H^2(\mathcal{D});\Gamma)}. \tag{3.28}$$

With the help of the H^1–projection operator in (3.24a) and Cauchy-Schwarz's inequality (2.11), taking $\chi = \mathcal{R}_h(v)$, it is obvious that

$$\|\mathcal{R}_h(v)\|_{L^2(\mathcal{D})} \leq \|v\|_{L^2(\mathcal{D})} \quad \text{and} \quad \|\nabla(\mathcal{R}_h(v))\|_{L^2(\mathcal{D})} \leq \|\nabla v\|_{L^2(\mathcal{D})}.$$

By the L^2–projection operator in (3.23) and the approximation in (3.22), the bound for the second term in (3.27) is

$$\begin{aligned}\|\mathcal{R}_h(v-\Pi_q(v))\|_{L^2(H^1(\mathcal{D});\Gamma)} &\leq C\|v-\Pi_q(v)\|_{L^2(H^1(\mathcal{D});\Gamma)}\\ &\leq C\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}v\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}.\end{aligned} \tag{3.29}$$

Combining (3.28) and (3.29) yields

$$\begin{aligned}\|v-\widetilde{v}\|_{L^2(H^1(\mathcal{D});\Gamma)} &\leq Ch^{\min(\ell+1,2)-1}\|v\|_{L^2(H^2(\mathcal{D});\Gamma)} \\ &\quad+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}v\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!},\end{aligned}$$

which implies (3.26a). For the derivation of (3.26b), the same strategy is followed as:

$$\begin{aligned}&\|\nabla^2(v-\widetilde{v})\|_{L^2(L^2(\mathcal{D});\Gamma)} \\ &\quad\leq \|\nabla^2(v-\mathcal{R}_h(v))\|_{L^2(L^2(\mathcal{D});\Gamma)}+\|\nabla^2(\mathcal{R}_h(v)-\Pi_q(\mathcal{R}_h(v))\|_{L^2(L^2(\mathcal{D});\Gamma)} \\ &\quad= \|\nabla^2(v-\mathcal{R}_h(v))\|_{L^2(L^2(\mathcal{D});\Gamma)}+\|\nabla^2(\mathcal{R}_h(v-\Pi_q(v))\|_{L^2(L^2(\mathcal{D});\Gamma)}.\end{aligned}$$

An application of the inverse inequality (2.9) on $\mathcal{R}_h(v)$, the definition of H^1–projection operator (3.24a), and the Cauchy–Schwarz inequality (2.11) yields

$$\|\nabla^2(\mathcal{R}_h(v))\|_{L^2(\mathcal{D})}\leq Ch^{-1}\|\nabla(\mathcal{R}_h(v))\|_{L^2(\mathcal{D})}\leq Ch^{-1}\|\nabla v\|_{L^2(\mathcal{D})}. \tag{3.30}$$

By (3.22), (3.30), and (3.25),

$$\begin{aligned}\|\nabla^2(v-\tilde{v})\|_{L^2(L^2(\mathcal{D});\Gamma)} &\leq \|\nabla^2(v-\mathcal{R}_h(v))\|_{L^2(L^2(\mathcal{D});\Gamma)} \\ &\quad+Ch^{-1}\|\nabla(v-\Pi_q(v))\|_{L^2(L^2(\mathcal{D});\Gamma)} \\ &\leq Ch^{\min(\ell+1,2)-2}\|v\|_{L^2(H^2(\mathcal{D});\Gamma)} \\ &\quad+Ch^{-1}\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}v\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!},\end{aligned}$$

which is the desired result. □

The next step is to use Theorem 3.2.2 together with the approximation estimate (3.25) to derive an upper bound for the error in the energy norm.

Theorem 3.2.3. *Assume $y\in L^2(H^2(\mathcal{D});\Gamma)\cap H^{q+1}(H^1(\mathcal{D});\Gamma)$ and $y_h\in V_h\otimes\mathcal{S}_k^q$. Then, there is a constant C independent of y, h, and k_n such that*

$$\begin{aligned}\|y-y_h\|_{\xi} &\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}\right. \\ &\quad\left.+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right).\end{aligned} \tag{3.31}$$

Proof. Decompose $\|y-y_h\|_\xi$ as

$$\|y-y_h\|_\xi\leq\|y_h-\widetilde{y}\|_\xi+\|y-\widetilde{y}\|_\xi, \tag{3.32}$$

where $\widetilde{y} \in V_h \otimes \mathcal{S}_k^q$ is an approximation of the solution y, satisfying the Theorem 3.2.2.

Firstly, we will find a bound for the first term in (3.32). By the coercivity of the bilinear form (3.13a), the Galerkin orthogonality, an integration by parts over the convective term in the bilinear form (3.11), and the assumption on the convective term $\nabla \cdot \mathbf{b} = 0$, we obtain

$$
\begin{aligned}
c_{cv}\|y_h - \widetilde{y}\|_\xi^2 \leq \;& a_\xi(y_h - \widetilde{y}, y_h - \widetilde{y}) \\
=\;& a_\xi(y - \widetilde{y}, \underbrace{y_h - \widetilde{y}}_{\chi \in V_h}) + \underbrace{a_\xi(y - y_h, y_h - \widetilde{y})}_{=0} \\
=\;& a_\xi(y - \widetilde{y}, \chi) \\
=\;& \int_\Gamma \rho(\xi) \Bigg[\sum_{K \in \mathcal{T}_h} \int_K a\nabla(y - \widetilde{y}) \cdot \nabla\chi \, dx - \sum_{E \in \mathcal{E}_h^0 \cup \mathcal{E}_h^\partial} \int_E \{\!\{ a\nabla(y - \widetilde{y}) \}\!\} [\![\chi]\!] \, ds \\
& - \sum_{E \in \mathcal{E}_h^0 \cup \mathcal{E}_h^\partial} \int_E \{\!\{ a\nabla\chi \}\!\} [\![y - \widetilde{y}]\!] \, ds + \sum_{E \in \mathcal{E}_h^0 \cup \mathcal{E}_h^\partial} \frac{\sigma}{h_E} \int_E [\![y - \widetilde{y}]\!] \cdot [\![\chi]\!] \, ds \\
& - \sum_{K \in \mathcal{T}_h} \int_K \mathbf{b} \cdot (y - \widetilde{y}) \nabla\chi \, dx - \sum_{K \in \mathcal{T}_h} \int_{\partial K^+ \setminus \partial\mathcal{D}} \mathbf{b} \cdot \mathbf{n}_\mathbf{E} (y - \widetilde{y})(\chi^e - \chi) \, ds \\
& + \sum_{K \in \mathcal{T}_h} \int_{\partial K^+ \cap \mathcal{D}^+} \mathbf{b} \cdot \mathbf{n}_\mathbf{E} (y - \widetilde{y}) \chi \, ds \Bigg] d\xi \\
\leq \;& |T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7|. \qquad (3.33)
\end{aligned}
$$

With the help of the bound on $a(x, \omega)$ (3.2), Cauchy–Schwarz inequality (2.11), Young's inequality (2.12), and Theorem 3.2.2, the following bound is obtained for the first term in (3.33)

$$
\begin{aligned}
|T_1| \;=\;& \left| \int_\Gamma \left[\sum_{K \in \mathcal{T}_h} \int_K a\nabla(y - \tilde{y}) \cdot \nabla\chi \, dx \right] \rho(\xi) d\xi \right| \\
\leq \;& \int_\Gamma \sqrt{a_{\max}} \left(\sum_{K \in \mathcal{T}_h} \|\nabla(y - \widetilde{y})\|_{L^2(K)}^2 \right)^{\frac{1}{2}} \left(\sum_{K \in \mathcal{T}_h} \|\sqrt{a_{\max}} \nabla\chi\|_{L^2(K)}^2 \right)^{\frac{1}{2}} \rho(\xi) d\xi \\
\leq \;& \int_\Gamma \left(\frac{2}{c_{cv}} a_{\max} \sum_{K \in \mathcal{T}_h} \|\nabla(y - \widetilde{y})\|_{L^2(K)}^2 + \frac{c_{cv}}{8} \|\chi\|_e^2 \right) \rho(\xi) d\xi \\
\leq \;& C \sum_{K \in \mathcal{T}_h} \|\nabla(y - \widetilde{y})\|_{L^2(L^2(K);\Gamma)}^2 + \frac{c_{cv}}{8} \|\chi\|_\xi^2
\end{aligned}
$$

$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)$$
$$+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2. \tag{3.34}$$

Next, we derive estimates for the second and third terms in (3.33). An application of Cauchy–Schwarz inequality (2.11), Young's inequality (2.12), the trace inequality (2.8) for $E\in K_1^E\cap K_2^E$, and Theorem 3.2.2 yields

$$\begin{aligned}
|T_2|=&\left|\int_\Gamma\left[\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\int_E\{\!\{a\nabla(y-\widetilde{y})\}\!\}[\![\chi]\!]\,ds\right]\rho(\xi)d\xi\right|\\
\leq&\int_\Gamma\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\left(\frac{h_E}{\sigma a_{\max}}\|\{\!\{a\nabla(y-\widetilde{y})\}\!\}\|_{L^2(E)}^2\right)^{\frac{1}{2}}\left(\frac{\sigma a_{\max}}{h_E}\|[\![\chi]\!]\|_{L^2(E)}^2\right)^{\frac{1}{2}}\rho(\xi)d\xi\\
\leq&\int_\Gamma\left[\frac{c_{cv}}{8}\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{\sigma}{h_E}\|\chi\|_{L^2(E)}^2+\frac{2}{c_{cv}}\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{h_E}{\sigma}\|\{\!\{a\nabla(y-\widetilde{y})\}\!\}\|_{L^2(E)}^2\right]\rho(\xi)d\xi\\
\leq&C\int_\Gamma\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{h_E}{\sigma}h_E|K_1^E|^{-1}\left(\|\nabla(y-\widetilde{y})\|_{L^2(K_1^E)}+h_{K_1^E}\|\nabla^2(y-\widetilde{y})\|_{L^2(K_1^E)}\right)^2\\
&\times\rho(\xi)d\xi\\
&+C\int_\Gamma\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{h_E}{\sigma}h_E|K_2^E|^{-1}\left(\|\nabla(y-\widetilde{y})\|_{L^2(K_2^E)}+h_{K_2^E}\|\nabla^2(y-\widetilde{y})\|_{L^2(K_2^E)}\right)^2\\
&\times\rho(\xi)d\xi\\
&+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2\\
\leq&C\left(\|\nabla(y-\widetilde{y})\|_{L^2(L^2(\mathcal{D});\Gamma)}+h\|\nabla^2(y-\widetilde{y})\|_{L^2(H_0^1(\mathcal{D});\Gamma)}\right)^2+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2\\
\leq&C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2\\
&+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2,
\end{aligned}\tag{3.35}$$

$$\begin{aligned}
|T_3|=&\left|\int_\Gamma\left[\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\int_E\{\!\{a\nabla\chi\}\!\}[\![y-\widetilde{y}]\!]\,ds\right]\rho(\xi)d\xi\right|\\
\leq&\int_\Gamma\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\|\{\!\{a\nabla\chi\}\!\}\|_{L^2(E)}\|[\![y-\widetilde{y}]\!]\|_{L^2(E)}\,\rho(\xi)d\xi\\
\leq&\int_\Gamma\sum_{K\in\mathcal{T}_h}\left(C\left(\|y-\widetilde{y}\|_{L^2(K)}+h_K\|\nabla(y-\widetilde{y})\|_{L^2(K)}\right)a\|\nabla\chi\|_{L^2(K)}\right)\rho(\xi)d\xi
\end{aligned}$$

$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)\|\chi\|_{\xi}$$

$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2$$
$$+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2. \tag{3.36}$$

By Cauchy–Schwarz inequality (2.11), Young's inequality (2.12), the trace inequality (2.8), and Theorem 3.2.2, one can find an upper bound for T_4 in (3.33)

$$|T_4|\leq\int_{\Gamma}\left[\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^{\partial}}\frac{\sigma}{h_E}\int_E[\![(y-\widetilde{y})]\!]\cdot[\![\chi]\!]\right]\rho(\xi)d\xi$$

$$\leq\int_{\Gamma}\left(\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^{\partial}}\left(\frac{\sigma}{h_E}\right)\|[\![y-\widetilde{y}]\!]\|_{L^2(E)}^2\right)^{\frac{1}{2}}\left(\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^{\partial}}\left(\frac{\sigma}{h_E}\right)\|[\![\chi]\!]\|_{L^2(E)}^2\right)^{\frac{1}{2}}\rho(\xi)d\xi$$

$$\leq\frac{2}{c_{cv}}\int_{\Gamma}\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^{\partial}}\left(\frac{\sigma}{h_E}\right)\|[\![y-\widetilde{y}]\!]\|_{L^2(E)}^2\rho(\xi)d\xi$$
$$+\frac{c_{cv}}{8}\int_{\Gamma}\sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^{\partial}}\left(\frac{\sigma}{h_E}\right)\|[\![\chi]\!]\|_{L^2(E)}^2\rho(\xi)d\xi$$

$$\leq\frac{2}{c_{cv}}\int_{\Gamma}\sum_{K\in\mathcal{T}_h}C\left(\|y-\widetilde{y}\|_{L^2(K)}+h_K\|\nabla(y-\widetilde{y})\|_{L^2(K)}\right)^2\rho(\xi)d\xi+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2$$

$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2$$
$$+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2. \tag{3.37}$$

Now, for the convective terms in (3.33), which are T_5, T_6, and T_7, the estimates are derived by following the similar steps as done for T_1–T_4 in (3.34)-(3.37):

$$|T_5|=\left|\int_{\Gamma}\left[\sum_{K\in\mathcal{T}_h}\int_K\mathbf{b}\cdot(y-\widetilde{y})\nabla\chi\,dx\right]\rho(\xi)d\xi\right|$$

$$\leq\int_{\Gamma}\frac{\|\mathbf{b}\|_{L^\infty(\mathcal{D})}}{\sqrt{a_{\max}}}\left(\sum_{K\in\mathcal{T}_h}\|y-\widetilde{y}\|_{L^2(K)}^2\right)^{\frac{1}{2}}\left(\sum_{K\in\mathcal{T}_h}\|\sqrt{a_{\max}}\nabla\chi\|_{L^2(K)}^2\right)^{\frac{1}{2}}\rho(\xi)d\xi$$

$$\leq\int_{\Gamma}\left(\frac{2}{c_{cv}}\frac{\|\mathbf{b}\|_{L^\infty(\mathcal{D})}}{\sqrt{a_{\max}}}\sum_{K\in\mathcal{T}_h}\|y-\widetilde{y}\|_{L^2(K)}^2+\frac{c_{cv}}{8}\sum_{K\in\mathcal{T}_h}\|\sqrt{a_{\max}}\nabla\chi\|_{L^2(K)}^2\right)\rho(\xi)d\xi$$

$$\leq\frac{2}{c_{cv}}C\sum_{K\in\mathcal{T}_h}\int_{\Gamma}\|y-\widetilde{y}\|_{L^2(K)}^2\,\rho(\xi)d\xi+\frac{c_{cv}}{8}\int_{\Gamma}\|\chi\|_e^2\,\rho(\xi)d\xi$$

$$\leq \quad C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2$$
$$+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2, \tag{3.38}$$

$$|T_6| = \left|\int_{\Gamma}\left[\sum_{K\in\mathcal{T}_h}\int_{\partial K^+\setminus\partial\mathcal{D}}\mathbf{b}\cdot\mathbf{n_E}(y-\widetilde{y})(\chi^e-\chi)\,ds\right]\rho(\xi)d\xi\right|$$
$$\leq \int_{\Gamma}\left(\sum_{E\in\mathcal{E}_h^0}\|\sqrt{\mathbf{b}\cdot\mathbf{n_E}}(y-\widetilde{y})\|_{L^2(E)}^2\right)^{\frac{1}{2}}\left(\sum_{E\in\mathcal{E}_h^0}\|\sqrt{\mathbf{b}\cdot\mathbf{n_E}}(\chi^e-\chi)\|_{E}^2\right)^{\frac{1}{2}}\rho(\xi)d\xi$$
$$\leq \frac{2}{c_{cv}}C\sum_{K\in\mathcal{T}_h}\int_{\Gamma}\left(\|y-\widetilde{y}\|_{L^2(K)} + h_K\|\nabla(y-\widetilde{y})\|_{L^2(K)}\right)^2\rho(\xi)d\xi$$
$$+\frac{c_{cv}}{8}\int_{\Gamma}\|\chi\|_e^2\,\rho(\xi)d\xi$$
$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2$$
$$+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2, \tag{3.39}$$

$$|T_7| = \left|\int_{\Gamma}\left[\sum_{K\in\mathcal{T}_h}\int_{\partial K^+\cup\mathcal{D}^+}\mathbf{b}\cdot\mathbf{n_E}(y-\widetilde{y})\chi\,ds\right]\rho(\xi)d\xi\right|$$
$$\leq \int_{\Gamma}\left(\sum_{E\in\mathcal{E}_h^0}\|\sqrt{\mathbf{b}\cdot\mathbf{n_E}}(y-\widetilde{y})\|_{L^2(E)}^2\right)^{\frac{1}{2}}\left(\sum_{E\in\mathcal{E}_h^\partial}\|\sqrt{\mathbf{b}\cdot\mathbf{n_E}}\chi\|_{E}^2\right)^{\frac{1}{2}}\rho(\xi)d\xi$$
$$\leq \frac{2}{c_{cv}}C\sum_{K\in\mathcal{T}_h}\int_{\Gamma}\left(\|y-\widetilde{y}\|_{L^2(K)} + h_K\|\nabla(y-\widetilde{y})\|_{L^2(K)}\right)^2\rho(\xi)d\xi$$
$$+\frac{c_{cv}}{8}\int_{\Gamma}\|\chi\|_e^2\,\rho(\xi)d\xi$$

$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2$$
$$+\frac{c_{cv}}{8}\|\chi\|_{\xi}^2. \tag{3.40}$$

Combining the bounds of T_1–T_7 (3.34)-(3.40), the following result is obtained

$$\|y_h-\widetilde{y}\|_{\xi} \quad \leq \quad C\Bigg(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}$$
$$+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\Bigg). \tag{3.41}$$

Now, the second term in (3.32), i.e., $\|y-\widetilde{y}\|_\xi$ will be discussed. By the definition of energy norm in (3.12), it yields

$$\begin{aligned}
\|y-\widetilde{y}\|_\xi^2 &= \int_\Gamma \|y-\widetilde{y}\|_e^2\,\rho(\xi)d\xi \\
&= \int_\Gamma \Bigg[\sum_{K\in\mathcal{T}_h}\int_K a(.,\omega)(\nabla(y-\widetilde{y}))^2\,dx + \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{\sigma}{h_E}\int_E [\![y-\widetilde{y}]\!]^2\,ds \\
&\quad +\frac{1}{2}\sum_{E\in\mathcal{E}_h^\partial}\int_E \mathbf{b}(.,\omega)\cdot\mathbf{n_E}(y-\widetilde{y})^2 ds \\
&\quad +\frac{1}{2}\sum_{E\in\mathcal{E}_h^0}\int_E \mathbf{b}(.,\omega)\cdot\mathbf{n_E}((y-\widetilde{y})^e-(y-\widetilde{y}))^2 ds\Bigg]\rho(\xi)d\xi \\
&= A_1+A_2+A_3+A_4.
\end{aligned}$$

One can easily derive the following estimates as done in the previous steps

$$\begin{aligned}
A_1 \le & \int_\Gamma a_{\max}\sum_{K\in\mathcal{T}_h}\|\nabla(y-\widetilde{y})\|_{L^2(K)}^2\,\rho(\xi)d\xi \\
=& C\|\nabla(y-\widetilde{y})\|_{L^2(L^2(\mathcal{D});\Gamma)}^2 \\
\le & C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^N\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2,
\end{aligned} \tag{3.42}$$

$$\begin{aligned}
A_2 \le & \int_\Gamma \sum_{E\in\mathcal{E}_h^0\cup\mathcal{E}_h^\partial}\frac{\sigma}{h_E}\Bigg(Ch_E^{\frac{1}{2}}|K_1^E|^{-\frac{1}{2}}\left(\|y-\widetilde{y}\|_{L^2(K_1^E)}+h_{K_1^E}\|\nabla(y-\widetilde{y})\|_{L^2(K_1^E)}\right)\rho(\xi)d\xi \\
& +Ch_E^{\frac{1}{2}}|K_2^E|^{-\frac{1}{2}}\left(\|y-\widetilde{y}\|_{L^2(K_2^E)}+h_{K_2^E}\|\nabla(y-\widetilde{y})\|_{L^2(K_2^E)}\right)\Bigg)^2\rho(\xi)d\xi \\
\le & \int_\Gamma \sum_{K\in\mathcal{T}_h} C\Big(\|y-\widetilde{y}\|_{L^2(K)}+h_K\|\nabla(y-\widetilde{y})\|_{L^2(K)}\Big)^2\rho(\xi)d\xi \\
\le & C\Big(\|y-\widetilde{y}\|_{L^2(L^2(\mathcal{D});\Gamma)}+h\|\nabla(y-\widetilde{y})\|_{L^2(L^2(\mathcal{D});\Gamma)}\Big)^2 \\
\le & C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^N\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2,
\end{aligned} \tag{3.43}$$

$$\begin{aligned}
A_3 \le & \frac{1}{2}\int_\Gamma\sum_{E\in\mathcal{E}_h^\partial}|\mathbf{b}\cdot\mathbf{n_E}|\|y-\widetilde{y}\|_{L^2(E)}^2\,\rho(\xi)d\xi \\
\le & C\int_\Gamma\left(\|y-\widetilde{y}\|_{L^2(\mathcal{D})}+h\|\nabla(y-\widetilde{y})\|_{L^2(L^2(\mathcal{D});\Gamma)}\right)^2\rho(\xi)d\xi
\end{aligned}$$

$$\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2, \tag{3.44}$$

$$\begin{aligned}A_4 &\leq \frac{1}{2}\int_{\Gamma}\sum_{E\in\mathcal{E}_h^0}|\mathbf{b}\cdot\mathbf{n}_{\mathbf{E}}|\left(\|(y-\widetilde{y})^e\|_{L^2(E)}+\|(y-\widetilde{y})\|_{L^2(E)}\right)^2\rho(\xi)d\xi\\ &\leq C\big(\|y-\widetilde{y}\|_{L^2(L^2(\mathcal{D});\Gamma)}+h\|\nabla(y-\widetilde{y})\|_{L^2(L^2(\mathcal{D});\Gamma)}\big)^2\\ &\leq C\left(h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right)^2.\end{aligned} \tag{3.45}$$

Summation of the bounds of A_1–A_4 in (3.42)–(3.45) gives

$$\begin{aligned}\|y-\widetilde{y}\|_{\xi}^2 \;\leq\; C\Bigg(&h^{\min(\ell+1,2)-1}\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}\\ &+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\Bigg)^2.\end{aligned} \tag{3.46}$$

Finally, the desired result is obtained from (3.41) and (3.46). □

The length N of the random vector ξ in real applications, such as transport phenomena in random media, can be large, especially given the small correlation length in the random input's covariance function. This rapidly increases the size of multivariate stochastic basis polynomials J, which is a phenomenon known as the curse of dimensionality. In the following section, the curse of dimensionality is broken by utilizing a Kronecker–product structure of system matrices defined in (3.15), which decreases both storage requirements and computing complexity.

3.3 Low–Rank Approximation

In this section, efficient Krylov subspace solvers with suitable preconditioners, where the solution is approximated using a low–rank representation, are proposed in order to reduce memory requirements and computational effort. Fundamental operations associated with the low–rank format are significantly less expensive, and when the Krylov subspace technique converges, it constructs a sequence of the low–rank approximations to the system solution.

The idea behind a low-rank solution of the linear system is explained by giving the basic notation related to Kronecker products and low-rank approach. Let $\mathbf{y} = [y_1^T, \ldots, y_J^T]^T \in \mathbb{R}^{N_d J}$ with each y_i of length N_d and $\mathbf{Y} = [y_1, \ldots, y_J] \in \mathbb{R}^{N_d \times J}$, where N_d and J are the degree of freedoms for the spatial discretization and the total degree of the multivariate stochastic basis polynomials, respectively.

Following the properties in Section 2.1.1, the system (3.15) can be interpreted as $\mathcal{A}(\mathbf{Y}) = \mathcal{F}$ for the matrix $Y \in \mathbb{R}^{N_d \times J}$ with $\mathbf{y} = \mathtt{vec}(\mathbf{Y})$, where $\mathcal{A}(\mathbf{Y})$ is defined as the linear operator satisfying $\mathtt{vec}(\mathcal{A}(\mathbf{Y})) = \mathcal{A}\mathtt{vec}(\mathbf{Y})$. Assuming low–rank decomposition of $\mathbf{Y} = WV^T$ with

$$W = [w_1, \ldots, w_k] \in \mathbb{R}^{N_d \times r}, \quad V = [v_1, \ldots, v_k] \in \mathbb{R}^{J \times r}, \quad r \ll N_d, J$$

and

$$\mathtt{vec}(\mathbf{Y}) = \mathtt{vec}\left(\sum_{i=1}^{r} w_i v_i^T \right) = \sum_{i=1}^{r} v_i \otimes w_i,$$

we have

$$\begin{aligned} \mathcal{A}\mathtt{vec}(\mathbf{Y}) &= \left(\sum_{k=0}^{N} \mathcal{G}_k \otimes \mathcal{K}_k \right) \left(\sum_{i=1}^{r} v_i \otimes w_i \right) \\ &= \sum_{k=0}^{N} \sum_{i=1}^{r} (\mathcal{G}_k v_i) \otimes (\mathcal{K}_k w_i) \in \mathbb{R}^{N_d J}. \end{aligned}$$

This implies

$$\mathcal{A}(\mathbf{Y}) := \mathtt{mat}(\mathcal{A}\mathtt{vec}(\mathbf{Y})) \in \mathbb{R}^{N_d \times J}.$$

In [20, 69], it is shown that the solution to (3.15) can be approximated by its low–rank approximation if the system matrix and the right-hand side have a Kronecker-product structure. Thus, the focus of this section will be on a low–rank approximation of the solution $\mathbf{y}$ to the system (3.15).

3.3.1 Low-Rank Preconditioned Iterative Methods

In this section, the solution of the linear system is addressed by a tensor variant of the Krylov subspace methods combined with the low–rank approximation so that the

computational costs and memory requirements can be significantly reduced. Iterates in the algorithm are truncated depending on the decay of their singular values in the low–rank process. As a result, the iterates are formatted in low–rank form at each iteration; see, e.g., [16, 102].

In the following, the low–rank variants of the some Krylov subspace solvers are applied, namely, conjugate gradient (CG) method [89], bi–conjugate gradient stabilized (BiCGstab) [144], quasi–minimal residual variant of the bi–conjugate gradient stabilized (QMRCGstab) method [43], and generalized minimal residual (GMRES) [133] based on the low–rank approximation, where the advantage is taken of the Kronecker product of the matrix $\mathcal{A}$. Algorithms 1, 2, 3, and 4 show a low–rank implementation of the classical left preconditioned CG, BiCGstab, QMRCGstab, and GMRES methods, respectively. In principle, the low–rank truncation steps can affect the convergence of the Krylov method and the well–established properties of Krylov subspace may no longer hold. Therefore, in the implementations, a rather small truncation tolerance ϵ_{trunc} is used to maintain a very accurate representation of what the full–rank representation would like.

Algorithm 1 Low-rank preconditioned conjugate gradient (LRPCG)[20]

Input: Matrix functions $\mathcal{A}, \mathcal{P} : \mathbb{R}^{N_d \times J} \to \mathbb{R}^{N_d \times J}$, right–hand side $\mathcal{F}^n$ in low–rank format, truncation operator $\mathcal{T}$ with respect to given tolerance ϵ_{trunc}.

Output: Matrix $\mathbf{Y} \in \mathbb{R}^{N_d \times J}$, $\|\mathcal{A}(\mathbf{Y}) - \mathcal{F}\|_F \leq \varepsilon$.

$\mathbf{Y}_0 = 0$, $R_0 = \mathcal{F}^n$, $Z_0 = \mathcal{P}^{-1}(R_0)$, $S_0 = Z_0$, $Q_0 = \mathcal{A}(S_0)$
$\upsilon_0 = \langle S_0, Q_0 \rangle$, $k = 0$
while $\|R_k\|_F > \varepsilon$ **do**
  $\omega_k = \langle R_k, S_k \rangle / \upsilon_k$
  $\mathbf{Y}_{k+1} = \mathbf{Y}_k^n + \omega_k S_k$, $\mathbf{Y}_{k+1} \leftarrow \mathcal{T}(\mathbf{Y}_{k+1})$
  $R_{k+1} = \mathcal{F}^n - \mathcal{A}(\mathbf{Y}_{k+1})$, $R_{k+1} \leftarrow \mathcal{T}(R_{k+1})$
  $Z_{k+1} = \mathcal{P}^{-1}(R_{k+1})$
  $\beta_k = -\langle Z_{k+1}, Q_k \rangle / \upsilon_k$
  $S_{k+1} = Z_{k+1} + \beta_k S_k$, $S_{k+1} \leftarrow \mathcal{T}(S_{k+1})$
  $Q_{k+1} = \mathcal{A}(S_{k+1})$, $Q_{k+1} \leftarrow \mathcal{T}(Q_{k+1})$
  $\upsilon_{k+1} = \langle S_{k+1}, Q_{k+1} \rangle$
  $k = k + 1$
end while
$\mathbf{Y} = \mathbf{Y}_k$

Algorithm 2 Low–rank preconditioned BiCGstab (LRPBiCGstab)

Input: Matrix functions $\mathcal{A}, \mathcal{P} : \mathbb{R}^{N_d \times J} \to \mathbb{R}^{N_d \times J}$, right–hand side $\mathcal{F}$ in low–rank format, truncation operator $\mathcal{T}$ with respect to given tolerance ϵ_{trunc}.

Output: Matrix $\mathbf{Y} \in \mathbb{R}^{N_d \times J}$ satisfying $\|\mathcal{A}(\mathbf{Y}) - \mathcal{F}\|_F \leq \epsilon_{tol}$.

$\mathbf{Y}_0 = 0$, $R_0 = \mathcal{F}$, $\widetilde{R} = \mathcal{F}$, $\rho_0 = \langle \widetilde{R}, R_0 \rangle$, $S_0 = R_0$, $\widetilde{S}_0 = \mathcal{P}^{-1}(S_0)$, $V_0 = \mathcal{A}(\widetilde{S}_0)$, $k = 0$

while $\|R_k\|_F > \epsilon_{tol}$ **do**

  $\omega_k = \langle \widetilde{R}, R_k \rangle / \langle \widetilde{R}, V_k \rangle$

  $Z_k = R_k - \omega_k V_k$, $\quad Z_k \leftarrow \mathcal{T}(Z_k)$

  $\widetilde{Z}_k = \mathcal{P}^{-1}(Z_k)$, $\quad \widetilde{Z}_k \leftarrow \mathcal{T}(\widetilde{Z}_k)$

  $T_k = \mathcal{A}(\widetilde{Z}_k)$, $\quad T_k \leftarrow \mathcal{T}(T_k)$

  if $\|Z_k\|_F \leq \epsilon_{tol}$ **then**

    $\mathbf{Y} = \mathbf{Y}_k + \omega_k \widetilde{S}_k$

    return

  end if

  $\xi_k = \langle T_k, Z_k \rangle / \langle T_k, T_k \rangle$

  $\mathbf{Y}_{k+1} = \mathbf{Y}_k + \omega_k \widetilde{S}_k + \xi_k \widetilde{Z}_k$, $\quad \mathbf{Y}_{k+1} \leftarrow \mathcal{T}(\mathbf{Y}_{k+1})$

  $R_{k+1} = \mathcal{F} - \mathcal{A}(\mathbf{Y}_{k+1})$, $\quad R_{k+1} \leftarrow \mathcal{T}(R_{k+1})$

  if $\|R_{k+1}\|_F \leq \epsilon_{tol}$ **then**

    $\mathbf{Y} = \mathbf{Y}_{k+1}$

    return

  end if

  $\rho_{k+1} = \langle \widetilde{R}, R_{k+1} \rangle$

  $\beta_k = \frac{\rho_{k+1}}{\rho_k} \frac{\omega_k}{\xi_k}$

  $S_{k+1} = R_{k+1} + \beta_k (S_k - \xi_k V_k)$, $\quad S_{k+1} \leftarrow \mathcal{T}(S_{k+1})$

  $\widetilde{S}_{k+1} = \mathcal{P}^{-1}(S_{k+1})$, $\quad \widetilde{S}_{k+1} \leftarrow \mathcal{T}(\widetilde{S}_{k+1})$

  $V_{k+1} = \mathcal{A}(\widetilde{S}_{k+1})$, $\quad V_{k+1} \leftarrow \mathcal{T}(V_{k+1})$

  $k = k + 1$

end while

Truncation operators are used at each iteration step of the algorithm, and these operations have a significant impact on the entire solution process. One of the issues in the low–rank approximation is that the rank of the low–rank factors can increase either via matrix vector products or vector (matrix) additions. The crucial way to prevent this case is to truncate the iterates and force their ranks to remain low. Therefore, it is needed to find new low–rank approximations $\widetilde{W}$ and $\widetilde{V}$ that approximate the old approximations $\mathbf{Y} \approx WV^T \approx \widetilde{W}\widetilde{V}^T$ by using the truncation operator $\mathcal{T}$. As a result, rank–reduction techniques are required to keep costs under control, such as truncation based on singular values [102] or truncation based on coarse–grid rank reduction [109]. As pointed out in [102], we define a truncation operator for a given matrix in the following.

The truncation operator $\mathbf{Y} \leftarrow \mathcal{T}(\mathbf{Y})$ compresses a matrix $\mathbf{Y} \approx WV^T$ in the low–rank format with $W \in \mathbb{R}^{N_d \times r},\ V \in \mathbb{R}^{J \times r}$ such that

$$\|WV^T - \widetilde{W}\widetilde{V}^T\|_F \leq \epsilon_{trunc}.$$

For this purpose, QR factorizations of both matrices $W = Q_W R_W$, $V = Q_V R_V$ are computed and then it can be written as

$$\mathbf{Y} = Q_W R_W R_V^T Q_V^T.$$

Then, applying singular value decomposition (SVD) [81]

$$R_W R_V^T = B\ \text{diag}(\sigma_1, \ldots, \sigma_r)\, C^T,$$

one can obtain a new low–rank representation. Here, the truncation rank $\tilde{r} \leq r$ is chosen provided that the smallest integer satisfy

$$\sqrt{\sigma_{\tilde{r}+1}^2 + \ldots + \sigma_r} \leq \epsilon_{trunc} \sqrt{\sigma_1^2 + \ldots + \sigma_r},$$

where ϵ_{trunc} is truncation tolerance. In MATLAB notation, the new low–rank factors are set by

$$\widetilde{W} = Q_W\, C(:, 1:\widetilde{r})\ \text{diag}(\sigma_1, \ldots, \sigma_{\widetilde{r}}), \quad \text{and} \quad \widetilde{V} = Q_V\, B(:, 1:\widetilde{r}).$$

In this thesis, following the discussion in [20, 139], a more economical alternative could be possible to compute singular values a truncated SVD of $\mathbf{Y} = WV^T \approx$

$B\,\mathrm{diag}(\sigma_1,\ldots,\sigma_r)C^T$ associated to the r singular values that are larger than the given truncation threshold. In this way, the new low–rank representation $\mathbf{Y} \approx \widetilde{W}\widetilde{V}^T$ is obtained by keeping both the rank of low-rank factor and cost under control.

The other issue in the low–rank process is the computation of the inner product. It can also be done easily by applying the following strategy:

$$\langle X, Z\rangle = \mathtt{vec}(X)^T\mathtt{vec}(Z) = \mathrm{trace}(X^TZ)$$

for the low-rank matrices

$$\begin{aligned} X &= W_X V_X^T \quad W_X \in \mathbb{R}^{N_d\times r_X},\ V_X \in \mathbb{R}^{J\times r_X},\\ Z &= W_Z V_Z^T \quad W_Z \in \mathbb{R}^{N_d\times r_Z},\ V_Z \in \mathbb{R}^{J\times r_Z}. \end{aligned}$$

Then, one can easily show that

$$\mathrm{trace}(X^TZ) = \mathrm{trace}\left((W_XV_X^T)^T(W_ZV_Z^T)\right) = \mathrm{trace}\left((V_Z^TV_X)(W_X^TW_Z)\right)$$

allows us to compute the trace of small matrices rather than of the ones from the full discretization.

Preconditioning is known to be necessary for Krylov subspace techniques in order to get a fast convergence in terms of the number of iterations, and low–rank Krylov methods are no exception. The preconditioning operator decreases the number of iterations at an acceptable computational cost, but it must not significantly increase the memory requirements of the solution process. The following well-known preconditioners in the context of PDEs with uncertainty are listed here:

i) Mean-based preconditioner

$$\mathcal{P}_0 = \mathcal{G}_0 \otimes \mathcal{K}_0$$

is one of the most commonly used preconditioners for solving PDEs with random data; see, e.g., [75, 128]. One can easily observe that $\mathcal{P}_0$ is block diagonal matrix since $\mathcal{G}_0$ is a diagonal matrix due to the orthogonality of the stochastic basis functions Ψ_i.

Algorithm 3 Low–rank preconditioned QMRCGstab (LRPQMRCGstab)

Input: Matrix functions $\mathcal{A}, \mathcal{P} : \mathbb{R}^{N_d \times J} \to \mathbb{R}^{N_d \times J}$, right–hand side $\mathcal{F}$ in low–rank format, truncation operator $\mathcal{T}$ with respect to given tolerance ϵ_{trunc}.

Output: Matrix $\mathbf{Y} \in \mathbb{R}^{N_d \times J}$ satisfying $\|\mathcal{A}(\mathbf{Y}) - \mathcal{F}\|_F \leq \epsilon_{tol}$.

$R_0 = \mathcal{F} - \mathcal{A}(\mathbf{Y}_0)$, for some initial guess $\mathbf{Y}_0$.

$Z_0 = \mathcal{P}^{-1}(R_0)$

Choose $\widetilde{R}_0$ such that $\left\langle Z_0, \widetilde{R}_0 \right\rangle \neq 0$ (for example, $\widetilde{R}_0 = R_0$).

$Q_0 = V_0 = D_0 = 0$

$\rho_0 = \alpha_0 = \omega_0 = 1, \tau_0 = \|Z_0\|_F, \theta_0 = 0, \eta_0 = 0, k = 0$

while $\sqrt{k+1}|\widetilde{\tau}|/\|R_0\|_F > \epsilon_{tol}$ **do**

$\rho_{k+1} = \left\langle Z_k, \widetilde{R}_0 \right\rangle,\ \beta_{k+1} = \frac{\rho_{k+1}}{\rho_k}\frac{\alpha_k}{\omega_k}$

$Q_{k+1} = Z_k + \beta_{k+1}(Q_k - \omega_k V_k),$ $\quad Q_{k+1} \leftarrow \mathcal{T}(Q_{k+1})$

$\widetilde{Q}_{k+1} = \mathcal{A}(Q_{k+1}),$ $\quad \widetilde{Q}_{k+1} \leftarrow \mathcal{T}(\widetilde{Q}_{k+1})$

if $\|\widetilde{Q}_{k+1}\|_F \leq \epsilon_{tol}$ **then**

$\mathbf{Y} = \mathbf{Y}_k$

return

end if

$V_{k+1} = \mathcal{P}^{-1}(\widetilde{Q}_{k+1}),$ $\quad V_{k+1} \leftarrow \mathcal{T}(V_{k+1})$

$\alpha_{k+1} = \rho_{k+1} / \left\langle V_{k+1}, \widetilde{R}_0 \right\rangle$

$S_{k+1} = Z_k - \alpha_{k+1} V_{k+1},$ $\quad S_{k+1} \leftarrow \mathcal{T}(S_{k+1})$

$\widetilde{\tau} = \tau\widetilde{\theta}_{k+1}c,\ \widetilde{\eta}_{k+1} = c^2\alpha_{k+1}$

$\widetilde{D}_{k+1} = Q_{k+1} + \dfrac{\theta_k^2\eta_k}{\alpha_{k+1}} D_k,$ $\quad \widetilde{D}_{k+1} \leftarrow \mathcal{T}(\widetilde{D}_{k+1})$

$\widetilde{\mathbf{Y}}_{k+1} = \mathbf{Y}_k + \widetilde{\eta}_{k+1}\widetilde{D}_{k+1},$ $\quad \widetilde{\mathbf{Y}}_{k+1} \leftarrow \mathcal{T}(\widetilde{\mathbf{Y}}_{k+1})$

$\widetilde{S}_{k+1} = \mathcal{A}(S_{k+1}),$ $\quad \widetilde{S}_{k+1} \leftarrow \mathcal{T}(\widetilde{S}_{k+1})$

$T_{k+1} = \mathcal{P}^{-1}(\widetilde{S}_{k+1}),$ $\quad T_{k+1} \leftarrow \mathcal{T}(T_{k+1})$

$\omega_{k+1} = \langle S_{k+1}, T_{k+1} \rangle \,/\, \langle T_{k+1}, T_{k+1} \rangle$

$Z_{k+1} = S_{k+1} - \omega_{k+1} T_{k+1}$

$\theta_{k+1} = \|Z_{k+1}\|_F / \widetilde{\tau},\ c = \dfrac{1}{\sqrt{1 + \theta_{k+1}^2}}$

$\tau = \widetilde{\tau}\theta_{k+1}c,\ \eta_{k+1} = c^2\omega_{k+1}$

$D_{k+1} = S_{k+1} + \dfrac{\widetilde{\theta}_{k+1}^2\widetilde{\eta}_{k+1}}{\omega_{k+1}} \widetilde{D}_{k+1},$ $\quad D_{k+1} \leftarrow \mathcal{T}(D_{k+1})$

$\mathbf{Y}_{k+1} = \widetilde{\mathbf{Y}}_{k+1} + \eta_{k+1} D_{k+1},$ $\quad \mathbf{Y}_{k+1} \leftarrow \mathcal{T}(\mathbf{Y}_{k+1})$

$k = k + 1$

end while

$\mathbf{Y} = \mathbf{Y}_k$

Algorithm 4 Low–rank preconditioned GMRES (LRPGMRES)

Input: Matrix functions $\mathcal{A}, \mathcal{P} : \mathbb{R}^{N_d \times J} \to \mathbb{R}^{N_d \times J}$, right–hand side $\mathcal{F}$ in low–rank format, truncation operator $\mathcal{T}$ with respect to given tolerance ϵ_{trunc}.

Output: Matrix $\mathbf{Y} \in \mathbb{R}^{N_d \times J}$ satisfying $\|\mathcal{A}(\mathbf{Y}) - \mathcal{F}\|_F \leq \epsilon_{tol}$.

$R_0 = \mathcal{F} - \mathcal{A}(\mathbf{Y}_0)$, for some initial guess $\mathbf{Y}_0$.
$V_1 = R_0 / \|R_0\|_F$
$\xi = [\xi_1, 0, \ldots, 0], \qquad \xi_1 = \|V_1\|_F$
for $k = 1, \ldots,$ maxit **do**
  $Z_k = \mathcal{P}^{-1}(V_k),$ $\qquad Z_k \leftarrow \mathcal{T}(Z_k)$
  $W = \mathcal{A}(Z_k),$ $\qquad W \leftarrow \mathcal{T}(W)$
  for $i = 1, \ldots, k$ **do**
    $h_{i,k} = \langle W, V_i \rangle$
    $W = W - h_{i,k} V_i,$ $\qquad W \leftarrow \mathcal{T}(W)$
  end for
  $h_{k+1,k} = \|W\|_F$
  $V_{k+1} = W / h_{k+1,k}$
  Apply Givens rotations to kth column of h, i.e.,
  for $i = 1, \ldots, k-1$ **do**
    $$\begin{bmatrix} h_{i,k} \\ h_{i+1,k} \end{bmatrix} = \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} h_{i,k} \\ h_{i+1,k} \end{bmatrix}$$
  end for
  Compute kth rotation, and apply to ξ and last column of h.
  $$\begin{bmatrix} \xi_k \\ \xi_{k+1} \end{bmatrix} = \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} \xi_k \\ 0 \end{bmatrix}$$
  $h_{k,k} = c_k h_{k,k} + s_k h_{k+1,k}, \qquad h_{k+1,k} = 0$
  if $|\xi_{k+1}|$ sufficiently small **then**
    Solve $Hq = \xi$, where the entries of H are $h_{j,k}$.
    $Q = [q_1 V_1, \ldots, q_k V_k],$ $\qquad Q \leftarrow \mathcal{T}(Q)$
    $\widetilde{Q} = \mathcal{P}^{-1}(Q),$ $\qquad \widetilde{Q} \leftarrow \mathcal{T}(\widetilde{Q})$
    $\mathbf{Y} = \mathbf{Y}_0 + \widetilde{Q},$ $\qquad \mathbf{Y} \leftarrow \mathcal{T}(\mathbf{Y})$
    return
  end if
end for

ii) Ullmann preconditioner, which is of the form

$$\mathcal{P}_1 = \underbrace{\mathcal{G}_0 \otimes \mathcal{K}_0}_{:=\mathcal{P}_0} + \sum_{k=1}^{N} \frac{\text{trace}(\mathcal{K}_k^T \mathcal{K}_0)}{\text{trace}(\mathcal{K}_0^T \mathcal{K}_0)} \mathcal{G}_k \otimes \mathcal{K}_0,$$

can be considered as a modified version of $\mathcal{P}_0$ as discussed in [143]. One of the advantages of this preconditioner is keeping the structure of the coefficient matrix, which is, in this case, the sparsity pattern. Moreover, unlike the mean–based preconditioner, it uses the whole information in the coefficient matrix. However, this advantage causes $\mathcal{P}_1$ being more expensive since it is not block diagonal anymore.

3.4 Unsteady Model Problem with Random Coefficients

Next, the discussion in Section 3.1 is extended to unsteady convection diffusion equations with random coefficients: find $y : \overline{\mathcal{D}} \times \Omega \times [0,T] \to \mathbb{R}$ such that $\mathbb{P}$-almost surely in Ω

$$\begin{aligned}
\frac{\partial y(\boldsymbol{x},\omega,t)}{\partial t} - \nabla \cdot (a(\boldsymbol{x},\omega)\nabla y(\boldsymbol{x},\omega,t)) & \\
+\mathbf{b}(\boldsymbol{x},\omega)\cdot\nabla y(\boldsymbol{x},\omega,t) &= f(\boldsymbol{x},t), \quad \text{in } \mathcal{D}\times\Omega\times(0,T], && \text{(3.47a)}\\
y(\boldsymbol{x},\omega,t) &= 0, \quad \text{on } \partial\mathcal{D}\times\Omega\times[0,T], && \text{(3.47b)}\\
y(\boldsymbol{x},\omega,0) &= y^0(\boldsymbol{x}), \quad \text{in } \mathcal{D}\times\Omega, && \text{(3.47c)}
\end{aligned}$$

where $y^0(\boldsymbol{x}) \in L^2(\mathcal{D})$ corresponds to the deterministic initial condition.

By following the methodologies introduced for the stationary problem in Section 3.1 and the backward Euler method in temporal space with the uniform time step $\Delta t = T/N$, the fully discrete form can be written as

$$\frac{1}{\Delta t}\big(y^{n+1} - y^n, v\big)_\xi + a_\xi(y^{n+1}, v) = l_\xi(t_{n+1}, v), \quad \forall w, v \in V_h \otimes \mathcal{Y}_k^q, \qquad (3.48)$$

where

$$\big(w, v\big)_\xi = \int_\Gamma \int_\mathcal{D} wv \; d\boldsymbol{x} \; \rho(\xi)\, d\xi \quad \forall w, v \in V_h \otimes \mathcal{Y}_k^q.$$

Then, one can obtain the following system of ordinary equations with block structure:

$$(\mathcal{G}_0 \otimes M)\left(\frac{\mathbf{y}^{n+1} - \mathbf{y}^n}{\Delta t}\right) + \left(\sum_{k=0}^{N} \mathcal{G}_k \otimes \mathcal{K}_k\right)\mathbf{y}^{n+1} = \Big(g_0 \otimes f_0\Big)^{n+1},$$

or, equivalently,

$$\mathcal{M}\left(\frac{\mathbf{y}^{n+1}-\mathbf{y}^{n}}{\Delta t}\right)+A\mathbf{y}^{n+1}=F^{n+1}, \tag{3.49}$$

where

$$A=\sum_{k=0}^{N}\mathcal{G}_k\otimes\mathcal{K}_k,\quad \mathcal{M}=\mathcal{G}_0\otimes M,\quad F^{n+1}=\Big(g_0\otimes f_0\Big)^{n+1}.$$

Rearranging (3.49), the following matrix form of the discrete systems is obtained:

$$\mathcal{A}\mathbf{y}^{n+1}=\mathcal{F}^{n+1}, \tag{3.50}$$

where for $k=1,\ldots,N$

$$\begin{aligned}\mathcal{A} &= \mathcal{G}_0\otimes\underbrace{(M+\Delta t\mathcal{K}_0)}_{\widehat{\mathcal{K}}_0}+\Big(\sum_{k=1}^{N}\mathcal{G}_k\otimes\underbrace{(\Delta t\mathcal{K}_k)}_{\widehat{\mathcal{K}}_k}\Big),\\ \mathcal{F}^{n+1} &= \mathcal{M}\mathbf{y}^{n}+\Delta t F^{n+1}.\end{aligned}$$

It is also noted that the time dependence of the problem introduces additional complexity of solving a large linear system for each time step. Therefore, it is necessary to apply the low–rank approximation technique introduced in Section 3.3 for each fixed time step.

Next, the following sections state the stability analysis and a priori error estimates of the proposed method on the energy norm defined in (3.12), respectively.

3.4.1 Stability Analysis

It is generally of interest to know something about the long term behaviours of the solution to a time-dependent equation. In particular, one would like to know if the solution grows with time or if it can be bounded by the known data of the equation. For the unsteady problem, stability estimates are crucial. Theorem 3.4.1 shows that the proposed method is bounded by the known data, which are the initial condition y^0 and the right hand side function f.

Theorem 3.4.1. *There exists a constant C independent of h and Δt such that for all $m>0$*

$$\|y^m\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\Delta t\sum_{n=1}^{m}\|y^n\|^2_{\xi}\leq C\Big(\|y^0\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\Delta t\sum_{n=1}^{m}\|f^n\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\Big).$$

Proof. Taking $v = y^{n+1}$ in the fully discrete system (3.48), it is obtained

$$\frac{1}{\Delta t}\int_{\Gamma}\int_{\mathcal{D}}(y^{n+1}-y^{n})\,y^{n+1}\,d\boldsymbol{x}\,\rho(\xi)\,d\xi + a_{\xi}(y^{n+1},y^{n+1}) = l_{\xi}(t_{n+1},y^{n+1}).$$

An application of the polarization identity

$$\forall x,y\in\mathbb{R},\quad \frac{1}{2}(x^2-y^2)\leq\frac{1}{2}(x^2-y^2+(x-y)^2)=(x-y)x, \tag{3.51}$$

yields

$$\frac{1}{2\Delta t}\left(\|y^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}-\|y^{n}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\right)+a_{\xi}(y^{n+1},y^{n+1})=l_{\xi}(t_{n+1},y^{n+1}). \tag{3.52}$$

From the coercivity of a_ξ (3.13a), Cauchy-Schwarz (2.11), and Young's inequalities (2.12), the expression (3.52) reduces to

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|y^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}-\|y^{n}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\right)+\frac{c_{cv}}{2}\|y^{n+1}\|^2_{\xi}\leq|l_{\xi}(t_{n+1},y^{n+1})|\\
&\quad\leq\|f^{n+1}\|_{L^2(L^2(\mathcal{D});\Gamma)}\|y^{n+1}\|_{L^2(L^2(\mathcal{D});\Gamma)}\\
&\quad\leq\frac{1}{2}\|f^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\frac{1}{2}\|y^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}.
\end{aligned}$$

Multiplying by $2\Delta t$ and summing from $n=0$ to $n=m-1$, one can get

$$\begin{aligned}
\|y^{m}\|^2_{L^2(L^2(\mathcal{D});\Gamma)} &- \|y^{0}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\Delta t c_{cv}\sum_{n=1}^{m}\|y^{n}\|^2_{\xi}\\
&\leq \Delta t\sum_{n=1}^{m}\|f^{n}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\Delta t\sum_{n=1}^{m}\|y^{n}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}.
\end{aligned}$$

After applying discrete Gronwall inequality given in Section 2.4.4, the desired result is obtained

$$\|y^{m}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\Delta t\sum_{n=1}^{m}\|y^{n}\|^2_{\xi}\leq C\left(\|y^{0}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\Delta t\sum_{n=1}^{m}\|f^{n}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\right),$$

where the constant C is independent of h and Δt. □

3.4.2 Error Analysis

In the following, a priori error estimates is derived for the unsteady stochastic problem (3.47) by the following procedure as done for the stationary problem in Section 3.2.

Assuming that $y \in L^2(H^2(\mathcal{D});\Gamma;[0,T]) \cap H^{q+1}(H^1(\mathcal{D});\Gamma;[0,T])$ and the best approximation $\widetilde{y} \in V_h \otimes \mathcal{S}_k^q$, it is known from Theorem (3.2.3) that for $\forall t \geq 0$,

$$\begin{aligned} \|y(t)-\widetilde{y}(t)\|_{L^2(H^1(\mathcal{D});\Gamma)} &\leq C\Bigg(h^{\min(\ell+1,2)-1}\|y(t)\|_{L^2(H^2(\mathcal{D});\Gamma)} \\ &\quad + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y(t)\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\Bigg), \end{aligned} \tag{3.53}$$

and

$$\begin{aligned} \|y(t)-\widetilde{y}(t)\|_{\xi} &\leq C\Bigg(h^{\min(\ell+1,2)-1}\|y(t)\|_{L^2(H^2(\mathcal{D});\Gamma)} \\ &\quad + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y(t)\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\Bigg). \end{aligned} \tag{3.54}$$

Using the linearity of the bilinear form, once can easily show that

$$a_\xi\Big(\frac{d}{dt}(y-\tilde{y})(t), v\Big) = 0, \qquad \forall t \geq 0,\ \forall v \in V_h \otimes \mathcal{S}_k^q.$$

Theorem 3.4.2. *Assume that the exact solution to problem* (3.47) *satisfies*

$$\begin{aligned} y &\in L^2(H^2(\mathcal{D});\Gamma;[0,T]) \cap H^{q+1}(H^1(\mathcal{D});\Gamma;[0,T]), \\ \frac{\partial^2 y}{\partial t^2} &\in L^2(L^2(\mathcal{D});\Gamma;[0,T]). \end{aligned}$$

Then, there exists a constant C independent of h and Δt such that for all $m > 0$

$$\begin{aligned} \|y_h^m - y^m\|_{L^2(L^2(\mathcal{D});\Gamma)} &\leq C\Delta t\Big\|\frac{\partial^2 y}{\partial t^2}\Big\|_{L^2(L^2(\mathcal{D});\Gamma;[0,T])} \\ &\quad + C\sqrt{\Delta t}\Bigg(h^{\min(\ell+1,2)-1}\Big\|\frac{\partial y}{\partial t}\Big\|_{L^2(H^2(\mathcal{D});\Gamma;[0,T])} \\ &\quad + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}\frac{\partial y}{\partial t}\|_{L^2(H^1(\mathcal{D});\Gamma;[0,T])}}{(q_n+1)!}\Bigg), \end{aligned} \tag{3.55a}$$

$$\begin{aligned} \left(\Delta t\sum_{n=1}^{m}\|y_h^n - y^n\|_{\xi}^2\right)^{1/2} &\leq C\Delta t\Big\|\frac{\partial^2 y}{\partial t^2}\Big\|_{L^2(L^2(\mathcal{D});\Gamma;[0,T])} \\ &\quad + C\sqrt{\Delta t}\Bigg(h^{\min(\ell+1,2)-1}\Big\|\frac{\partial y}{\partial t}\Big\|_{L^2(H^2(\mathcal{D});\Gamma;[0,T])} \\ &\quad + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}\frac{\partial y}{\partial t}\|_{L^2(H^1(\mathcal{D});\Gamma;[0,T])}}{(q_n+1)!}\Bigg). \end{aligned} \tag{3.55b}$$

Proof. Let $\tilde{y}$ be the best approximation of y. It is obvious that

$$\Big(\frac{y^{n+1}-y^n}{\Delta t},v\Big)_\xi-\Big(\frac{y_h^{n+1}-y_h^n}{\Delta t},v\Big)_\xi+a_\xi(y^{n+1},v)-a_\xi(y_h^{n+1},v)=0. \quad (3.56)$$

Denoting $y^n-y_h^n=\rho^n-\chi^n$ with $\rho^n=y^n-\tilde{y}^n$ and $\chi^n=y_h^n-\tilde{y}^n$, substituting into (3.56) and by the Galerkin orthogonality, one can have

$$\begin{aligned}\Big(\frac{\chi^{n+1}-\chi^n}{\Delta t},v\Big)_\xi+a_\xi(\chi^{n+1},v) &= \Big(\frac{\partial y}{\partial t}^{n+1}-\frac{y^{n+1}-y^n}{\Delta t},v\Big)_\xi\\ &\quad+\Big(\frac{\rho^{n+1}-\rho^n}{\Delta t},v\Big)_\xi+\underbrace{a_\xi(\rho^{n+1},v)}_{=0}.\end{aligned}$$

Setting $v=\chi^{n+1}$ and $\Theta^{n+1}=\frac{\partial y}{\partial t}^{n+1}-\frac{y^{n+1}-y^n}{\Delta t}$, and using the polarization identity (3.51), the coercivity of $a_\xi(\cdot,\cdot)$ (3.13a), Cauchy-Schwarz inequality (2.11), and Young's inequality (2.12), it holds

$$\begin{aligned}&\frac{1}{2\Delta t}\Big(\|\chi^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}-\|\chi^n\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\Big)+c_{cv}\|\chi^{n+1}\|^2_\xi\\ &\qquad\le\frac{c_{cv}}{2}\|\chi^{n+1}\|^2_\xi+C\Big(\|\Theta^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\|\frac{\rho^{n+1}-\rho^n}{\Delta t}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\Big).\end{aligned}$$

Then, it is obvious

$$\begin{aligned}&\frac{1}{2\Delta t}\Big(\|\chi^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}-\|\chi^n\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\Big)+\frac{c_{cv}}{2}\|\chi^{n+1}\|^2_\xi\\ &\qquad\le C\Big(\|\Theta^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}+\|\frac{\rho^{n+1}-\rho^n}{\Delta t}\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\Big). \quad (3.57)\end{aligned}$$

The definition of Θ^{n+1} and Taylor series expansion yield

$$\Theta^{n+1}=\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}(t-t_n)\frac{\partial^2 y(t)}{\partial t^2}\,dt, \quad \text{and} \quad \rho^{n+1}-\rho^n=\int_{t_n}^{t_{n+1}}\frac{\partial\rho}{\partial t}\,dt.$$

Using Cauchy-Schwarz inequality (2.11), one can easily prove that

$$\|\Theta^{n+1}\|^2_{L^2(L^2(\mathcal{D});\Gamma)} \le \frac{\Delta t}{3}\int_{t_n}^{t_{n+1}}\Big\|\frac{\partial^2 y(t)}{\partial t^2}\Big\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\,dt, \quad (3.58a)$$

$$\|\rho^{n+1}-\rho^n\|^2_{L^2(L^2(\mathcal{D});\Gamma)} \le \int_{t_n}^{t_{n+1}}\Big\|\frac{\partial\rho}{\partial t}\Big\|^2_{L^2(L^2(\mathcal{D});\Gamma)}\,dt. \quad (3.58b)$$

Inserting (3.58a) and (3.58b) into (3.57), multiplying the inequality by $2\Delta t$, and summing from $n = 0$ to $n = m - 1$, it can be obtained

$$\|\chi^m\|^2_{L^2(L^2(\mathcal{D});\Gamma)} - \|\chi^0\|^2_{L^2(L^2(\mathcal{D});\Gamma)} + c_{cv}\Delta t \sum_{n=1}^{m} \|\chi^{n+1}\|^2_{\xi} \leq C\Delta t^2 \int_0^T \left\| \frac{\partial^2 y}{\partial t^2} \right\|^2_{L^2(L^2(\mathcal{D});\Gamma)} dt + C\Delta t \int_0^T \left\| \frac{\partial \rho}{\partial t} \right\|^2_{L^2(L^2(\mathcal{D});\Gamma)} dt.$$

Since χ^0 is zero and $\frac{\partial \rho}{\partial t}$ satisfies the error estimates (3.53), it yields

$$\begin{aligned}\|\chi^m\|^2_{L^2(L^2(\mathcal{D});\Gamma)} + c_{cv}\Delta t \sum_{n=1}^{m} \|\chi^{n+1}\|^2_{\xi} \leq & C\Delta t^2 \int_0^T \left\| \frac{\partial^2 y}{\partial t^2} \right\|^2_{L^2(L^2(\mathcal{D});\Gamma)} dt \\ & + C\Delta t \int_0^T \left(h^{\min(\ell+1,2)-1} \left\| \frac{\partial y(t)}{\partial t} \right\|_{L^2(H^2(\mathcal{D});\Gamma)} \right. \\ & \left. + \sum_{n=1}^{N} \left(\frac{k_n}{2} \right)^{q_n+1} \frac{\|\partial^{q_n+1}_{\xi_n} \frac{\partial y(t)}{\partial t}\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!} \right)^2 dt.\end{aligned}$$

Using triangle inequality, the desired results in (3.55) are obtained. □

3.5 Numerical Results

This section presents several numerical results to examine the quality of the proposed numerical approaches. As mentioned before, we are here interested in the statistical moments of the solution $y(\boldsymbol{x}, \omega)$ in (3.1), such as mean and variance rather than, the solution $y(\boldsymbol{x}, \omega)$. The numerical experiments are performed on an Ubuntu Linux machine with 32 GB RAM using MATLAB R2020a. To compare the performance of the solution methods, the rank of the computed solution, the number of performed iterations, the computational time, the relative residual, that is, $\|\mathcal{A}\mathbf{y} - \mathcal{F}\|_F / \|\mathcal{F}\|_F$, and the memory demand of the solution are reported. Unless otherwise stated, in all simulations, iterative methods are terminated when the residual, measured in the Frobenius norm, is reduced to $\epsilon_{tol} = 10^{-4}$ or the maximum iteration number ($\#iter_{max} = 100$) is reached. It is noted that the tolerance ϵ_{tol} should be chosen, such that $\epsilon_{trunc} \leq \epsilon_{tol}$; otherwise, one would be essentially iterating on the noise from the low–rank truncations.

In the numerical experiments, the random input z is characterized by the covariance function

$$C_z(\boldsymbol{x}, \boldsymbol{y}) = \kappa_z^2 \prod_{n=1}^{2} e^{-|x_n - y_n|/\ell_n} \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D} \tag{3.59}$$

with the correlation length ℓ_n and the standard deviation κ_z. This section uses linear elements to generate discontinuous Galerkin basis and Legendre polynomials as stochastic basis functions since the underlying random variables have uniform distribution over $[-\sqrt{3}, \sqrt{3}]$, that is, it is chosen as $\xi = \{\xi_1, \ldots, \xi_N\}$ such that $\xi_j \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$. The eigenpair (λ_j, ϕ_j) corresponding to covariance function (3.59) are given explicitly in Section 2.5.1. Moreover, all parameters used in the simulations are described in Table 3.2.

Table 3.2: Descriptions of the parameters used in the simulations.

Parameter	Description
N_d	degree of freedoms for the spatial discretization
N	truncation number in KL expansion
Q	highest order of basis polynomials for the stochastic domain
ν	viscosity parameter
ℓ	correlation length
κ_z	standard deviation
ϵ_{tol}	stopping tolerance of iterative methods
ϵ_{trunc}	truncation tolerance of the low–rank approximation
$\#iter_{\max}$	maximum iteration number of iterative methods

3.5.1 Stationary Problem with Random Diffusion Parameter

As a first benchmark problem, a two-dimensional stationary convection diffusion equation with random diffusion parameter [109] defined on $\mathcal{D} = [-1, 1]^2$ with the deterministic source function $f(\boldsymbol{x}) = 0$, the constant convection parameter $\mathbf{b}(\boldsymbol{x}) = (0, 1)^T$, and the Dirichlet boundary condition

$$y_{DB}(\boldsymbol{x}) = \begin{cases} y_{DB}(x_1, -1) = x_1, & y_{DB}(x_1, 1) = 0, \\ y_{DB}(-1, x_2) = -1, & y_{DB}(1, x_2) = 1, \end{cases}$$

is considered.

The random diffusion parameter is defined by $a(\boldsymbol{x}, \omega) = \nu\, z(\boldsymbol{x}, \omega)$, where the random input $z(\boldsymbol{x}, \omega)$ has the unity mean, i.e., $\overline{z}(\boldsymbol{x}) = 1$ and ν is the viscosity parameter. Around $x_2 = 1$, when the value of the solution alters dramatically, the solution displays an exponential boundary layer. As an alternative to traditional finite element techniques, discontinuous Galerkin discretization in the spatial domain may thus be preferable; the mean and variance of solutions for various values of the viscosity parameter ν are shown in Figure 3.1. As ν decreases, the boundary layer becomes more visible.

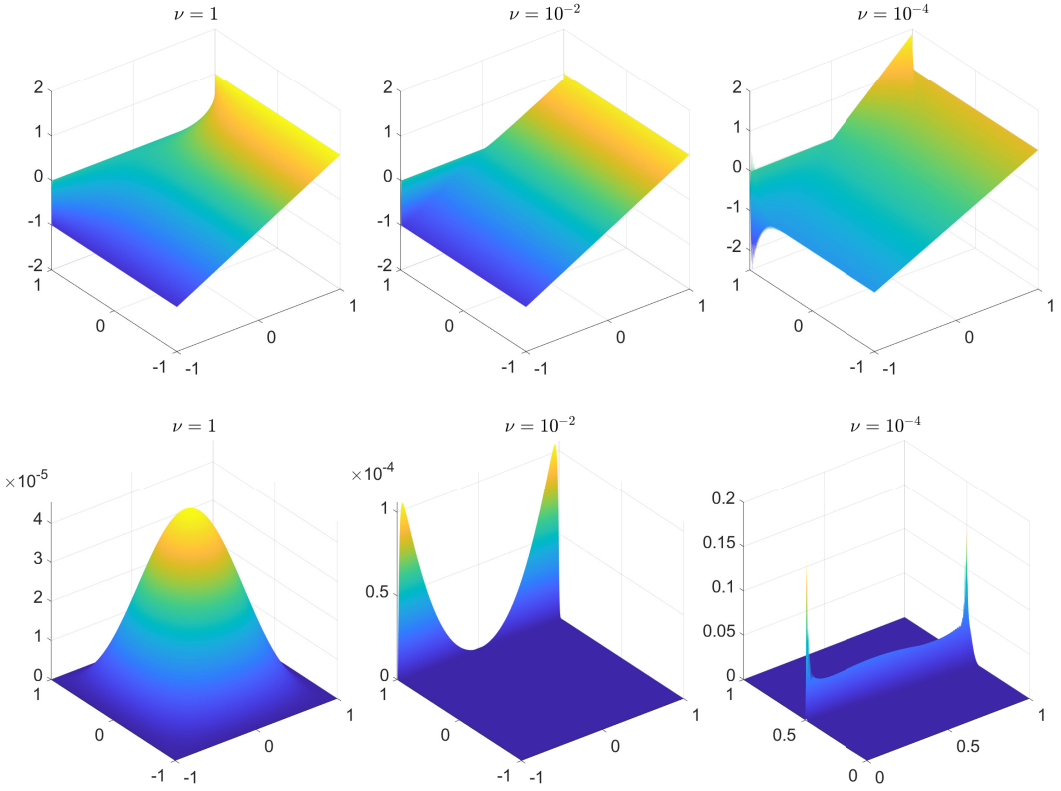

Figure 3.1: Example 3.5.1: Mean (top) and variance (bottom) of SG solutions obtained solving by $\mathcal{A}\backslash\mathcal{F}$ with $\ell = 1$, $\kappa_z = 0.05$, $N_d = 393216$, $N = 3$, and $Q = 2$ for various values of viscosity parameter ν.

In Table 3.3, 3.4, and 3.5, the results of the simulations are reported by taking into account numerous data sets. Table 3.3 shows the results for changing the truncation number in the KL expansion N while keeping all other parameters fixed. It is obvious that the complexity of the problem increases as N increases. As expected, decreasing the truncation tolerance ϵ_{trunc} increases the cost of computational time and memory requirement, especially for large N. Another important finding from the Table 3.3

is that LRPGMRES performs better in terms of CPU time and memory requirement when compared to other iterative methods. Table 3.4 shows how low–rank Krylov subspace algorithms with the mean–based preconditioner $\mathcal{P}_0$ perform for different viscosity values ν . The problem becomes increasingly convection dominated as the values of ν decrease. As a result, all iterative solvers use more memory, and the rank of the low-rank solution increases.

Afterward, the effect of the standard deviation parameter κ_z on the numerical simulations is investigated. Figure 3.2 displays the convergence behaviour of the low–rank variants of iterative solvers for varying values of ν. For relatively large κ_z, one can observe that LRPBiCGstab and LRPGMRES provide better convergence behaviour, although the LRPCG method does not converge since the dominance of the nonsymmetric part of $\mathcal{A}$ increases.

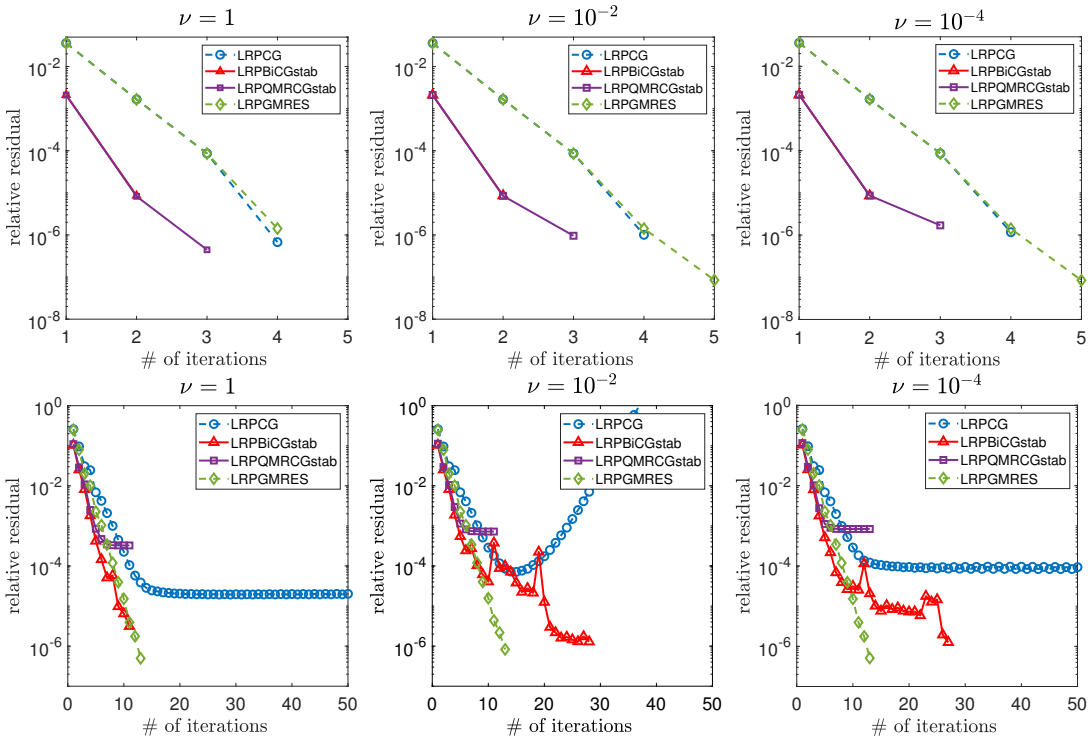

Figure 3.2: Example 3.5.1: Convergence of low–rank variants of iterative solvers with $\kappa_z = 0.05$ (top) and $\kappa_z = 0.5$ (bottom) for varying values of viscosity ν. The mean-based preconditioner $\mathcal{P}_0$ is used with the parameters $N = 5$, $Q = 3$, $\ell = 1$, $N_d = 6144$, and $\epsilon_{trunc} = 10^{-6}$.

In Table 3.5, the effect of the standard deviation parameter κ_z is examined with $\mathcal{P}_0$ and $\mathcal{P}_1$ preconditioners for only LRPBiCGstab and LRPGMRES since they exhibit

Table 3.3: Example 3.5.1: Simulation results showing ranks of truncated solutions, total number of iterations, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $\nu = 10^{-4}$, and the mean-based preconditioner $\mathcal{P}_0$ for varying values of N.

Method	LRPCG	LRPBiCGstab	LRPQMRCGstab	LRPGMRES
ϵ_{trunc}	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)
N=3				
Ranks	10 (10)	10 (10)	9 (10)	10 (10)
#iter	5 (5)	3 (3)	3 (3)	4 (4)
CPU	7.0 (7.7)	8.8 (8.9)	7.8 (10.0)	5.4 (5.2)
Resi.	1.8160e-07 (3.2499e-07)	5.5982e-06 (5.5413e-06)	2.9222e-05 (3.1021e-05)	6.3509e-07 (6.3509e-07)
Memory	481.6 (481.6)	481.6 (481.6)	433.4 (481.6)	481.6 (481.6)
N=4				
Ranks	12 (18)	17 (18)	17 (17)	17 (17)
#iter	4 (5)	3 (3)	3 (3)	5 (5)
CPU	9.2 (13.3)	15.3 (15.3)	14.2 (14.4)	12.1 (11.7)
Resi.	1.2367e-06 (1.4311e-07)	7.7090e-06 (7.7030e-06)	1.0819e-05 (4.3167e-06)	8.1316e-08 (8.1316e-08)
Memory	579.3 (868.9)	820.7 (868.9)	820.7 (820.7)	820.7 (820.7)
N=5				
Ranks	18 (28)	21 (28)	22 (28)	19 (28)
#iter	4 (5)	3 (3)	3 (3)	5 (5)
CPU	15.2 (20.8)	24.9 (25.1)	25.4 (26.0)	20.3 (20.6)
Resi.	1.1705e-06 (8.2045e-08)	8.5525e-06 (8.5527e-06)	1.6985e-06 (8.6182e-07)	8.4680e-08 (8.4680e-08)
Memory	871.9 (1356.3)	1017.2 (1356.3)	1065.6 (1356.3)	920.3 (1356.3)
N=6				
Ranks	26 (42)	26 (42)	25 (42)	25 (42)
#iter	4 (4)	3 (3)	3 (3)	4 (4)
CPU	25.6 (32.0)	42.4 (43.7)	49.4 (51.2)	29.7 (31.5)
Resi.	9.2495e-07 (1.0605e-06)	9.6694e-06 (9.6649e-06)	7.7812e-07 (4.1770e-07)	1.0476e-06 (1.0476e-06)
Memory	1265.1 (2043.6)	1265.1 (2043.6)	1216.4 (2043.6)	1216.4 (2043.6)
N=7				
Ranks	30 (60)	32 (60)	32 (60)	28 (47)
#iter	4 (4)	3 (3)	3 (3)	4 (4)
CPU	52.5 (58.8)	69.3 (73.8)	86.1 (87.9)	57.9 (57.7)
Resi.	1.0719e-06 (1.1205e-06)	9.9865e-06 (9.9880e-06)	6.5595e-07 (2.0226e-07)	1.1075e-06 (1.1075e-06)
Memory	1468.1 (2936.3)	1566 (2936.3)	1566 (2936.3)	1370.3 (2300.1)

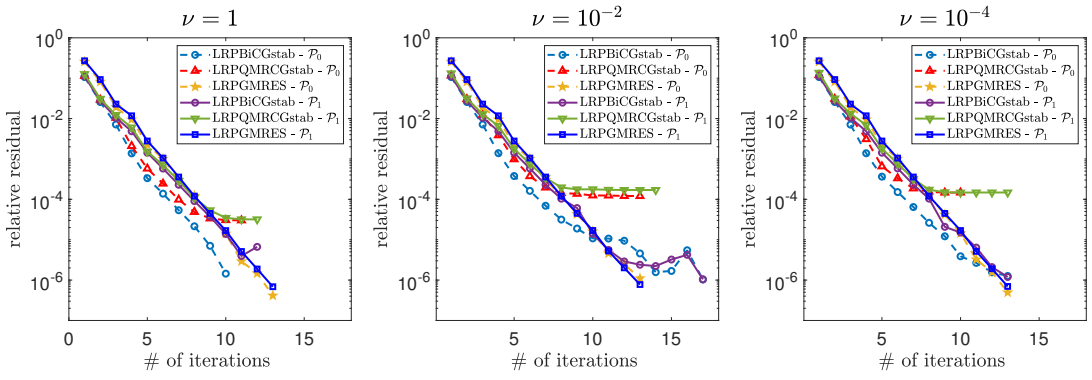

Figure 3.3: Example 3.5.1: Convergence of low–rank variants of LRPBiCGstab, LRPQMRCGstab, and LRPGMRES with $N = 7$, $Q = 3$, $\ell = 1$, $N_d = 6144$, $\epsilon_{trunc} = 10^{-8}$, and $\kappa_z = 0.5$ for the mean-based preconditioner $\mathcal{P}_0$ and the Ullmann preconditioner $\mathcal{P}_1$.

Table 3.4: Example 3.5.1: Simulation results showing ranks of truncated solutions, total number of iterations, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $N = 7$, and the mean-based preconditioner $\mathcal{P}_0$ for various values of viscosity parameter ν.

Method	LRPCG	LRPBiCGstab	LRPQMRCGstab	LRPGMRES
ϵ_{trunc}	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)
$\nu = 1$				
Ranks	17 (44)	20 (51)	20 (42)	22 (39)
#iter	4 (4)	3 (3)	3 (3)	4 (4)
CPU	54.3 (62.0)	68.7 (75.2)	87.9 (91.2)	56.4 (56.3)
Resi.	8.2189e-07 (1.1215e-06)	9.9896e-06 (9.9897e-06)	7.3503e-07 (3.6458e-08)	1.1062e-06 (1.1062e-06)
Memory	831.9 (2300.1)	978.8 (2495.8)	978.8 (2055.4)	1076.6 (1908.6)
$\nu = 10^{-2}$				
Ranks	21 (60)	26 (60)	25 (59)	23 (39)
#iter	4 (4)	3 (3)	3 (3)	4 (4)
CPU	52.4 (64.5)	65.9 (72.3)	89.5 (94.3)	52.3 (52.6)
Resi.	7.7284e-07 (1.1225e-06)	9.9906e-06 (9.9918e-06)	1.6268e-06 (8.4171e-08)	1.1074e-06 (1.1074e-06)
Memory	1027.7 (2936.3)	1272.4 (2936.3)	1223.4 (2887.3)	1125.6 (1908.6)
$\nu = 10^{-4}$				
Ranks	30 (60)	32 (60)	32 (60)	28 (47)
#iter	4 (4)	3 (3)	3 (3)	4 (4)
CPU	52.5 (58.8)	69.3 (73.8)	86.1 (87.9)	57.9 (57.7)
Resi.	1.0719e-06 (1.1205e-06)	9.9865e-06 (9.9880e-06)	6.5595e-07 (2.0226e-07)	1.1075e-06 (1.1075e-06)
Memory	1468.1 (2936.3)	1566 (2936.3)	1566 (2936.3)	1370.3 (2300.1)

Table 3.5: Example 3.5.1: Simulation results showing ranks of truncated solutions, total number of iterations, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $N = 7$, $Q = 3$, $\ell = 1$, $\epsilon_{trunc} = 10^{-6}$, and $\nu = 10^{-4}$ for different choices of preconditioners.

Method	LRPBiCGstab	LRPGMRES	LRPBiCGstab	LRPGMRES
Preconditioner	$\mathcal{P}_0$	$\mathcal{P}_0$	$\mathcal{P}_1$	$\mathcal{P}_1$
$\kappa_z = 0.05$				
Ranks	32	28	31	27
#iter	3	4	3	5
CPU	69.3	57.9	69.0	72.7
Resi.	9.9865e-06	1.1075e-06	6.0448e-06	8.7712e-08
Memory	1566	1370.3	1517.1	1321.3
$\kappa_z = 0.5$				
Ranks	60	60	60	60
#iter	13	13	15	13
CPU	781.7	248.5	913.6	245.8
Resi.	1.2629e-06	4.9417e-07	1.8697e-06	6.9625e-07
Memory	2936.3	2936.3	2936.3	2936.3

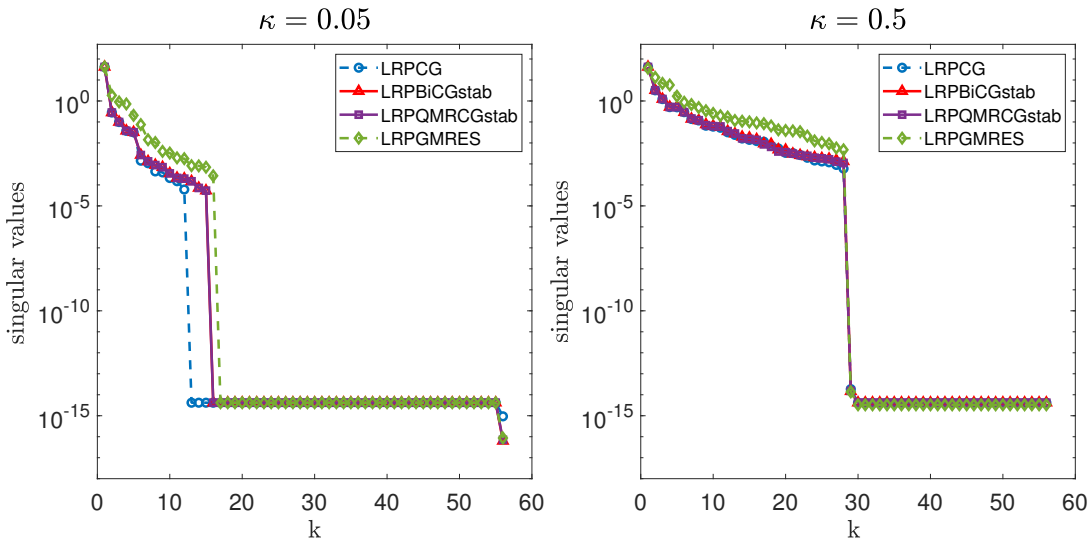

Figure 3.4: Example 3.5.1: Decay of singular values of low–rank solution matrix $\mathbf{Y}$ obtained by using the mean-based preconditioner $\mathcal{P}_0$ with $N = 5$, $Q = 3$, $\ell = 1$, $N_d = 6144$, $\nu = 1$, and $\epsilon_{trunc} = 10^{-6}$ for $\kappa_z = 0.05$ (left) and $\kappa_z = 0.5$ (right).

better convergence behaviour; see Figure 3.2. As κ_z increases, the low-rank solutions indicate deteriorating performance, regardless of which the preconditioner or iterative solver is used. Also, the effect of preconditioners on the iterative solvers is shown in Figure 3.3 in terms of convergence of iterative solvers. Since LRPCG does not converge for large values of κ_z, they are not included. The results show that the mean–based preconditioner $\mathcal{P}_0$ exhibits better convergence behaviour compared to the Ullmann preconditioner $\mathcal{P}_1$ for LRPBiCGstab and LRPQMRCGstab, whereas they are almost the same for LRPGMRES.

Figure 3.4 shows the decay of singular values of low–rank solution matrix $\mathbf{Y}$ obtained by using the mean-based preconditioner $\mathcal{P}_0$. Keeping other parameters fixed, increasing the value of κ_z slows down the decay of the singular values of the obtained solutions. Thus, the total time for solving the system and the time spent on truncation will also increase; see Table 3.5.

Last, the performance of $\mathcal{A}\backslash\mathcal{F}$ is displayed in terms of total CPU times (in seconds) and memory requirements (in KB) in Table 3.6. Due to "out of memory" termination, which has been marked as "OoM," some numerical results are not provided. A significant finding from numerical simulations is that low–rank variants of Krylov subspace algorithms provide better computational savings, particularly in terms of memory.

Table 3.6: Example 3.5.1: Total CPU times (in seconds) and memory (in KB) for $N_d = 6144$, $Q = 3$, $\ell = 1$, and $\kappa_z = 0.05$.

$\mathcal{A}\backslash\mathcal{F}$	$\nu = 10^0$	$\nu = 10^{-2}$	$\nu = 10^{-4}$
N	CPU (Memory)	CPU (Memory)	CPU (Memory)
2	10.8 (960)	10.7 (960)	10.8 (960)
3	1463.7 (1920)	1464.2 (1920)	1463.7 (1920)
4	OoM	OoM	OoM

3.5.2 Stationary Problem with Random Convection Parameter

Our second example is a two-dimensional stationary convection diffusion equation with random velocity. To be more specific, the choice is the deterministic diffusion parameter $a(\boldsymbol{x}, \omega) = \nu > 0$, the deterministic source function $f(\boldsymbol{x}) = 0$, and the spatial domain $\mathcal{D} = [0, 1]^2$. The random velocity field $\mathbf{b}(\boldsymbol{x}, \omega)$ is

$$\mathbf{b}(\boldsymbol{x}, \omega) := \left(\cos\left(\frac{1}{5}z(\boldsymbol{x}, \omega)\right), \sin\left(\frac{1}{5}z(\boldsymbol{x}, \omega)\right)\right)^T, \tag{3.60}$$

where the mean of the random field is $\overline{z}(\boldsymbol{x}) = 0$. The Dirichlet boundary condition $y_{DB}(\boldsymbol{x})$ is given by

$$y_{DB}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in S, \\ 0, & \boldsymbol{x} \in \partial\mathcal{D}\backslash S, \end{cases}$$

where the set S is the subset of $\partial\mathcal{D}$ defined by

$$\{x_1 = 0, x_2 \in [0, 0.5]\} \cup \{x_1 \in [0, 1], x_2 = 0\} \cup \{x_1 = 1, x_2 \in [0, 0.5]\}.$$

The solution features sharp transitions in the domain $\mathcal{D}$ due to the random velocity, $\mathbf{b}(\boldsymbol{x}, \omega)$, and later spurious oscillations will spread into the stochastic domain Ω. As ν decreases, the interior layer becomes more visible; see Figure 3.5 for the mean and variance of solution for various values of ν.

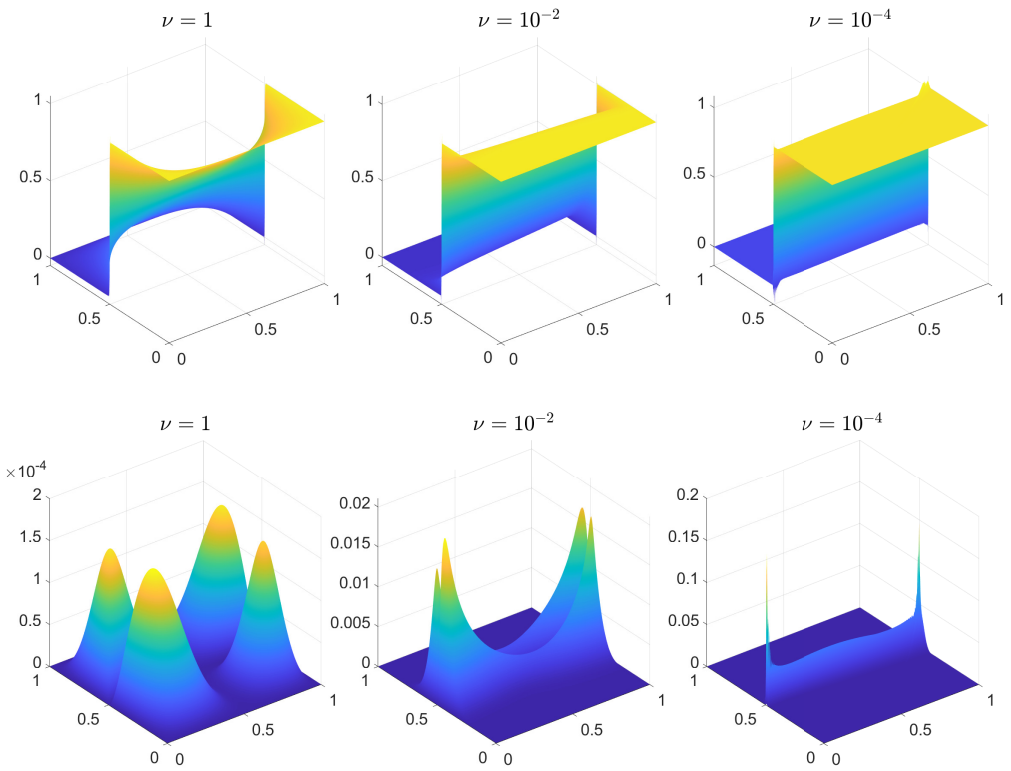

Figure 3.5: Example 3.5.2: Mean (top) and variance (bottom) of SG solutions obtained by solving $\mathcal{A}\backslash\mathcal{F}$ with $N = 2$, $Q = 2$, $\ell = 1$, $N_d = 393216$, and $\kappa_z = 0.05$, for various values of ν.

Table 3.7: Example 3.5.2: Simulation results showing ranks of truncated solutions, total number of iterations, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $N = 7$, and the mean-based preconditioner $\mathcal{P}_0$ for various values of viscosity parameter ν.

Method	LRPCG	LRPBiCGstab	LRPQMRCGstab	LRPGMRES
ϵ_{trunc}	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)
$\nu = 1$				
Ranks	8 (19)	10 (23)	9 (22)	6 (8)
#iter	10 (10)	3 (3)	4 (4)	5 (5)
CPU	108.6 (120.4)	49.7 (56.0)	76.1 (82.4)	60.7 (59.7)
Resi.	1.3666e-06 (1.4307e-06)	5.7868e-07 (5.7870e-07)	6.8606e-06 (6.8424e-06)	1.2811e-06 (1.2811e-06)
Memory	391.5 (929.8)	489.4 (1125.6)	440.4 (1076.6)	293.6 (391.5)
$\nu = 10^{-2}$				
Ranks	15 (34)	45 (60)	21 (60)	6 (14)
#iter	100 (100)	100 (100)	100 (100)	100 (100)
CPU	992.2 (1202.8)	1782.7 (2133.9)	2602.3 (2815.0)	8798.7 (8864.7)
Resi.	3.0059e+26 (3.0060e+26)	1.4807e-01 (2.6669e-02)	9.0173e-03 (9.3659e-03)	1.0002e-03 (1.0002e-03)
Memory	734.1 (1663.9)	2202.2 (2936.3)	1027.7 (2936.3)	293.6 (685.1)
$\nu = 10^{-4}$				
Ranks	29 (60)	60 (60)	35 (60)	12 (27)
#iter	100 (100)	100 (100)	100 (100)	100 (100)
CPU	1102.3 (1382.6)	2124.4 (2135.5)	2802.1 (2869.9)	8401.0 (8394.8)
Resi.	8.7243e+26 (8.7242e+26)	2.2085e-02 (3.2792e-02)	3.6713e-03 (3.3074e-03)	1.2075e-03 (1.2075e-03)
Memory	1419.2 (2936.3)	2936.3 (2936.3)	1712.8 (2936.3)	587.3 (1321.3)

Table 3.8: Example 3.5.2: Simulation results showing ranks of truncated solutions, total number of iterations, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N = 7$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $\nu = 10^{-4}$, and the mean-based preconditioner $\mathcal{P}_0$ for various values of N_d.

	LRPCG 1e-06 (1e-08)	LRPBiCGstab 1e-06 (1e-08)	LRPQMRCGstab 1e-06 (1e-08)	LRPGMRES 1e-06 (1e-08)
$N_d = 384$				
Ranks	17 (37)	58 (60)	21 (60)	26 (42)
#iter	100 (100)	100 (100)	100 (100)	100 (100)
CPU	213.3 (205.7)	329.8 (329.4)	487.6 (481.7)	2119.7 (2135.3)
Resi.	1.1637e+28 (1.1637e+28)	1.5964e-02 (3.3908e-01)	1.0163e-02 (1.0857e-02)	7.4718e-06 (7.2284e-06)
Memory	66.9 (145.7)	228.4 (236.3)	82.7 (236.3)	102.4 (165.4)
$N_d = 1536$				
Ranks	25 (56)	60 (60)	21 (55)	30 (49)
#iter	100 (100)	100 (100)	100 (100)	65 (65)
CPU	305.4 (338.8)	497.8 (503.0)	699.4 (709.0)	1278.3 (1286.1)
Resi.	2.9340e+27 (2.9339e+27)	5.3887e-02 (1.9615e-02)	6.8316e-03 (7.0652e-03)	1.8606e-06 (1.8606e-06)
Memory	323.4 (724.5)	776.3 (776.3)	271.7 (711.6)	388.1 (633.9)
$N_d = 6144$				
Ranks	29 (60)	60 (60)	35 (60)	12 (27)
#iter	100 (100)	100 (100)	100 (100)	100 (100)
CPU	1102.3 (1382.6)	2124.4 (2135.5)	2802.1 (2869.9)	8401.0 (8394.8)
Resi.	8.7243e+26 (8.7242e+26)	2.2085e-02 (3.2792e-02)	3.6713e-03 (3.3074e-03)	1.2075e-03 (1.2075e-03)
Memory	1419.2 (2936.3)	2936.3 (2936.3)	1712.8 (2936.3)	587.3 (1321.3)
$N_d = 24576$				
Ranks	25 (60)	60 (60)	23 (60)	7 (19)
#iter	100 (100)	100 (100)	100 (100)	100 (100)
CPU	7276.2 (10498.6)	16726.6 (16936.7)	19960.3 (20646.5)	41929.8 (41778.9)
Resi.	3.5040e+26 (3.5100e+26)	5.8952e-03 (1.7120e-02)	1.7710e-03 (1.6015e-03)	7.3550e-04 (7.3550e-04)
Memory	4823.4 (11576.3)	11576.3 (11576.3)	4437.6 (11576.2)	1350.6 (3665.8)

Next, Table 3.7 and 3.8 display the performance of low–rank of Krylov subspace methods with the mean–based precondition $\mathcal{P}_0$ by taking into account numerous data sets. It is obvious that the complexity of the problem increases in terms of the rank of the truncated solutions, total CPU times (in seconds), and memory demand of the solution (in KB) when ν decreases; see Table 3.7. As seen in the previous example, the LRPCG approach performs poorly for smaller values of ν, however the LRPGMRES method performs better.

Then, as shown in Figure 3.6, the convergence behaviour of low-rank versions of iterative solvers with different values of ν is explored. The relative residuals obtained by LRPQMRCGstab and LRPGMRES decline monotonically, whereas the LRPBiCGstab method exhibits oscillatory behaviour.

Figure 3.7 shows the decay of singular values of low–rank solution matrix $\mathbf{Y}$ obtained

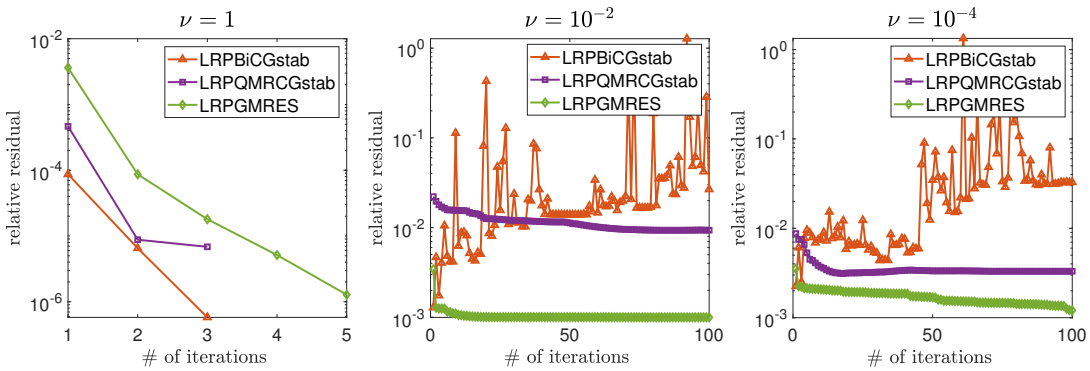

Figure 3.6: Example 3.5.2: Convergence of low–rank variants of iterative solvers for varying values of viscosity ν. The mean-based preconditioner $\mathcal{P}_0$ is used with the parameters $N = 7$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $N_d = 6144$, and $\epsilon_{trunc} = 10^{-8}$.

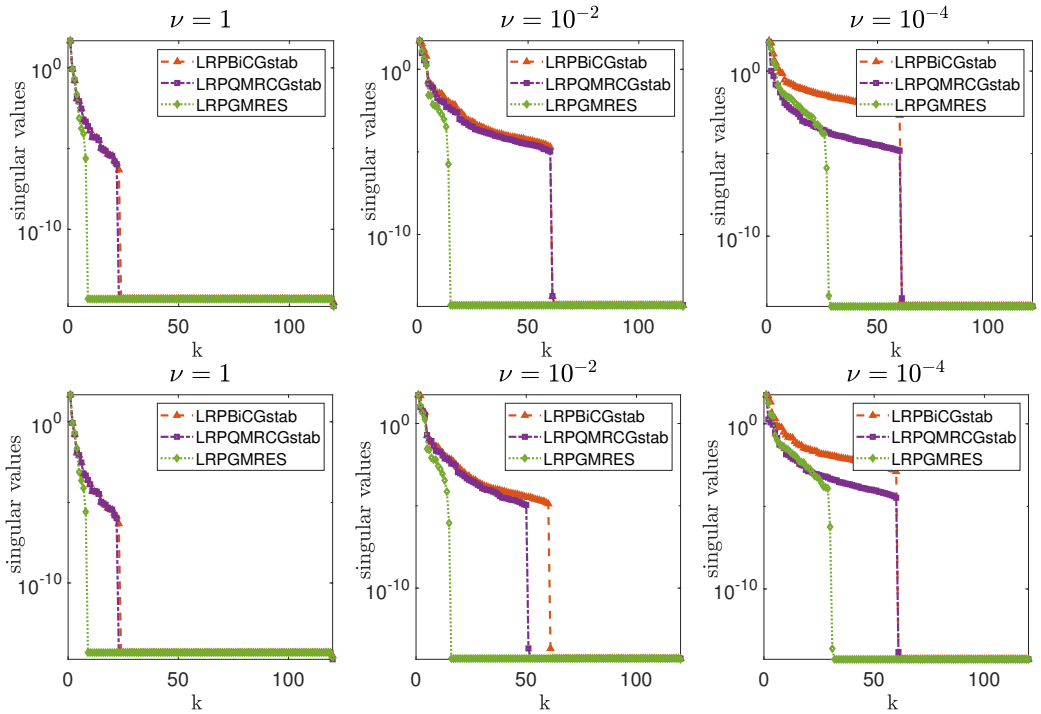

Figure 3.7: Example 3.5.2: Decay of singular values of solution matrix $\mathbf{Y}$ with $N = 7$, $Q = 3$, $\ell = 1$, $N_d = 6144$, $\kappa_z = 0.05$, and $\epsilon_{trunc} = 10^{-8}$ with the mean-based preconditioner $\mathcal{P}_0$ (top) and the Ullmann preconditioner $\mathcal{P}_1$ (bottom) for various values of ν.

by using $\mathcal{P}_0$ and $\mathcal{P}_1$ preconditioners. When the other parameters remain constant, reducing the value of ν decreases the decay of the singular values of the resulting solutions. As a result, the overall time required to solve the system and the time spent on truncation will rise; see Table 3.7.

Large–scale simulations with a high degree of freedoms (DOFs) are generally more interesting in practical applications. Another experiment in Table 3.9 is the memory requirement of the solution (in KB) produced from full–rank and low–rank variations of the GMRES solver. Low–rank approximation, as predicted, considerably decreases the amount of computer memory required to solve the huge system.

Table 3.9: Example 3.5.2: Memory demand of the solution (in KB) obtained full–rank and low–rank variants of GMRES solver with $N = 7$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $\epsilon_{trunc} = 10^{-6}$ ($\epsilon_{trunc} = 10^{-8}$), and the mean-based preconditioner $\mathcal{P}_0$ for various values of degree of freedoms (DOFs).

DOFs	46080	184320	737280	2949120
Low–Rank	94.5 (157.5)	323.4 (556.3)	587.3 (1468.1)	1543.5 (3665.8)
Full–Rank	360	1440	5760	23040

3.5.3 Unsteady Problem with Random Diffusion Parameter

Last, an unsteady convection diffusion equation with random diffusion parameter defined on $\mathcal{D} = [0, 1]^2$ is considered. The rest of data is as follows

$$T = 0.5, \quad \mathbf{b}(\boldsymbol{x}) = (1, 1)^T, \quad f(\boldsymbol{x}, t) = 0, \quad y^0(\boldsymbol{x}) = 0$$

with the Dirichlet boundary condition

$$y_{DB}(\boldsymbol{x}) = \begin{cases} y_{DB}(0, x_2) = x_2(1 - x_2), & y_{DB}(1, x_2) = 0, \\ y_{DB}(x_1, 0) = 0, & y_{DB}(x_1, 1) = 0. \end{cases}$$

In this example, the random diffusion coefficient $a(\boldsymbol{x}, w)$ is chosen as $a(\boldsymbol{x}, w) = z(\boldsymbol{x}, w)$ having the unity mean. In the numerical simulations, the number of time points is chosen as $N_T = 32$. As mentioned in Section 2.5.1, it is known that decreasing the correlation length slows down the decay of the eigenvalues in the KL expansion of the random variable $a(\boldsymbol{x}, \omega)$ and therefore, more random variables are

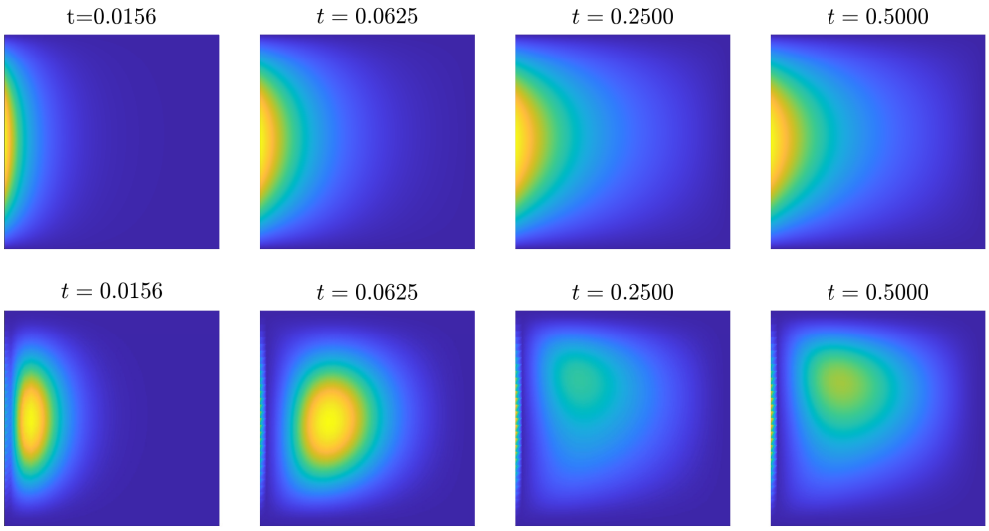

Figure 3.8: Example 3.5.3: Mean and variance of computed solution at various time steps obtained by LRPBiCGstab with $N = 17$, $Q = 3$, $\ell = 1.5$, $\kappa_z = 0.15$, $\epsilon_{trunc} = 10^{-6}$, and the mean–based preconditioner $\mathcal{P}_0$.

required to capture the randomness sufficiently; see, e.g., [117]. In other words, it causes the truncation parameter N to increase: When the correlation length is increased, the situation is the reverse. Consequently, the main focus for this benchmark problem is the effect of correlation length on the low–rank variants of the iterative solver. With the help of the following computation as done in [63],

$$\left(\sum_{i=1}^{N} \lambda_i\right) / \left(\sum_{i=1}^{M_\ell} \lambda_i\right) > 0.97,$$

suitable truncation number N can be chosen for the given correlation length ℓ. Here, M_ℓ is a large number which is set 1000. Computed mean and variance of the solution are displayed in Figure 3.8 for various time steps.

In Table 3.10, the results of numerical simulations for the mean-based preconditioner $\mathcal{P}_0$ is presented for varying values of the correlation length ℓ. No matter whether an iterative solver is employed, the small correlation length raises the rank of obtained low–rank solutions and the number of iterations as long as 97% of the total variance is captured. As expected, decreasing the truncation tolerance ϵ_{trunc} increases the cost of comparatively more computational time and memory requirements, nonetheless it does not affect the relative residuals. Next, Table 3.11 displays numerical results obtained by using the standard Krylov subspace iterative solvers with the mean-based preconditioner $\mathcal{P}_0$. Compared to the full–rank solvers in Ta-

Table 3.10: Example 3.5.3: Simulation results showing ranks of truncated solutions, total number of iterations, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $Q = 3$, $\kappa_z = 0.15$, and the mean–based preconditioner $\mathcal{P}_0$ for various values of correlation length ℓ at final time $T = 0.5$.

Method	LRPCG	LRPBiCGstab	LRPQMRCGstab	LRPGMRES
ϵ_{trunc}	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)	1e-06 (1e-08)
$\ell = 3, N = 9$				
Ranks	25 (56)	25 (55)	22 (52)	25 (38)
#iter	4 (4)	3 (3)	4 (4)	4 (4)
CPU	5348.9 (6543.5)	5072.2 (6869.4)	9459.5 (11060.9)	5974.5 (6005.6)
Resi.	2.7590e-04 (2.7601e-04)	1.2268e-03 (1.2268e-03)	8.5901e-05 (8.5657e-05)	2.3936e-04 (2.3936e-04)
Memory	1243 (2784.3)	1243 (2734.5)	1093.8 (2585.4)	1243 (1889.3)
$\ell = 2.5, N = 10$				
Ranks	27 (61)	27 (60)	25 (55)	27 (43)
#iter	4 (4)	3 (3)	4 (4)	4 (4)
CPU	7564.4 (9310.7)	7030.5 (9558.5)	13158.6 (15636.3)	9219.1 (9158.8)
Resi.	2.7397e-04 (2.7405e-04)	1.2665e-03 (1.2665e-03)	9.5085e-05 (9.4831e-05)	2.3810e-04 (2.3810e-04)
Memory	1356.3 (3064.3)	1356.3 (3014.1)	1255.7 (2762.9)	1356.3 (2160.1)
$\ell = 2, N = 13$				
Ranks	28 (68)	29 (66)	27 (63)	32 (52)
#iter	4 (4)	3 (3)	4 (4)	4 (4)
CPU	10813.5 (15435.4)	10745.0 (15709.9)	18748.2 (24049.6)	18496.9 (18284.7)
Resi.	2.6686e-04 (2.6703e-04)	1.2994e-03 (1.2994e-03)	1.0372e-04 (1.0345e-04)	2.4580e-04 (2.4580e-04)
Memory	1466.5 (3561.5)	1518.9 (3456.8)	1414.1 (3299.6)	1676 (2723.5)
$\ell = 1.5, N = 17$				
Ranks	32 (78)	33 (77)	31 (73)	38 (62)
#iter	4 (4)	3 (3)	4 (4)	4 (4)
CPU	20658.3 (36422.4)	22889.2 (33851.9)	36876.4 (50963.2)	58974.4 (57828.0)
Resi.	2.3545e-04 (2.3548e-04)	1.3217e-03 (1.3217e-03)	1.0444e-04 (1.0425e-04)	2.9531e-04 (2.9531e-04)
Memory	1821 (4438.7)	1877.9 (4381.8)	1764.1 (4154.2)	2162.4 (3528.2)

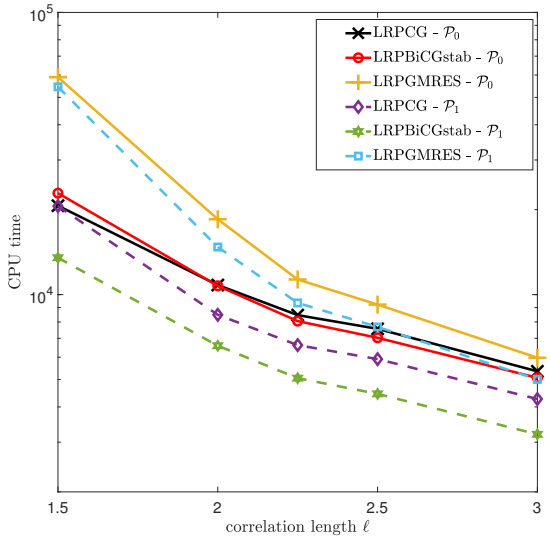

Figure 3.9: Example 3.5.3: CPU times of LRPCG, LRPBiCGstab, and LRPGMRES iterative solvers obtained by the the mean–based preconditioner $\mathcal{P}_0$ and the Ullmann preconditioner $\mathcal{P}_1$ with $Q = 3$, $N_d = 6144$, and $\kappa_z = 0.15$ for various values of correlation length ℓ.

Table 3.11: Example 3.5.3: Simulation results showing total number of iterations, total CPU times (in seconds), and memory demand of the full–rank solution (in KB) with $N_d = 6144$, $Q = 3$, $\kappa_z = 0.15$, and the mean-based preconditioner $\mathcal{P}_0$ for various values of correlation length ℓ at final time $T = 0.5$.

Method	PCG	PBiCGstab	PGMRES
ϵ_{tol}	1e-04	1e-04	1e-04
$\ell = 3$, $N = 9$			
#iter	11	5.5	10
CPU	7910.7	7863.4	8727.0
Memory	10560	10560	10560
$\ell = 2$, $N = 13$			
#iter	11	5.5	10
CPU	20252.0	20148.5	22318.8
Memory	26880	26880	26880
$\ell = 1.5$, $N = 17$			
#iter	11	5.5	10
CPU	41345.0	41207.4	45499.8
Memory	54720	54720	54720

ble 3.11, low–rank Krylov subspace solvers generally exhibit better performance; see Table 3.10. Regarding the preconditioners, the Ullmann preconditioner $\mathcal{P}_1$ produces better performance in terms of computational time; see Figure 3.9.

3.6 Discussion

In this chapter, the statistical moments of a convection diffusion equation having random coefficients have been numerically investigated. The original problem is converted into a system consisting of deterministic convection diffusion equations for each realization of the random coefficients by means of the stochastic Galerkin approach. Due to its local mass conservativity, the symmetric interior penalty Galerkin approach is then employed to discretize the deterministic problem. According to the literature, the analysis of stochastic discontinuous Galerkin techniques using convection diffusion equations has not been studied, so this study intends to fill this gap; see [41]. Moreover, the a priori error estimates in the energy norm is provided for the steady and unsteady models, while the stability analysis for the time-dependent models is discussed in the energy norm. To reduce computational time and memory

requirements, low–rank variants of various Krylov subspace methods, such as CG, BiCGstab, QMRCGstab, and GMRES with suitable preconditioners, have been employed. On the contrary to the studies in literature [20, 55, 64, 109, 110, 122], where randomness is generally defined in the diffusion parameter, the randomness is, here, considered both in diffusion or convection parameters. The numerical simulations have demonstrated that LRPGMRES performs better, especially for the convection-dominated models.

CHAPTER 4

ADAPTIVE STOCHASTIC DISCONTINUOUS GALERKIN FOR CONVECTION DIFFUSION EQUATIONS WITH RANDOM COEFFICIENTS

The aim of this chapter is the development, and in part the analysis, of adaptive stochastic discontinuous Galerkin method, which consists of successive loops of the following sequence [138]:

$$\textbf{SOLVE} \rightarrow \textbf{ESTIMATE} \rightarrow \textbf{MARK} \rightarrow \textbf{REFINE} \tag{4.1}$$

for convection diffusion equations containing random coefficients. The **SOLVE** step stands for the numerical solution of the underlying PDE in a finite dimensional tensor product space defined on the given spatial mesh and stochastic domain. The **ESTIMATE** step is the key point of the adaptive stochastic discontinuous Galerkin method. In this step, in order to control the error in the Galerkin solution, spatial estimators and parametric estimators are computed in terms of the discrete solutions without knowledge of the exact solutions. Based on the information of the indicators, the **MARK** step selects a subset of elements subject to each refinement. At each iteration, the **REFINE** module either performs a local refinement of the current triangulation in the spatial domain or enriches the current index set of the stochastic domain; see, e.g., [29, 60].

The efficient numerical solution of the convection diffusion equations with uncertainty discussed in Chapter 3 presents a number of theoretical and practical challenges. One challenging class of problems is that in the convection dominated problems, some layers and oscillations occur, so these cause difficulties to compute nu-

merical solutions. Another particularly challenging class of problems is represented by PDEs whose inputs and outputs depend on infinitely many uncertain parameters. For this class of problems, numerical algorithms are sought that are able to identify a finite set of most important parameters to be incorporated into the basis of the approximation space. These difficulties motivates that the combination of discontinuous Galerkin adaptivity in the spatial domain and the adaptivity in the (stochastic) parameter domain for the solution of convection dominated PDEs with random inputs.

In this chapter, we begin with our investigation in the next section by introducing the model problem. In Section 4.2, the numerical schemes, which are Karhunen–Loève (KL) expansion, stochastic Galerkin method, and symmetric interior penalty Galerkin method (SIPG), are briefly recalled by referring the Chapter 3. A residual–based error estimator is developed in Section 4.3. Section 4.4 represents the adaptive algorithm and reports the results of numerical experiments. Finally, Section 4.5 concludes this chapter by giving some conclusions and discussions.

4.1 Problem Formulation

In this chapter, we study a convection diffusion equation with random coefficients: find a random function $y : \overline{\mathcal{D}} \times \Omega \to \mathbb{R}$ such that $\mathbb{P}$-almost surely in Ω

$$-\nabla \cdot (a(\boldsymbol{x},\omega)\nabla y(\boldsymbol{x},\omega)) + \mathbf{b}(\boldsymbol{x},\omega)\cdot\nabla y(\boldsymbol{x},\omega) = f(\boldsymbol{x}) \qquad \text{in } \mathcal{D}\times\Omega, \tag{4.2a}$$

$$y(\boldsymbol{x},\omega) = y^{DB}(\boldsymbol{x}) \quad \text{on } \partial\mathcal{D}\times\Omega, \tag{4.2b}$$

where $a : (\mathcal{D}\times\Omega) \to \mathbb{R}$ and $\mathbf{b} : (\mathcal{D}\times\Omega) \to \mathbb{R}^2$ are random diffusivity and velocity coefficients. As done in Chapter 3, source function and Dirichlet boundary condition denoted by $f(\boldsymbol{x}) \in L^2(\mathcal{D})$ and $y^{DB}(\boldsymbol{x}) \in H^{1/2}(\partial\mathcal{D})$, respectively, are given in a deterministic way. The well–posedness of the model problem (4.2) can be shown by proceeding the classical Lax–Milgram lemma, see, e.g., [11], under the assumptions given in Section 3.1.

4.2 Stochastic Galerkin Discretization

After representing the given random coefficients $a(\boldsymbol{x}, \omega)$ and $b(\boldsymbol{x}, \omega)$ by Karhunen–Loève (KL) expansion (2.16), we seek for the solution $y(\boldsymbol{x}, \omega) = y(\boldsymbol{x}, \xi_1(\omega), \xi_2(\omega), \ldots, \xi_N(\omega))$ on the probability space $(\Gamma, \mathcal{B}(\Gamma), \rho(\xi)d\xi)$, where $\Gamma = \prod_{n=1}^{N} \Gamma_n$ is the support of probability density in finite dimensional space, $\mathcal{B}(\Gamma)$ denotes Borel σ–algebra, and $\rho(\xi)d\xi$ is the distribution measure of the vector ξ. For any integers $n, m \in \mathbb{N}$, $\{\psi_m^n(\xi_n)\}_{m=0}^{\infty}$ denotes the set of univariate polynomials of degree m on Γ_n that are orthogonal in $L^2(\Gamma_n)$. By following the discussion in [25, 59], the set of finitely supported multi–indices (or called as index set) is

$$\mathfrak{U} = \{q = (q_1, q_2, \ldots) \in \mathbb{N} \ : \ |\operatorname{supp} q| < \infty\},$$

where $\operatorname{supp} q = \{n \in \mathbb{N} \mid q_n \neq 0\}$ and $|q| := \sum_{i \in \operatorname{supp} q} q_i$. Then, the countable tensor product polynomial $\Psi_q(\xi)$ is corresponding to a product of univariate orthogonal polynomials, i.e.,

$$\Psi_q(\xi) = \prod_{n \in \operatorname{supp} q} \psi_{q_n}^n(\xi_n),$$

for any $q \in \mathfrak{U}$ and $\xi \in \Gamma$. In this setting, for a given finite index set $\mathfrak{B} \subset \mathfrak{U}$, the (discontinuous) finite element approximation space $\mathcal{S}_k^q \subset L^2(\Gamma)$ having at most q_n degree on each direction ξ_n is denoted by

$$\mathcal{S}_k^q := \operatorname{span}\{\Psi_q : \ q \in \mathfrak{B}\}, \tag{4.3}$$

where its dimension is equivalent to the cardinality of set, that is, $\dim(\mathcal{S}_k^q) = \#\mathfrak{B}$. In addition, we set a subspace $\mathcal{S}_k^{q_n} \subset \mathcal{S}_k^q$ for $n = 1, \ldots, N$ consisting of the family of polynomials $\{\widetilde{\Psi}_q(\xi)\}_{q \in \mathfrak{B}}$ defined by

$$\widetilde{\Psi}_q(\xi) = \prod_{i=1}^{N} \widetilde{\psi}_{q_i}^i(\xi_i), \quad \text{where} \quad \widetilde{\psi}_{q_i}^i(\xi_i) = \begin{cases} \widetilde{\psi}_{q_n}^n(\xi_n), & i = n, \\ 1, & i \neq n. \end{cases}$$

Hence, the approximate solution of the model problem (4.2), $y(\boldsymbol{x}, \omega) \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, represented by a generalized polynomial chaos (gPC) approximation, see Section 2.5.2, is of the form

$$y(\boldsymbol{x}, \omega) \approx y_k(\boldsymbol{x}, \omega) = \sum_{q \in \mathfrak{B}} y_q(\boldsymbol{x}) \Psi_q(\xi(\omega)), \tag{4.4}$$

where $y_q(\boldsymbol{x})$, the deterministic modes of the expansion, are given by

$$y_q(\boldsymbol{x}) = \frac{\langle y(\boldsymbol{x},\omega)\Psi_q(\xi)\rangle}{\langle \Psi_q^2(\xi)\rangle}.$$

Last, we discretize the spatial domain $\mathcal{D}$ by using the SIPG discretization introduced in Section 3.1.1 so that the variational formulation of (4.2) obtained by the stochastic discontinuous Galerkin discretization is as follows: Find $y_h \in V_h \otimes \mathcal{S}_k^q$ such that

$$a_\xi(y_h, v) = l_\xi(v) \qquad \forall v \in V_h \otimes \mathcal{S}_k^q, \tag{4.5}$$

where

$$a_\xi(y,v) = \int_\Gamma a_h(y,v,\xi)\rho(\xi)\,d\xi, \quad l_\xi(v) = \int_\Gamma l_h(v,\xi)\rho(\xi)\,d\xi.$$

Here, the (bi)–linear forms $a_h(\cdot,\cdot,\cdot)$ and $l_h(\cdot,\cdot)$ for a finite dimensional vector ξ are corresponding to (3.9) and 3.10, respectively.

Due the jump terms $[\![\cdot]\!]$, the bilinear form $a_\xi(y,v)$ (4.5) is not well–defined for the functions $y, v \in H^1(\mathcal{D}) \otimes \mathcal{S}_k^q$. However, when the bilinear form $a_\xi(y,v)$ (4.5) is decomposed as follows [136]

$$a_\xi(y,v) = \widetilde{a}_\xi(y,v) + j_\xi(y,v) \quad \forall y, v \in V_h \otimes \mathcal{S}_k^q, \tag{4.6}$$

where

$$j_\xi(y,v) = -\int_\Gamma \Big(\sum_{E\in\mathcal{E}_h} \int_E \big(\{\!\{a(.,\xi)\nabla y\}\!\} [\![v]\!] + \{\!\{a(.,\xi)\nabla v\}\!\} [\![y]\!] \big)\, ds \Big) \rho(\xi)\,d\xi,$$

the bilinear form $\widetilde{a}_\xi(y,v)$ becomes well–defined for the functions $y, v \in H^1(\mathcal{D}) \otimes \mathcal{S}_k^q$ and the following relation holds

$$a_\xi(y,v) = \widetilde{a}_\xi(y,v), \qquad y, v \in H^1(\mathcal{D}) \otimes \mathcal{S}_k^q. \tag{4.7}$$

In the following, a residual–based error estimator in the energy norm, which separates the effects of stochastic Galerkin discretization in parametric spaces and of the SIPG discretization in physical space, will be derived.

4.3 Residual–Based Error Estimator

Throughout this section, the symbol $\lesssim$ is used to denote bounds that are valid up to positive constants independent of the local mesh sizes and the penalty parameter σ,

provided that $\sigma \geq 1$. To derive a reliable residual–based error estimator, we follow the energy norm defined in (3.12).

By following [47, 137], we introduce an L^2–projection operator $\Pi_q : L^2(\Gamma) \to \mathcal{S}_k^q$ by

$$(\Pi_q(\xi) - \xi, \zeta)_{L^2(\Gamma)} = 0 \qquad \forall \zeta \in \mathcal{S}_k^q, \qquad \forall \xi \in L^2(\Gamma), \tag{4.8a}$$

and an local L^2–projection operator $\Pi_{q_n} : L^2(\Gamma) \to \mathcal{S}_k^{q_n}$ for $n = 1, 2, \ldots, N$ by

$$(\Pi_{q_n}(\xi) - \xi, \zeta)_{L^2(\Gamma)} = 0 \qquad \forall \zeta \in \mathcal{S}_k^{q_n}, \qquad \forall \xi \in L^2(\Gamma). \tag{4.8b}$$

Setting $\zeta = \Pi_q(\xi)$ and $\zeta = \Pi_{q_n}(\xi)$ in (3.23) and (4.8b), respectively, and applying Cauchy–Schwarz inequality (2.11), one can easily show that

$$\|\Pi_q(\xi)\|_{L^2(\Gamma)} \leq C\|\xi\|_{L^2(\Gamma)} \quad \text{and} \quad \|\Pi_{q_n}(\xi)\|_{L^2(\Gamma)} \leq C\|\xi\|_{L^2(\Gamma)}, \tag{4.9}$$

where C is a generic positive constant depending on the parameter space Γ. Moreover, since $\mathcal{S}_k^{q_n} \subset \mathcal{S}_k^q$, we have

$$(\Pi_q(\xi) - \xi, \Pi_{q_n}(\xi))_{L^2(\Gamma)} = 0 \qquad \forall \xi \in L^2(\Gamma). \tag{4.10}$$

Let f_h, y_h^{DB}, $a_{h,N} \in V_h \otimes \mathcal{S}_k^q$, and $\mathbf{b}_{h,N} \in (V_h \otimes \mathcal{S}_k^q)^2$ denote the piecewise polynomial approximations to the right–hand side function f, the Dirichlet boundary condition y^{DB}, and the random coefficient functions $a(\boldsymbol{x}, \omega)$ and $\mathbf{b}(\boldsymbol{x}, \omega)$, respectively. By extending the work of Schötzau and Zhu [136] for a single convection diffusion equation to the parametric setting, the total error estimator becomes

$$\eta_T = \left(\underbrace{\sum_{K \in \mathcal{T}_h} \eta_{h,K}^2}_{\eta_h^2} + \underbrace{\sum_{K \in \mathcal{T}_h} \eta_{\theta,K}^2}_{\eta_\theta^2} + \underbrace{\sum_{K \in \mathcal{T}_h} \eta_{q,K}^2}_{\eta_q^2} \right)^{1/2}, \tag{4.11}$$

where the spatial error estimator for each element $K \in \mathcal{T}_h$ is

$$\begin{aligned}
\eta_{h,K}^2 = & \int_\Gamma \left(h_K^2 \|a\|_{L^2(K)}^{-1} \|f_h + \nabla \cdot (a_{h,N} \nabla y_h) - \mathbf{b}_{h,N} \cdot \nabla y_h\|_{L^2(K)}^2 \right) \rho(\xi) d\xi \\
& + \int_\Gamma \left(\sum_{E \in \partial K \setminus \partial \mathcal{D}} h_E \|a\|_{L^2(E)}^{-1} \|[\![a_{h,N} \nabla y_h]\!]\|_{L^2(E)}^2 \right) \rho(\xi) d\xi \\
& + \int_\Gamma \sum_{E \in \partial K \setminus \partial \mathcal{D}} \left(\frac{\sigma}{h_E} \|a_{h,N}\|_{L^2(E)} + \|\mathbf{b}_{h,N}\|_{L^2(E)} + \frac{h_E \|\mathbf{b}_{h,N}\|_{L^2(E)}^2}{\|a\|_{L^2(E)}} \right) \\
& \qquad \times \|[\![y_h]\!]\|_{L^2(E)}^2 \, \rho(\xi) d\xi
\end{aligned}$$

$$+\int_{\Gamma}\sum_{E\in\partial K\cap\partial\mathcal{D}}\left(\frac{\sigma}{h_E}\|a\|_{L^2(E)}+\|\mathbf{b}\|_{L^2(E)}\right)\|y_h-y_h^{DB}\|^2_{L^2(E)}\,\rho(\xi)d\xi,$$

data approximation terms caused by the discontinuity of the functions or parameters is

$$\begin{aligned}\eta^2_{\theta,K}=&\int_{\Gamma}\Big(h_K^2\|a\|^{-1}_{L^2(K)}\Big(\|f-f_h\|^2_{L^2(K)}+\|\nabla\cdot((a-a_{h,N})\nabla y_h)\|^2_{L^2(K)}\\&\qquad+\|(\mathbf{b}-\mathbf{b}_{h,N})\cdot\nabla y_h\|^2_{L^2(K)}\Big)\Big)\,\rho(\xi)d\xi,\\&+\int_{\Gamma}\Big(\sum_{E\in\partial K\setminus\partial\mathcal{D}}h_E\|a\|^{-1}_{L^2(E)}\|[\![(a-a_{h,N})\nabla y_h]\!]\|^2_{L^2(E)}\Big)\rho(\xi)d\xi\\&+\int_{\Gamma}\sum_{E\in\partial K\setminus\partial\mathcal{D}}\Big(\frac{\sigma}{h_E}\|a-a_{h,N}\|_{L^2(E)}+\|\mathbf{b}-\mathbf{b}_{h,N}\|_{L^2(E)}\\&\qquad+\frac{h_E\|\mathbf{b}-\mathbf{b}_{h,N}\|^2_{L^2(E)}}{\|a\|_{L^2(E)}}\Big)\|[\![y_h]\!]\|^2_{L^2(E)}\,\rho(\xi)d\xi\\&+\int_{\Gamma}\sum_{E\in\partial K\cap\partial\mathcal{D}}\Big(\frac{\sigma}{h_E}\|a\|_{L^2(E)}+\|\mathbf{b}\|_{L^2(E)}\Big)\|y_h^{DB}-y^{DB}\|^2_{L^2(E)}\,\rho(\xi)d\xi,\end{aligned}$$

and the parametric error estimator is equivalent to

$$\begin{aligned}\eta^2_{q,K}=&\int_{\Gamma}\|a\|^{-1}_{L^2(K)}\,\frac{1}{N}\sum_{n=1}^{N}\|\Pi_{q_n}(a\nabla y_h)-a\nabla y_h\|^2_{L^2(K)}\,\rho(\xi)\,d\xi\\&+\int_{\Gamma}\|a\|^{-1}_{L^2(K)}\,\frac{1}{N}\sum_{n=1}^{N}\|\Pi_{q_n}(\mathbf{b}\cdot\nabla y_h)-\mathbf{b}\cdot\nabla y_h\|^2_{L^2(K)}\,\rho(\xi)\,d\xi\\&+\int_{\Gamma}\sum_{E\in\partial K\setminus\partial\mathcal{D}}\|a\|^{-1}_{L^2(E)}\frac{1}{N}\sum_{n=1}^{N}\|\Pi_{q_n}\left(\mathbf{b}\cdot[\![y_h]\!]\right)-\mathbf{b}\cdot[\![y_h]\!]\|^2_{L^2(E)}\,\rho(\xi)\,d\xi.\end{aligned}$$

Before the derivation of reliability of the proposed error estimates (4.11), an operator $\mathfrak{T}_h : V_h\otimes\mathcal{S}_k^q\to H^1\otimes\mathcal{S}_k^q$ is constructed, following the discussion in [96, Theorem 2.1] for the deterministic models, satisfying $\mathfrak{T}_h v|_{\partial\mathcal{D}}=\tilde{v}\ \forall v\in V_h\otimes\mathcal{S}_k^q$ and

$$\begin{aligned}\int_{\Gamma}\sum_{K\in\mathcal{T}_h}\|v-\mathfrak{T}_h v\|^2_{L^2(K)}\,\rho(\xi)d\xi\lesssim&\int_{\Gamma}\sum_{E\in\mathcal{E}_h^0}h_E\,\|[\![v]\!]\|^2_{L^2(E)}\,\rho(\xi)d\xi\\&+\int_{\Gamma}\sum_{E\in\mathcal{E}_h^{\partial}}h_E\,\|v-\tilde{v}\|^2_{L^2(E)}\,\rho(\xi)d\xi,\end{aligned}\tag{4.12a}$$

$$\int_\Gamma \sum_{K\in\mathcal{T}_h} \|\nabla(v-\mathfrak{T}_h v)\|^2_{L^2(K)}\,\rho(\xi)d\xi \lesssim \int_\Gamma \sum_{E\in\mathcal{E}_h^0} h_E^{-1}\,\|[\![v]\!]\|^2_{L^2(E)}\,\rho(\xi)d\xi + \int_\Gamma \sum_{E\in\mathcal{E}_h^\partial} h_E^{-1}\,\|v-\tilde{v}\|^2_{L^2(E)}\,\rho(\xi)d\xi. \quad (4.12b)$$

Moreover, a Clément type interpolation $I_h : H_0^1(\mathcal{D}) \to V_h^c$, where $V_h^c = V_h \cap H_0^1(\mathcal{D})$, is a conforming subspace of V_h satisfies for $0 \le k \le l \le 2$ [39, 145]

$$\|\nabla^k(v-I_h v)\|_{L^2(K)} \lesssim h_K^{l-k}\|\nabla^l v\|_{L^2(\Delta_K)}, \quad (4.13a)$$

$$\|v-I_h v\|_{L^2(E)} \lesssim h_E^{1/2}\|\nabla v\|_{L^2(\Delta_E)}, \quad (4.13b)$$

where Δ_K and Δ_E are the union of elements that share at least one vertex with the element K and the edge E, respectively. By the inequalities (4.9) and (4.13), for $\forall v \in L^2(H^1(\mathcal{D});\Gamma)$, it further holds that

$$\|\Pi_q(v-I_h v)\|^2_{L^2(L^2(K);\Gamma)} \lesssim \|v-I_h v\|^2_{L^2(L^2(K);\Gamma)} \lesssim h_K^2\|\nabla v\|^2_{L^2(L^2(\Delta_K);\Gamma)} \quad (4.14)$$

and

$$\|\Pi_q(v-I_h v)\|^2_{L^2(L^2(E);\Gamma)} \lesssim \|v-I_h v\|^2_{L^2(L^2(E);\Gamma)} \lesssim h_E\|\nabla v\|^2_{L^2(L^2(\Delta_E);\Gamma)}. \quad (4.15)$$

Now, we derive an upper bound for the error between the continuous solution y of (4.2) and the discrete solution y_h obtained from the stochastic discontinuous Galerkin (4.5), which shows the reliability of the proposed estimator in (4.11).

Theorem 4.3.1. *Let y be the solution of* (4.2) *and $y_h \in V_h \otimes \mathcal{S}_k^q$ be its SDG approximation* (4.5)*. Then, there exists the following upper bound*

$$\|y-y_h\|_\xi \lesssim \eta_T.$$

Proof. Split the discretized solution y_h into a conforming part plus a remainder in the spatial domain as done in [95], i.e., $y_h = y_h^c + y_h^r$, where $y_h^c = \mathfrak{T}_h y_h \in V_h^c \otimes \mathcal{S}_k^q$. Then, by the triangle inequality, we have

$$\|y-y_h\|_\xi \le \|y_h^r\|_\xi + \|y-y_h^c\|_\xi. \quad (4.16)$$

First, an upper bound for the remainder term in (4.16) is derived in terms of the error estimator (4.11). By the fact that $[\![y_h^r]\!] = [\![y_h]\!]$ and the definition of the energy norm

(3.12), we have

$$\begin{aligned}\|y_h^r\|_\xi^2 = &\int_\Gamma \sum_{K\in\mathcal{T}_h}\int_K a(.,\xi)(\nabla y_h^r)^2\,dx\,\rho(\xi)d\xi + \int_\Gamma \sum_{E\in\mathcal{E}_h}\frac{\sigma a(.,\xi)}{h_E}\int_E [\![y_h]\!]^2\,ds\,\rho(\xi)d\xi\\ &+\int_\Gamma \frac{1}{2}\sum_{E\in\mathcal{E}_h^0}\int_E \mathbf{b}(.,\xi)\cdot\mathbf{n}_E((y_h^r)^e - y_h^r)^2\,ds\,\rho(\xi)d\xi\\ &+\int_\Gamma \frac{1}{2}\sum_{E\in\mathcal{E}_h^\partial}\int_E \mathbf{b}(.,\xi)\cdot\mathbf{n}_E(y_h^r)^2 ds\,\rho(\xi)d\xi. \end{aligned} \tag{4.17}$$

An application of Cauchy-Schwarz inequality (2.11), the inequalities in (4.12) with $y_h^r = y_h - \mathfrak{T}_h y_h$, adding/subtracting the data approximation terms, and Young's inequality (2.12) into the first term in (4.17) yields

$$\begin{aligned}&\int_\Gamma \sum_{K\in\mathcal{T}_h}\int_K a(.,\xi)(\nabla y_h^r)^2\,dx\,\rho(\xi)d\xi\\ &\lesssim \int_\Gamma \frac{1}{\sigma}\sum_{E\in\mathcal{E}_h^0}\frac{\sigma}{h_E}\Big(\|a_{h,N}\|_{L^2(E)} + \|a-a_{h,N}\|_{L^2(E)}\Big)\|[\![y_h]\!]\|_{L^2(E)}^2\,\rho(\xi)d\xi\\ &\quad+\int_\Gamma \frac{1}{\sigma}\sum_{E\in\mathcal{E}_h^\partial}\frac{\sigma}{h_E}\|a\|_{L^2(E)}\Big(\|y_h - y_h^{DB}\|_{L^2(E)}^2 + \|y_h^{DB}-y^{DB}\|_{L^2(E)}^2\Big)\,\rho(\xi)d\xi\\ &\lesssim \frac{1}{\sigma}\Big(\sum_{K\in\mathcal{T}_h}\eta_{h,K}^2 + \sum_{K\in\mathcal{T}_h}\eta_{\theta,K}^2\Big). \end{aligned} \tag{4.18}$$

A similar upper bound is also obtained for the second term in (4.17). For the remaining terms in (4.17), Cauchy-Schwarz (2.11) and Young's (2.12) inequalities give us

$$\begin{aligned}&\int_\Gamma \frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E \mathbf{b}(.,\xi)\cdot\mathbf{n}_E((y_h^r)^e - y_h^r)^2\,ds\,\rho(\xi)d\xi\\ &\lesssim \int_\Gamma\sum_{E\in\mathcal{E}_h^0}\big(\|\mathbf{b}_{h,N}(.,\xi)\|_{L^2(E)} + \|\mathbf{b}(.,\xi)-\mathbf{b}_{h,N}(.,\xi))\|_{L^2(E)}\big)\,\|[\![y_h]\!]\|_{L^2(E)}^2\,\rho(\xi)d\xi\\ &\quad+\int_\Gamma\sum_{E\in\mathcal{E}_h^\partial}\|\mathbf{b}(.,\xi)\|_{L^2(E)}\Big(\|y_h - y_h^{DB}\|_{L^2(E)}^2 + \|y_h^{DB}-y^{DB}\|_{L^2(E)}^2\Big)\,ds\,\rho(\xi)d\xi\\ &\lesssim \sum_{K\in\mathcal{T}_h}\eta_{h,K}^2 + \sum_{K\in\mathcal{T}_h}\eta_{\theta,K}^2. \end{aligned} \tag{4.19}$$

So combining (4.18) and (4.19), it is obtained that

$$\|y_h^r\|_\xi^2 \lesssim \sum_{K\in\mathcal{T}_h}\eta_{h,K}^2 + \sum_{K\in\mathcal{T}_h}\eta_{\theta,K}^2. \tag{4.20}$$

Next, a bound for the second term in (4.16) will be derived. By the fact that $y|_{\partial\mathcal{D}} = y_h^c|_{\partial\mathcal{D}} = y^{DB}$ due to the construction of $\mathfrak{T}_h$, we have $y - y_h^c \in H_0^1(\mathcal{D}) \otimes \mathcal{S}_k^q$. Then, the following inf–sup condition holds for all $v \in L^2(H_0^1(\mathcal{D}); \Gamma)$ [136, Lemma 4.4]

$$\|y - y_h^c\|_\xi \lesssim \sup_{v \in L^2(H_0^1(\mathcal{D});\Gamma)} \frac{\widetilde{a}_\xi(y - y_h^c, v)}{\|v\|_\xi}. \tag{4.21}$$

By the bilinear systems (4.5) and (4.6), Clément type interpolation estimates (4.13), and the continuity of the bilinear form (3.13b), one can have

$$\begin{aligned}
\widetilde{a}_\xi(y - y_h^c, v) &= \widetilde{a}_\xi(u, v) - \widetilde{a}_\xi(y_h^c, v) = \int_\Gamma \int_\mathcal{D} f v \, dx \, \rho(\xi) d\xi - \widetilde{a}_\xi(y_h^c, v) \\
&= \int_\Gamma \int_\mathcal{D} f v \, dx \, \rho(\xi) d\xi - \widetilde{a}_\xi(y_h, v) + \widetilde{a}_\xi(y_h^r, v) \\
&= \int_\Gamma \int_\mathcal{D} f I_h v \, dx \, \rho(\xi) d\xi + \int_\Gamma \int_\mathcal{D} f(v - I_h v) \, dx \, \rho(\xi) d\xi - \widetilde{a}_\xi(y_h, v) + \widetilde{a}_\xi(y_h^r, v) \\
&= \int_\Gamma \int_\mathcal{D} f(v - I_h v) \, dx \, \rho(\xi) d\xi - \widetilde{a}_\xi(y_h, v - I_h v) + j_\xi(y_h, I_h v) + \widetilde{a}_\xi(y_h^r, v) \\
&\leq \|y_h^r\|_\xi \|v\|_\xi + j_\xi(y_h, I_h v) + \underbrace{\int_\Gamma \int_\mathcal{D} f(v - I_h v) \, dx \, \rho(\xi) d\xi - \widetilde{a}_\xi(y_h, v - I_h v)}_{T}.
\end{aligned} \tag{4.22}$$

The term $j_\xi(y_h, I_h v)$ in (4.22) is equivalent to

$$j_\xi(y_h, I_h v) = -\int_\Gamma \sum_{E \in \mathcal{E}_h} \int_E \{\!\{ a(., \xi) \nabla I_h v \}\!\} [\![y_h]\!] \, ds \, \rho(\xi) \, d\xi,$$

since $I_h v \in V_h^c$. Then, with the help of Cauchy–Schwarz (2.11) and inverse estimate (2.9)

$$\|v\|_{L^2(E)} \lesssim h_K^{-1/2} \|v\|_{L^2(K)} \qquad \forall v \in V_h$$

with $h_E \lesssim h_K$, and adding/subtracting the data approximation terms, it yields

$$\begin{aligned}
&j_\xi(y_h, I_h v) \\
&\lesssim \sum_{E \in \mathcal{E}_h} \|a \nabla I_h v\|_{L^2(L^2(E);\Gamma)} \|[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)} \\
&\lesssim \sum_{E \in \mathcal{E}_h} h_E^{1/2} \|a^{1/2} \nabla I_h v\|_{L^2(L^2(E);\Gamma)} \sum_{E \in \mathcal{E}_h} \|a^{1/2}\|_{L^2(L^2(E);\Gamma)} h_E^{-1/2} \|[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)}
\end{aligned}$$

$$\lesssim \sum_{K\in\mathcal{T}_h} \|a^{1/2}\nabla I_h v\|_{L^2(L^2(K);\Gamma)} \sum_{E\in\mathcal{E}_h} \|a^{1/2}\|_{L^2(L^2(E);\Gamma)} h_E^{-1/2} \|[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)}$$

$$\lesssim \sigma^{-1} \left(\sum_{K\in\mathcal{T}_h} \eta_{h,K}^2 + \sum_{K\in\mathcal{T}_h} \eta_{\theta,K}^2 \right)^{1/2} \|v\|_\xi. \tag{4.23}$$

To find a bound for the term denoted by T in (4.22), the L^2–projection operator Π_q (4.8) is first added/subtracted with $v \in L^2(H_0^1(\mathcal{D});\Gamma)$

$$T = \sum_{i=1}^{4} A_i, \tag{4.24}$$

where

$$\begin{aligned}
A_1 =& \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \big(f(v - I_h v) - \Pi_q(a\nabla y_h)\cdot\nabla(v - I_h v)\big) dx\, \rho(\xi) d\xi \\
&- \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \Pi_q(\mathbf{b}\cdot\nabla y_h)(v - I_h v) dx\, \rho(\xi) d\xi \\
&- \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\mathcal{D}} \Pi_q((\mathbf{b}\cdot\mathbf{n}_E)(y_h^e - y_h))(v - I_h v)\, ds\, \rho(\xi) d\xi, \\
A_2 =& \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \big(\Pi_q(a\nabla y_h) - a\nabla y_h\big)\cdot\nabla(v - I_h v)\, dx\, \rho(\xi) d\xi, \\
A_3 =& \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \big(\Pi_q(\mathbf{b}\cdot\nabla y_h) - \mathbf{b}\cdot\nabla y_h\big)(v - I_h v)\, dx\, \rho(\xi) d\xi, \\
A_4 =& \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\mathcal{D}} \big(\Pi_q((\mathbf{b}\cdot\mathbf{n}_E)(y_h^e - y_h)) - (\mathbf{b}\cdot\mathbf{n}_E)(y_h^e - y_h)\big)(v - I_h v) dx\, \rho(\xi) d\xi.
\end{aligned}$$

With the help of the definition of L^2–projection operator Π_q (4.8a), the fact that $\Pi_q f = f$, and the integration by parts over $\mathcal{D}$, it is obtained that

$$\begin{aligned}
A_1 =& \underbrace{\int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \big(f + \nabla\cdot(a\nabla y_h) - \mathbf{b}\cdot\nabla y_h\big)\Pi_q(v - I_h v)\, dx\, \rho(\xi) d\xi}_{A_{1,1}} \\
&\underbrace{- \int_\Gamma \sum_{E\in\mathcal{E}_h^0} \int_E a\nabla y_h\cdot\mathbf{n}_E \Pi_q(v - I_h v)\, ds\, \rho(\xi) d\xi}_{A_{1,2}} \\
&\underbrace{- \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_{\partial K^-\setminus\partial\mathcal{D}} ((\mathbf{b}\cdot\mathbf{n}_E)(y_h^e - y_h))\Pi_q(v - I_h v)\, ds\, \rho(\xi) d\xi}_{A_{1,3}}.
\end{aligned} \tag{4.25}$$

By adding/subtracting the data approximation terms and using Cauchy–Schwarz inequality (2.11) with (4.14) and (3.12), a bound for the first term in (4.25) is derived

$$
\begin{aligned}
A_{1,1} = &\int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \big(f_h + \nabla\cdot(a_{h,N}\nabla y_h) - \mathbf{b}_{h,N}\cdot\nabla y_h\big)\Pi_q(v - I_h v)\,dx\,\rho(\xi)d\xi \\
&+ \int_\Gamma \sum_{K\in\mathcal{T}_h} \int_K \big((f - f_h) + \nabla\cdot((a - a_{h,N})\nabla y_h) - (\mathbf{b} - \mathbf{b}_{h,N})\cdot\nabla y_h\big) \\
&\qquad\qquad \times \Pi_q(v - I_h v)\,dx\,\rho(\xi)d\xi \\
\lesssim &\left(\sum_{K\in\mathcal{T}_h} \eta_{h,K}^2 + \eta_{\theta,K}^2\right)^{1/2} \|v\|_\xi. \qquad (4.26a)
\end{aligned}
$$

Analogously, an application of the Cauchy–Schwarz inequality (2.11), the inequality (4.15), and the definition of the energy norm (3.12) yields

$$
\begin{aligned}
A_{1,2} &\lesssim \sum_{E\in\mathcal{E}^0} \|[\![a\nabla y_h]\!]\|_{L^2(L^2(E);\Gamma)}\, \|\Pi_q(v - I_h v)\|_{L^2(L^2(E);\Gamma)} \\
&\lesssim \sum_{E\in\mathcal{E}^0} \|[\![((a - a_{h,N}) + a_{h,N})\nabla y_h]\!]\|_{L^2(L^2(E);\Gamma)}\, h_E^{1/2}\|\nabla v\|_{L^2(L^2(\Delta_E);\Gamma)} \\
&\lesssim \left(\sum_{K\in\mathcal{T}_h} \eta_{h,K}^2 + \eta_{\theta,K}^2\right)^{1/2} \|v\|_\xi, \qquad (4.26b)
\end{aligned}
$$

$$
\begin{aligned}
A_{1,3} &\lesssim \sum_{E\in\mathcal{E}^0} \|\mathbf{b}\cdot[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)}\, \|\Pi_q(v - I_h v)\|_{L^2(L^2(E);\Gamma)} \\
&\lesssim \sum_{E\in\mathcal{E}^0} \|\big(\mathbf{b} - \mathbf{b}_{h,N}) + \mathbf{b}_{h,N}\big)\cdot[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)}\, h_E^{1/2}\|\nabla v\|_{L^2(L^2(\Delta_E);\Gamma)} \\
&\lesssim \left(\sum_{K\in\mathcal{T}_h} \eta_{h,K}^2 + \eta_{\theta,K}^2\right)^{1/2} \|v\|_\xi. \qquad (4.26c)
\end{aligned}
$$

Next, it will be shown that the rest of terms in (4.24) is bounded in terms of the parametric estimator given in (4.11). By $\mathcal{S}_k^{q_n} \subset \mathcal{S}_k^q \subset L^2(\Gamma)$ and (4.8), one can obtain for each $n = 1, \dots N$

$$
\begin{aligned}
&\|\Pi_q(a\nabla y_h) - a\nabla y_h\|_{L^2(L^2(\mathcal{D});\Gamma)} \\
&\leq \underbrace{\|\Pi_q(a\nabla y_h) - \Pi_{q_n}(a\nabla y_h)\|_{L^2(L^2(\mathcal{D});\Gamma)}}_{=0} + \|\Pi_{q_n}(a\nabla y_h) - a\nabla y_h\|_{L^2(L^2(\mathcal{D});\Gamma)}.
\end{aligned}
$$

Then,

$$N\|\Pi_q(a\nabla y_h) - a\nabla y_h\|_{L^2(L^2(\mathcal{D});\Gamma)} \leq \sum_{n=1}^{N}\|\Pi_{q_n}(a\nabla y_h) - a\nabla y_h\|_{L^2(L^2(\mathcal{D});\Gamma)}. \quad (4.27)$$

Hence, the Cauchy–Schwarz inequality (2.11), the inequalities (4.13) and (4.27) with $h_E \leq h_K < 1$ give us

$$\begin{aligned}
A_2 &= \int_\Gamma \sum_{K\in\mathcal{T}_h}\int_K \big(\Pi_q(a\nabla y_h) - a\nabla y_h\big)\cdot\nabla(v - I_h v)\,dx\,\rho(\xi)d\xi\\
&\leq \frac{1}{N}\sum_{n=1}^{N}\|\Pi_{q_n}(a\nabla y_h) - a\nabla y_h\|_{L^2(L^2(\mathcal{D});\Gamma)}\|\nabla v\|_{L^2(L^2(\mathcal{D});\Gamma)}\\
&\leq \frac{1}{N}\sum_{n=1}^{N}\|\Pi_{q_n}(a\nabla y_h) - a\nabla y_h\|_{L^2(L^2(\mathcal{D});\Gamma)}\,\|a^{-1/2}\|_{L^2(L^2(\mathcal{D});\Gamma)}\|v\|_\xi\\
&\leq \left(\sum_{K\in\mathcal{T}_h}\eta_{q,K}^2\right)^{1/2}\|v\|_\xi, \qquad (4.28a)\\
A_3 &= \int_\Gamma \sum_{K\in\mathcal{T}_h}\int_K \big(\Pi_q(\mathbf{b}\cdot\nabla y_h) - \mathbf{b}\cdot\nabla y_h\big)(v - I_h v)\,dx\,\rho(\xi)d\xi\\
&\leq \sum_{K\in\mathcal{T}_h}\|\big(\Pi_q(\mathbf{b}\cdot\nabla y_h) - \mathbf{b}\cdot\nabla y_h\big)\|_{L^2(L^2(K);\Gamma)}h_K\|\nabla v\|_{L^2(L^2(K);\Gamma)}\\
&\leq \frac{1}{N}\sum_{n=1}^{N}\|\big(\Pi_{q_n}(\mathbf{b}\cdot\nabla y_h) - \mathbf{b}\cdot\nabla y_h\big)\|_{L^2(L^2(\mathcal{D});\Gamma)}\,\|a^{-1/2}\|_{L^2(L^2(\mathcal{D});\Gamma)}\|v\|_\xi\\
&\leq \left(\sum_{K\in\mathcal{T}_h}\eta_{q,K}^2\right)^{1/2}\|v\|_\xi, \qquad (4.28b)\\
A_4 &= \int_\Gamma \sum_{K\in\mathcal{T}_h}\int_{\partial K^-\setminus\partial\mathcal{D}} \big(\Pi_q((\mathbf{b}\cdot\mathbf{n}_E)(y_h^e - y_h)) - (\mathbf{b}\cdot\mathbf{n}_E)(y_h^e - y_h)\big)(v - I_h v)dx\,\rho(\xi)d\xi\\
&\lesssim \sum_{E\in\mathcal{E}^0}\|\Pi_q(\mathbf{b}\cdot[\![y_h]\!]) - \mathbf{b}\cdot[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)}\|v - I_h v\|_{L^2(L^2(E);\Gamma)}\\
&\lesssim \sum_{E\in\mathcal{E}^0}\|\Pi_q(\mathbf{b}\cdot[\![y_h]\!]) - \mathbf{b}\cdot[\![y_h]\!]\|_{L^2(L^2(E);\Gamma)}h_E^{1/2}\|\nabla v\|_{L^2(L^2(\Delta_E);\Gamma)}\\
&\lesssim \left(\sum_{K\in\mathcal{T}_h}\eta_{q,K}^2\right)^{1/2}\|v\|_\xi. \qquad (4.28c)
\end{aligned}$$

Inserting the bounds obtained in (4.20), (4.23), (4.26), and (4.28) into (4.21), we get

$$\|y - y_h^c\|_\xi \lesssim \left(\sum_{K\in\mathcal{T}_h}\eta_{h,K}^2 + \eta_{\theta,K}^2 + \eta_{q,h}^2\right)^{1/2}. \quad (4.29)$$

Last, combining the results in (4.20) and (4.29), we obtain the desired result. □

4.4 Numerical Experiments

This section now uses the error estimator developed in Section 4.3 to propose an adaptive stochastic discontinuous Galerkin method for the numerical solutions of (4.2). In the next section, the adaptive algorithm is described in details.

4.4.1 Adaptive Loop

A generic adaptive refinement/enrichment procedure for the numerical solution of the model problem (4.2) consists of successive loops of the sequence given in (4.1). Starting with a given triangulation $\mathcal{T}_h^0$ and an initial index set $\mathfrak{B}^0$, the adaptive procedure generates a sequence of triangulations $\mathcal{T}_h^k \subseteq \mathcal{T}_h^{k+1}$ and of index sets $\mathfrak{B}^k \subseteq \mathfrak{B}^{k+1}$.

The **SOLVE** step (subroutine **solve**) corresponds to the numerical approximations of the statistical moments obtained by the stochastic discontinuous Galerkin numerical scheme in the current mesh and index set. After an application of the discretization techniques, one needs to solve the following linear system:

$$\underbrace{\left(\sum_{i=0}^{N} \mathcal{G}_i \otimes \mathcal{K}_i\right)}_{\mathcal{A}} \mathbf{y} = \underbrace{\left(\sum_{i=0}^{N} \mathbf{g}_i \otimes \mathbf{f}_i\right)}_{\mathcal{F}}, \tag{4.30}$$

where $\mathbf{y} = \big(y_0, \ldots, y_{(\#(\mathfrak{B}^k)-1)}\big)^T$ with $y_i \in \mathbb{R}^{N_d}$, $i = 0, 1, \ldots, (\#(\mathfrak{B}^k) - 1)$ and N_d corresponds to the degree of freedoms for the spatial discretization. Here, $\mathcal{K}_i \in \mathbb{R}^{N_d \times N_d}$ and $\mathcal{G}_i \in \mathbb{R}^{(\#(\mathfrak{B}^k))\times(\#(\mathfrak{B}^k))}$ represent the stiffness and stochastic matrices, respectively, whereas $\mathbf{f}_i \in \mathbb{R}^{N_d}$ and $\mathbf{g}_i \in \mathbb{R}^{(\#(\mathfrak{B}^k))}$ are the right–hand side and stochastic vectors, respectively; see Section 3.1.2 for the constructions of the matrices and vectors in details. In order to solve the matrix system (4.30), we use a generalized minimal residual (GMRES) solver, given in Algorithm 5, in a combination with the well known mean-based preconditioner [128], that is,

$$\mathcal{P}_0 = \mathcal{G}_0 \otimes \mathcal{K}_0,$$

where $\mathcal{G}_0$ and $\mathcal{K}_0$ are the mean stochastic and stiffness matrices, respectively. It is noted that $\mathcal{P}_0$ is one of the most commonly used preconditioners for solving PDEs with random data since it is a block diagonal matrix obtained by the orthogonality

of the stochastic basis functions. In the **ESTIMATE** step (subroutine **estimate**), we compute the error estimator (4.11) contributed from the physical domain, the data approximation terms, and the parametric domain in order to control the behaviour of the error. In the a posteriori error estimates (4.11), some terms still contain continuous data terms. To overcome this problem, continuous terms a, $\mathbf{b}$ are replaced by the discrete ones $a_{h,N}$, $\mathbf{b}_{h,N}$. By (2.17), there is also a bound for the data oscillation term caused by a random variable $z(\boldsymbol{x},\omega)$ as follows

$$\begin{aligned}\|z - z_{h,N}\|_{L^2(\Gamma;L^2(\mathcal{D}))} &\leq \|z - z_N\|_{L^2(\Gamma;L^2(\mathcal{D}))} + \|z_N - z_{h,N}\|_{L^2(\Gamma;L^2(\mathcal{D}))} \\ &\lesssim \underbrace{\sum_{i=N+1}^{N_\infty} \lambda_i}_{\Lambda_z} + \|z_N - z_{h,N}\|_{L^2(\Gamma;L^2(\mathcal{D}))},\end{aligned}$$

where N_∞ is chosen as the KL expansion can cover the 97% of sum of the eigenvalues λ_i. In order to equidistribute the error, the data estimator $\eta_{\theta,K}$ is decomposed as follows

$$\eta^2_{\theta,K} = \eta^2_{\theta,h,K} + \eta^2_{\theta,q,K}, \tag{4.31}$$

where

$$\begin{aligned}\eta^2_{\theta,h,K} = \int_\Gamma \Big(h_K^2 \|a_{h,N}\|^{-1}_{L^2(K)} \Big(&\|f - f_h\|^2_{L^2(K)} + \|\nabla\cdot((a_N - a_{h,N})\nabla y_h)\|^2_{L^2(K)} \\ &+ \|(\mathbf{b}_N - \mathbf{b}_{h,N})\cdot\nabla y_h\|^2_{L^2(K)} \Big)\Big)\rho(\xi)d\xi, \\ + \int_\Gamma \Big(\sum_{E\in\partial K\setminus\partial\mathcal{D}} h_E \|a_{h,N}\|^{-1}_{L^2(E)} &\|[\![(a_N - a_{h,N})\nabla y_h]\!]\|^2_{L^2(E)} \Big)\rho(\xi)d\xi \\ + \int_\Gamma \sum_{E\in\partial K\setminus\partial\mathcal{D}} \Big(\frac{\sigma}{h_E}\|a_N - a_{h,N}\|_{L^2(E)} &+ \|\mathbf{b}_N - \mathbf{b}_{h,N}\|_{L^2(E)} \\ &+ \frac{h_E\|\mathbf{b}_N - \mathbf{b}_{h,N}\|^2_{L^2(E)}}{\|a_{h,N}\|_{L^2(E)}}\Big)\|[\![y_h]\!]\|^2_{L^2(E)}\ \rho(\xi)d\xi \\ + \int_\Gamma \sum_{E\in\partial K\cap\partial\mathcal{D}} \Big(\frac{\sigma}{h_E}\|a_{h,N}\|_{L^2(E)} + \|\mathbf{b}_{h,N}\|_{L^2(E)}\Big)\ &\|y_h^{DB} - y^{DB}\|^2_{L^2(E)}\ \rho(\xi)d\xi\end{aligned}$$

and

$$\begin{aligned}\eta^2_{\theta,q,K} = \int_\Gamma \Big(h_K^2 \|a_{h,N}\|^{-1}_{L^2(K)} \Big(&\|\nabla\cdot(\Lambda_a\nabla y_h)\|^2_{L^2(K)} + \|\Lambda_b\cdot\nabla y_h\|^2_{L^2(K)} \Big)\Big)\rho(\xi)d\xi, \\ + \int_\Gamma \Big(\sum_{E\in\partial K\setminus\partial\mathcal{D}} h_E \|a_{h,N}\|^{-1}_{L^2(E)} &\|[\![\Lambda_a\nabla y_h]\!]\|^2_{L^2(E)} \Big)\rho(\xi)d\xi\end{aligned}$$

$$+\int_{\Gamma}\sum_{E\in\partial K\setminus\partial\mathcal{D}}\Big(\frac{\sigma\Lambda_a}{h_E}+\Lambda_b+\frac{h_E\Lambda_b^2}{\|a_{h,N}\|_{L^2(E)}}\Big)\|[\![y_h]\!]\|^2_{L^2(E)}\ \rho(\xi)d\xi.$$

Algorithm 5 Preconditioned GMRES (PGMRES) [133]

Input: Matrices $\mathcal{A},\mathcal{P}\in\mathbb{R}^{N_d\times(\#(\mathfrak{B}^k))}$, right–hand side vector $\mathcal{F}\in\mathbb{R}^{N_d(\#(\mathfrak{B}^k))}$.

Output: Vector $\mathbf{y}\in\mathbb{R}^{N_d(\#(\mathfrak{B}^k))}$ satisfying $\|\mathcal{A}\,\mathbf{y}-\mathcal{F}\|_2\leq\epsilon_{tol}$.

Solve r_0 from $\mathcal{P}\,r_0=\mathcal{F}-\mathcal{A}\,\mathbf{y}_0$ for some initial guess $\mathbf{y}_0$.
$v_1=r_0/\|r_0\|_2$
$\xi=[\xi_1,0,\ldots,0],\qquad \xi_1=\|v_1\|_2$
for $k=1,\ldots,$maxit **do**
  Solve w from $\mathcal{P}\,w=\mathcal{A}\,v_k$
  for $i=1,\ldots,k$ **do**
    $h_{i,k}=\langle w,v_i\rangle$
    $w=w-h_{i,k}v_i$
  end for
  $h_{k+1,k}=\|w\|_2$
  $v_{k+1}=w/h_{k+1,k}$
  Apply Givens rotations to kth column of h, i.e.,
  for $i=1,\ldots,k-1$ **do**
    $$\begin{bmatrix}h_{i,k}\\h_{i+1,k}\end{bmatrix}=\begin{bmatrix}c_i&s_i\\-s_i&c_i\end{bmatrix}\begin{bmatrix}h_{i,k}\\h_{i+1,k}\end{bmatrix}$$
  end for
  Compute kth rotation, and apply to ξ and last column of h.
  $$\begin{bmatrix}\xi_k\\\xi_{k+1}\end{bmatrix}=\begin{bmatrix}c_i&s_i\\-s_i&c_i\end{bmatrix}\begin{bmatrix}\xi_k\\0\end{bmatrix}$$
  $h_{k,k}=c_kh_{k,k}+s_kh_{k+1,k},\qquad h_{k+1,k}=0$
  if $|\xi_{k+1}|$ sufficiently small **then**
    Solve $Hq=\xi$, where the entries of H are $h_{j,k}$.
    $Q=[q_1v_1,\ldots,q_kv_k]$
    Solve $\mathcal{P}\,\widetilde{Q}=Q$
    $\mathbf{y}=\mathbf{y}_0+\widetilde{Q}$
    return
  end if
end for

In the computation of spatial terms of the estimator, we just make the Kronecker product of the spatial contributions with the stochastic matrix $\mathcal{G}_0$, whereas stochastic

mass matrix is used in the computation of parametric term by following the classical projection computation. In the step **MARK** (subroutine **mark**), a bulk criterion is used to specify the elements in $\mathcal{T}_h^k$ by using the a posteriori error estimators $\eta_{h,K}$ and $\eta_{\theta,h}$ for a fixed marking parameter $0 < \theta_h \leq 1$:

$$\theta_h \sum_{K \in \mathcal{T}_h^k} \left(\eta_{h,K}^2 + \eta_{\theta,h,K}^2\right) \leq \sum_{K \in \mathcal{M}_h^k \subset \mathcal{T}_h^k} \left(\eta_{h,K}^2 + \eta_{\theta,h,K}^2\right). \tag{4.32}$$

On the other hand, to build a minimal subset of marked indices, we first define a new index set

$$\mathfrak{R}^k := \{q \in \mathfrak{U} \backslash \mathfrak{B}^k \, : \, q = q_{\mathfrak{B}} \pm \epsilon^{(n)} \;\; \forall q_{\mathfrak{B}} \in \mathfrak{B}^k, \; \forall n = 1, \dots N_{\mathfrak{B}}\}, \tag{4.33}$$

where the counter parameter is

$$N_{\mathfrak{B}} = \begin{cases} 0, & \text{if } \mathfrak{B} = \{\mathbf{0}\}, \\ \max\{\max(\text{supp}(q_{\mathfrak{B}})) \, : q_{\mathfrak{B}} \in \mathfrak{B} \backslash \{\mathbf{0}\}\}, & \text{otherwise}, \end{cases}$$

and $\epsilon^{(n)} = \left(\epsilon_1^{(n)}, \epsilon_2^{(n)}, \dots\right)$ is the Kronecker delta sequence satisfying $\epsilon_j^{(n)} = \delta_{nj}$ for all $j \in \mathbb{N}$. Then, by using parametric error estimators for a fixed marking parameter $0 < \theta_q \leq 1$, we apply a bulk criterion in parametric setting as follows

$$\theta_q \sum_{q \in \mathfrak{R}^k} \eta_q^2 \leq \sum_{q \in \mathfrak{M}^k \subseteq \mathfrak{R}^k} \eta_q^2. \tag{4.34}$$

In spite of performing the marking for both triangulation and indices, we only enrich one direction by comparing the spatial and parametric error estimators. Finally, the **REFINE** step (subroutine **refine**) either performs a refinement of the current triangulation or an enrichment of the current index set. In the refinement of a triangulation, the marked elements are refined by longest edge bisection, whereas the elements of the marked edges are refined by bisection [44]. On the other hand, the enrichment of the polynomial space is made by adding all marked indices to the current index, i.e., $\mathfrak{B}^{k+1} = \mathfrak{B}^k \cup \mathfrak{M}^k$. Overall, the adaption process is summarized in Algorithm 6, which repeats until the prescribed tolerance TOL is met by the total estimator or maximum number of total degree of freedoms $\max_{dof}$ is reached.

Algorithm 6 Adaptive stochastic discontinuous Galerkin algorithm

Input: initial mesh $\mathcal{T}_h^0$, initial index set $\mathfrak{B}^0$, marking parameters θ_h and θ_q, data $f, y^{DB}, a, \mathbf{b}$, tolerance threshold TOL, and maximum degree of freedoms $\max_{dof}$.

for $k = 0, 1, 2, \ldots$ **do**
  $y_h^k = \textbf{solve}(\mathcal{T}_h^k, \mathfrak{B}^k, f, y^{DB}, a, \mathbf{b})$
  $(\eta_h^2, \eta_{\theta,h}^2, \eta_{\theta,q}^2, \eta_q^2) = \textbf{estimate}(y_h^k, \mathcal{T}_h^k, \mathfrak{B}^k, f, y^{DB}, a, \mathbf{b})$
  if $(\eta_T^2 \leq TOL)$ or $(\#(\mathfrak{B}^k) \times N_d \leq \max_{dof})$ **then**
    break;
  end if
  $\mathcal{M}_h^k = \textbf{mark}(\mathcal{T}_h^k, \eta_h^2, \eta_{\theta,h}^2, \theta_h)$ and $\mathfrak{M}^k = \textbf{mark}(\mathfrak{U}^k, \eta_q^2, \theta_q)$
  if $(\eta_h^2 + \eta_{\theta,h}^2) \geq (\eta_q^2 + \eta_{\theta,q}^2)$ **then**
    $\mathcal{T}_h^{k+1} = \textbf{refine}(\mathcal{T}_h^k, \mathcal{M}_h^k)$
  else
    $\mathfrak{B}^{k+1} = \textbf{refine}(\mathfrak{B}^k, \mathfrak{M}^k)$
  end if
end for

4.4.2 Numerical Results

This section now presents several numerical results in order to examine the quality of derived estimators in Section 4.3 and the performance of the adaptive loop proposed in Section 4.4.1. In the numerical experiments, the random input z is characterized by the covariance function in (3.59) with the correlation length ℓ_n and the eigenpair (λ_j, ϕ_j). Since the underlying random variables have been chosen based on the uniform distribution over $[-1, 1]$, that is, $\xi_j \sim \mathcal{U}[-1, 1]$ for $i = 1, \ldots, N$, Legendre polynomials are used as the stochastic basis functions; see Table 2.1. On the other hand, linear elements are employed to generate a discontinuous Galerkin basis. In the numerical implementations, the initial mesh and index set are chosen as $\mathcal{T}_h^0$ with $N_d = 384$ and $\mathfrak{B}^0 = \{(0, 0, 0, \ldots), (1, 0, 0, \ldots)\}$, respectively. Then, uniform meshes are constructed by dividing each triangle into four triangles, whereas uniform parametric spaces are generated by increasing the truncation number N and polynomial degree ψ by one unit. Unless otherwise stated, in all simulations, we take the correlation length $\ell = 1$ and standard deviation $\kappa = 0.05$. The Algorithm 6 is terminated when the total error estimator η_T is reduced to $TOL = 1e - 6$ or the degree of freedoms is reached to maximum ($\max_{dof} = 10^7$). Further, all parameters used in the simulations are described in Table 4.1.

Table 4.1: Descriptions of the parameters used in the simulations.

Parameter	Description
N_d	degree of freedoms for the spatial discretization
N	truncation number in KL expansion
Q	highest order of basis polynomials for the stochastic domain
ν	viscosity parameter
ℓ	correlation length
κ_z	standard deviation
θ_h	marking parameter for spatial discretization
θ_q	marking parameter for parametric discretization
$\#\,\mathfrak{B}$	size of index

4.4.2.1 Example with Random Diffusivity

As a first benchmark problem, a two-dimensional convection diffusion equation with random diffusion parameter examined in Section 3.5.1 is considered on $\mathcal{D} = [-1, 1]^2$ by choosing the deterministic source function $f(\boldsymbol{x}) = 0$, the deterministic convection parameter $\mathbf{b}(\boldsymbol{x}) = (0, 1)^T$, and the nonhomogeneous Dirichlet boundary condition

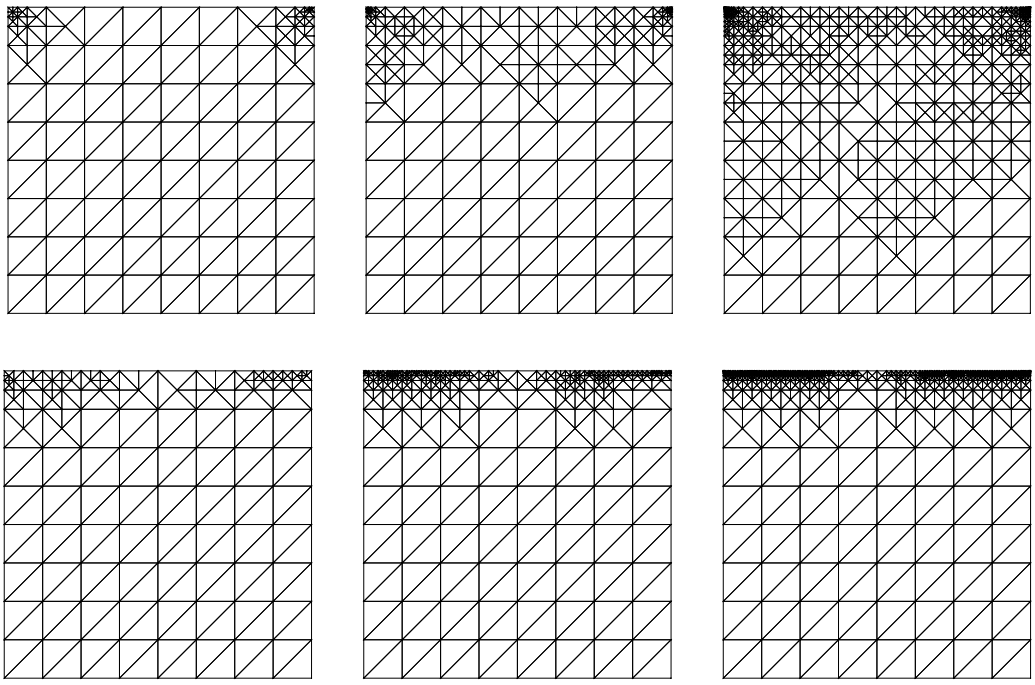

Figure 4.1: Example 4.4.2.1: Process of adaptively refined triangulations obtained by Algorithm 6 with the marking parameters $\theta_h = 0.5$, $\theta_q = 0.5$, the initial mesh $\mathcal{T}_h^0$, and the initial index set $\mathfrak{B}^0$ for the viscosity parameter $\nu = 10^0$ (top) and $\nu = 10^{-2}$ (bottom).

$$y^{DB}(\boldsymbol{x}) = \begin{cases} y^{DB}(x_1,-1) = x_1, & y^{DB}(x_1,1) = 0, \\ y^{DB}(-1,x_2) = -1, & y^{DB}(1,x_2) = 1. \end{cases}$$

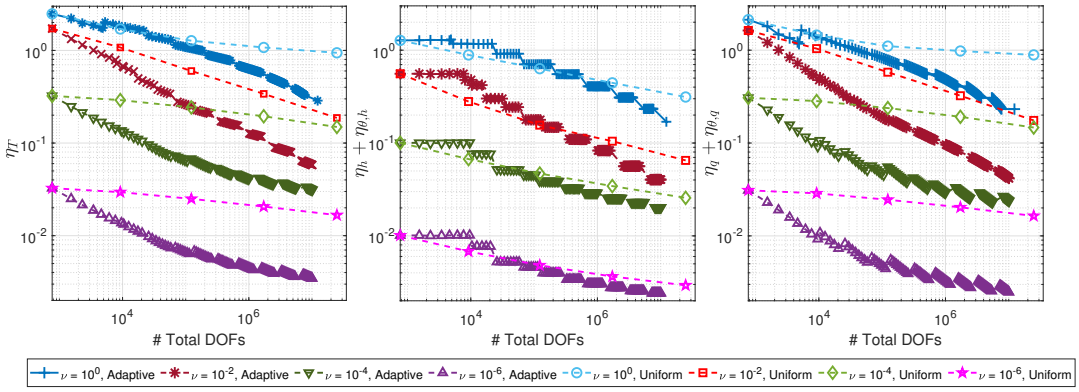

Figure 4.2: Example 4.4.2.1: Behaviour of error estimators on the adaptively and uniformly generated spatial/parametric spaces for various values of viscosity parameter ν.

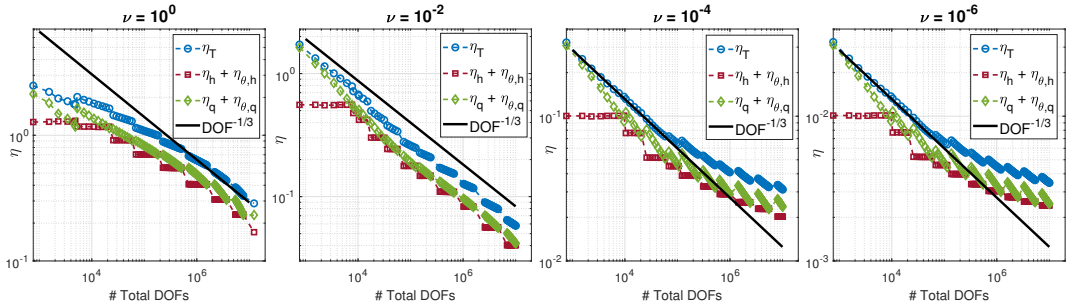

Figure 4.3: Example 4.4.2.1: Behaviour of error estimators on the adaptively generated spatial/parametric spaces with marking parameter $\theta_h = 0.5$, $\theta_q = 0.5$ for various values of viscosity parameter ν.

On the other hand, diffusion parameter $a(\boldsymbol{x},\omega)$ is defined in the form of $a(\boldsymbol{x},\omega) = \nu z(\boldsymbol{x},\omega)$, where $z(\boldsymbol{x},\omega)$ is a random variable having unity mean $\overline{z}(\boldsymbol{x}) = 1$ and ν is the viscosity parameter. As ν decreases, the solution of the underlying problem exhibits an exponential boundary layer near, where the value of the solution changes dramatically; see Figure 3.1. Locally refined triangulations generated by following the Algorithm 6 with the marking parameters $\theta_h = 0.5$, $\theta_q = 0.5$, the initial coarse mesh $\mathcal{T}_h^0$, and the initial index set $\mathfrak{B}^0$ are given in Figure 4.1. It is observed that our estimator η_T (4.11) detects the regions well where the mean of the solution is not sufficiently.

Figure 4.2 displays the behavior of estimator η_T and its spatial and parametric contributions $\eta_h + \eta_{\theta,h}$ and $\eta_q + \eta_{\theta,q}$, respectively, on the adaptively (with the marking parameters $\theta_h = 0.5$, $\theta_q = 0.5$) and uniformly generated spatial/parametric spaces for various values of viscosity parameter ν. Results on Figure 4.2 show that the estimators on the adaptively generated spaces are superior to the ones on the uniformly generated spaces. From Figure 4.3, we also observe that the total error estimates decay with an overall rate of about $\mathcal{O}(DOF^{-1/3})$ even for the smaller values of ν; see [27] for the convergence analysis of adaptive stochastic Galerkin applied to elliptic problems.

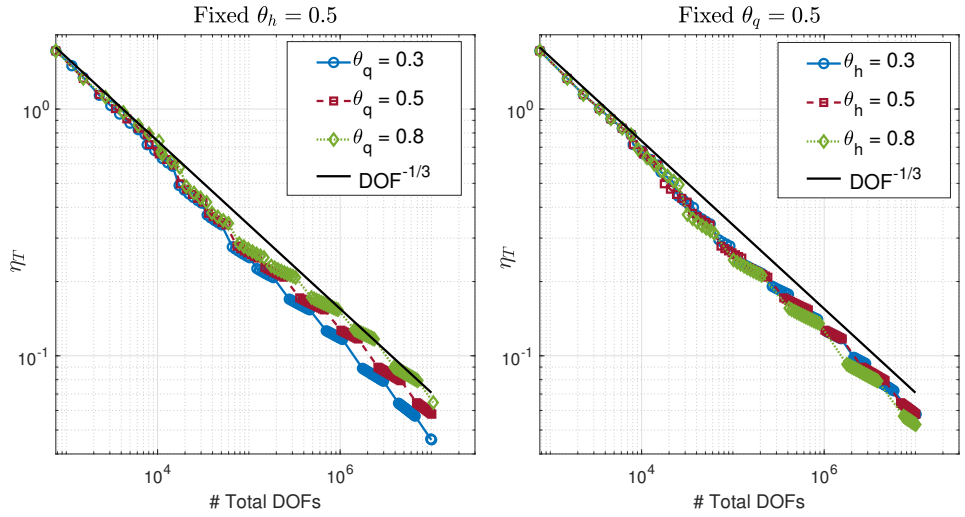

Figure 4.4: Example 4.4.2.1: Effect of the marking parameters θ_q (left) and θ_h (right) on the behaviour of estimator η_T for $\nu = 10^{-2}$.

Next, the effect of marking parameters θ_q and θ_h is investigated in Tables 4.2 and 4.3, respectively, for the viscosity parameter $\nu = 10^{-2}$. As expected, bigger values for the parameter θ_q result in more enrichment in one loop, whereas smaller θ_q yields more optimal index but more enrichment loops. When θ_q is fixed, although the iteration number decreases as θ_h increases, the process does not perform optimally; see also Figure 4.4. However, the adaptive algorithm converges with an overall rate of about $\mathcal{O}(DOF^{-1/3})$ regardless of the marking parameters.

Last, we investigate the effect of the correlation length ℓ and the standard deviation κ on the estimator η_T with the adaptively generated spatial/parametric spaces in Figure 4.5. As expected, the performance of the estimator becomes worse when we decrease the value of the correlation length ℓ and increase the value of the standard

Table 4.2: Example 4.4.2.1: Results of adaptive procedure with the viscosity parameter $\nu = 10^{-2}$ for varying marking parameter θ_q.

$\theta_h = 0.5$		$\theta_q = 0.3$		$\theta_q = 0.5$		$\theta_q = 0.8$
# iter		78		64		59
# Total DOFs		12,880,665		10,240,608		12,735,090
# $\mathfrak{B}$		721		784		1010
N_d		17865		13062		12609
η_T		4.7329e-02		6.1645e-02		6.4343e-02
$\eta_h + \eta_{\theta,h}$		3.1194e-02		4.1618e-02		4.3911e-02
$\eta_q + \eta_{\theta,q}$		3.5594e-02		4.5476e-02		4.7030e-02
$\mathfrak{B}$	iter = 1	(0 0) (1 0)	iter = 1	(0 0) (1 0)	iter = 1	(0 0) (1 0)
	iter = 2	(0 1)	iter = 2	(0 1) (1 1)	iter = 2	(0 1) (1 1)
	iter = 3	(0 0 1)	iter = 3	(0 0 1) (0 1 1)	iter = 3	(0 0 1) (0 1 1) (0 2 0)
		⋮		⋮		⋮
	iter = 6	(0 0 0 0 0 1) (0 0 0 0 1 1)	iter = 6	(0 0 0 0 0 1) (0 0 0 0 1 1) (0 0 0 0 2 0) (0 0 0 1 0 1)	iter = 6	(0 0 0 0 0 1) (0 0 0 0 1 1) (0 0 0 0 2 0) (0 0 0 1 0 1) (0 0 0 1 1 1)
		⋮		⋮		⋮
	iter = 9	(0 0 0 0 0 0 0 0 1) (0 0 0 0 0 0 0 1 1) (0 0 0 0 0 0 0 2 0)	iter = 9	(0 0 0 0 0 0 0 0 1) (0 0 0 0 0 0 0 1 1) (0 0 0 0 0 0 0 2 0) (0 0 0 0 0 0 1 0 1) (0 0 0 0 0 0 1 1 1)	iter = 9	(0 0 0 0 0 0 0 0 1) (0 0 0 0 0 0 0 1 1) (0 0 0 0 0 0 0 2 0) (0 0 0 0 0 0 1 0 1) (0 0 0 0 0 0 1 1 1) (0 0 0 0 0 0 1 2 0) (0 0 0 0 0 0 2 0 1) (0 0 0 0 0 0 2 1 0)
		⋮		⋮		⋮

Table 4.3: Example 4.4.2.1: Results of adaptive procedure with the viscosity parameter $\nu = 10^{-2}$ for varying marking parameter θ_h.

$\theta_q = 0.5$	$\theta_h = 0.3$	$\theta_h = 0.5$	$\theta_h = 0.8$
# iter	102	99	80
# Total DOFs	10,171,980	10,109,880	10,110,888
# $\mathfrak{B}$	2070	2070	1444
N_d	4914	4884	7002
η_T	5.7915e-02	5.7960e-02	5.2630e-02
$\eta_h + \eta_{\theta,h}$	4.0556e-02	4.0376e-02	3.4721e-02
$\eta_q + \eta_{\theta,q}$	4.1344e-02	4.1583e-02	3.9552e-02

deviation κ.

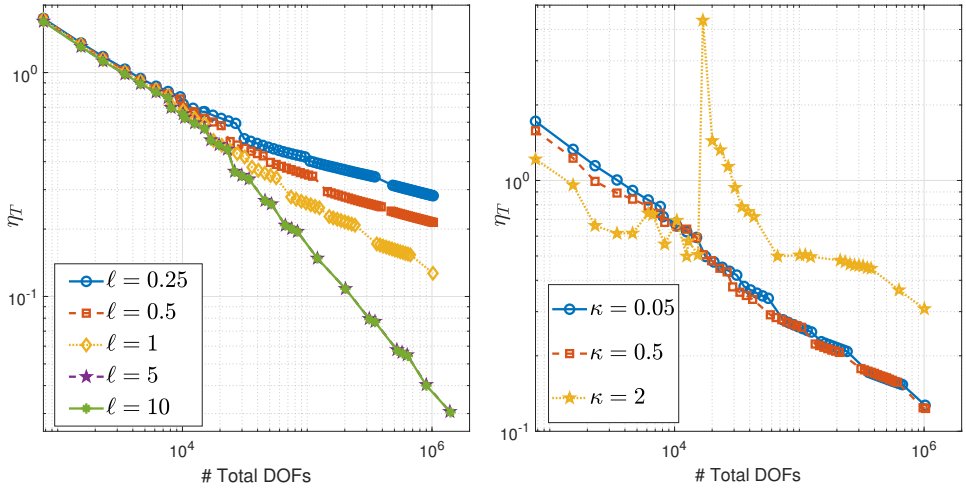

Figure 4.5: Example 4.4.2.1: Effect of the correlation length ℓ (left) and the standard deviation κ (right) on the behaviour of estimator η_T with adaptively generated spatial/parametric spaces for the viscosity parameter $\nu = 10^{-2}$.

4.4.2.2 Example with Random Convection

Our second example is a two-dimensional convection diffusion equation with random velocity in the domain $\mathcal{D} = [-1, 1]^2$. To be precise, we choose the deterministic diffusion parameter $a(\boldsymbol{x}, \omega) = \nu > 0$, the deterministic source function $f(\boldsymbol{x}) = 0.5$, and the homogeneous Dirichlet boundary condition. The random velocity field $\mathbf{b}(\boldsymbol{x}, \omega)$ is taken as

$$\mathbf{b}(\boldsymbol{x}, \omega) := (z(\boldsymbol{x}, \omega), z(\boldsymbol{x}, \omega))^T,$$

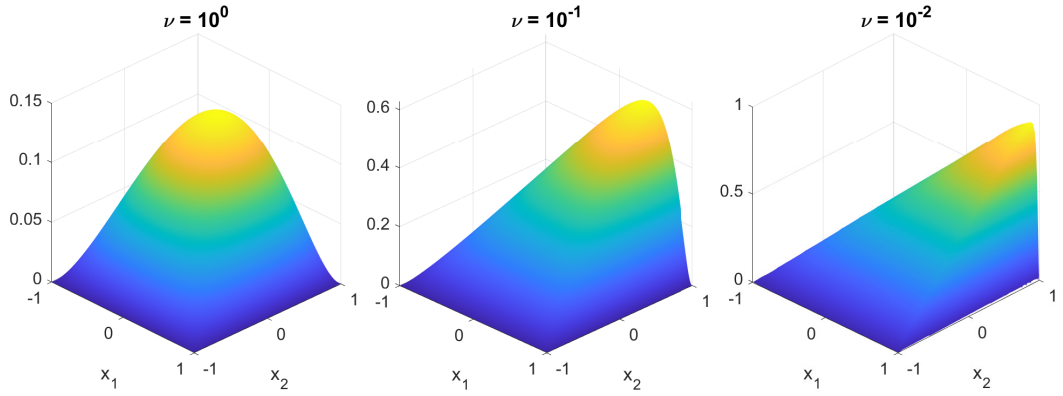

Figure 4.6: Example 4.4.2.2: Mean of SG solutions for various values of viscosity parameter ν.

where the mean of the random field is $\overline{z}(\boldsymbol{x}) = 1$. Figure 4.6 displays the mean of the computed discrete solution for various values of viscosity parameter ν. Adaptively generated triangulations obtained by Algorithm 6 for the different values of viscosity ν are displayed in Figure 4.7. As we expected, most refinements occur around the boundaries $x_1 = 1$ and $x_2 = 1$ for the smaller values of ν, where the solution exhibits boundary layers for the smaller values of ν.

Figure 4.8 shows that estimators exhibit a better convergence behavior for each value of viscosity parameter ν on the adaptively (with the marking parameters $\theta_h = 0.5$, $\theta_q = 0.5$) than the ones on uniformly generated spatial/parametric spaces. In addition, the overall convergence rate of the total estimator η_T is about $\mathcal{O}(DOF^{-1/3})$ for the smaller values of the viscosity ν, see Figure 4.9, as the previous Example 4.4.2.1. The performance of the estimator η_T on adaptively generated spatial/parametric spaces for different values of the correlation length ℓ and the standard deviation κ is given in Figure 4.10.

Table 4.4 shows results of adaptive algorithm with the viscosity parameter $\nu = 10^{-2}$ for varying marking parameter θ_q. When the value of the marking parameter θ_q is increased, we observe that the number of iterations decreases, while the size of the index set $\mathfrak{B}$ increases. However, as in the previous example, we could not obtain an optimal process for the making parameter θ_h; see, Table 4.5 and Figure 4.11.

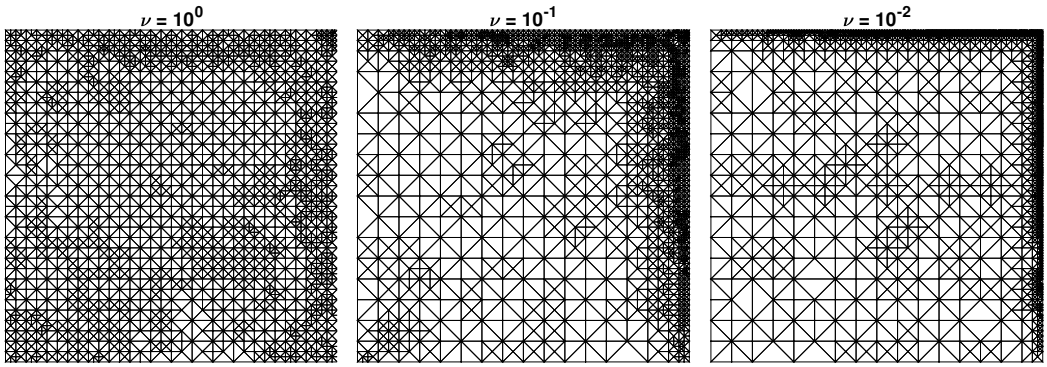

Figure 4.7: Example 4.4.2.2: Adaptively refined triangulations obtained by Algorithm 6 with the marking parameters $\theta_h = 0.5$, $\theta_q = 0.5$ for the viscosity parameter $\nu = 10^0$ with $iter = 8, N_d = 10593$ (left), $\nu = 10^{-1}$ with $iter = 30, N_d = 11385$ (middle), and $\nu = 10^{-2}$ with $iter = 65, N_d = 13062$ (right).

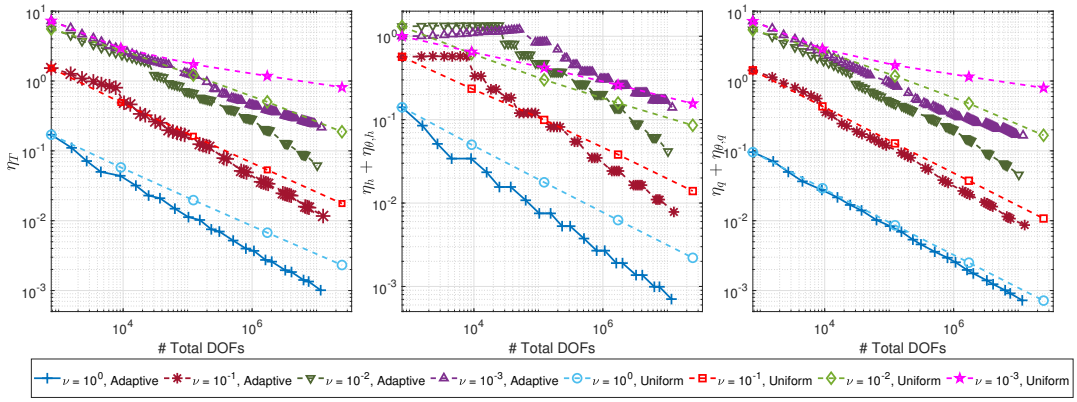

Figure 4.8: Example 4.4.2.2: Behaviour of error estimators on the adaptively and uniformly generated spatial/parametric spaces for various values of viscosity parameter ν.

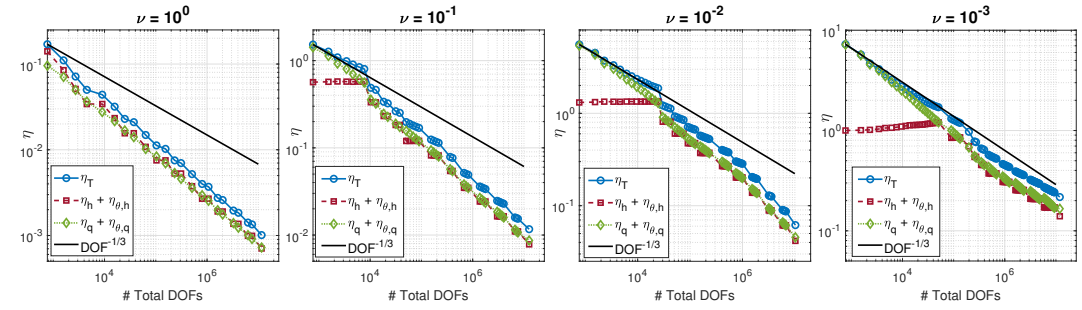

Figure 4.9: Example 4.4.2.2: Behaviour of error estimators on the adaptively generated spatial/parametric spaces with marking parameter $\theta_h = 0.5$, $\theta_q = 0.5$ for various values of viscosity parameter ν.

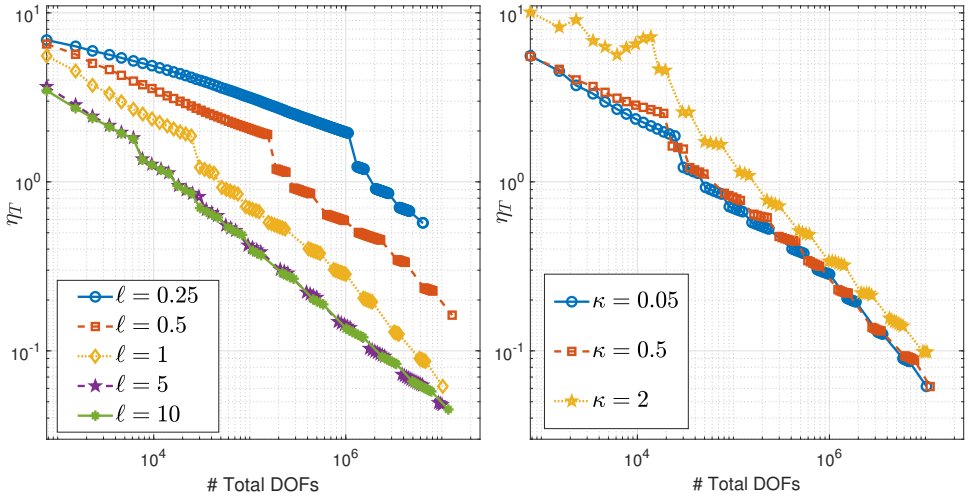

Figure 4.10: Example 4.4.2.2: Effect of the correlation length ℓ (left) and the standard deviation κ (right) on the behaviour of estimator η_T with adaptively generated spatial/parametric spaces for the viscosity parameter $\nu = 10^{-2}$.

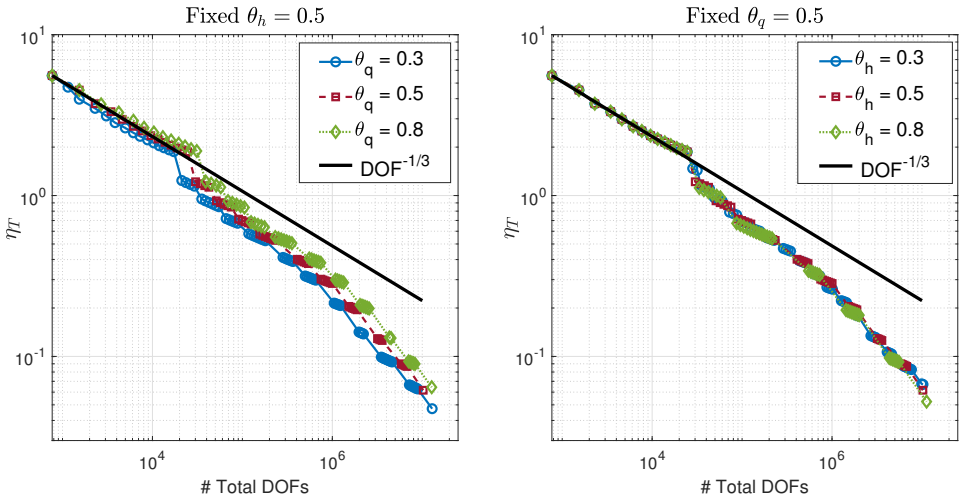

Figure 4.11: Example 4.4.2.2: Effect of the marking parameters θ_q (left) and θ_h (right) on the behaviour of estimator η_T for $\nu = 10^{-2}$.

4.5 Discussion

In this chapter, an efficient adaptive approach, based on mesh refinement or parametric enrichment, has been proposed for convection diffusion equations containing random coefficients. A parametric system of convection diffusion equations obtained by an application of stochastic Galerkin approach is discretized by using a symmetric in-

Table 4.4: Example 4.4.2.2: Results of adaptive algorithm with the viscosity parameter $\nu = 10^{-2}$ for varying marking parameter θ_q.

$\theta_h = 0.5$		$\theta_q = 0.3$		$\theta_q = 0.5$		$\theta_q = 0.8$
# iter		78		64		59
# Total DOFs		12,880,665		10,240,608		12,735,090
# $\mathfrak{B}$		721		784		1010
N_d		17865		13062		12609
η_T		4.7329e-02		6.1645e-02		6.4343e-02
$\eta_h + \eta_{\theta,h}$		3.1194e-02		4.1618e-02		4.3911e-02
$\eta_q + \eta_{\theta,q}$		3.5594e-02		4.5476e-02		4.7030e-02
$\mathfrak{B}$	iter = 1	(0 0) (1 0)	iter = 1	(0 0) (1 0)	iter = 1	(0 0) (1 0)
	iter = 2	(0 1)	iter = 2	(0 1) (1 1)	iter = 2	(0 1) (1 1)
	iter = 3	(0 0 1)	iter = 3	(0 0 1) (0 1 1)	iter = 3	(0 0 1) (0 1 1) (0 2 0)
		⋮		⋮		⋮
	iter = 6	(0 0 0 0 0 1) (0 0 0 0 1 1)	iter = 6	(0 0 0 0 0 1) (0 0 0 0 1 1) (0 0 0 0 2 0) (0 0 0 1 0 1)	iter = 6	(0 0 0 0 0 1) (0 0 0 0 1 1) (0 0 0 0 2 0) (0 0 0 1 0 1) (0 0 0 1 1 1)
		⋮		⋮		⋮
	iter = 9	(0 0 0 0 0 0 0 0 1) (0 0 0 0 0 0 0 1 1) (0 0 0 0 0 0 0 2 0)	iter = 9	(0 0 0 0 0 0 0 0 1) (0 0 0 0 0 0 0 1 1) (0 0 0 0 0 0 0 2 0) (0 0 0 0 0 0 1 0 1) (0 0 0 0 0 0 1 1 1)	iter = 9	(0 0 0 0 0 0 0 0 1) (0 0 0 0 0 0 0 1 1) (0 0 0 0 0 0 0 2 0) (0 0 0 0 0 0 1 0 1) (0 0 0 0 0 0 1 1 1) (0 0 0 0 0 0 1 2 0) (0 0 0 0 0 0 2 0 1) (0 0 0 0 0 0 2 1 0)
		⋮		⋮		⋮

Table 4.5: Example 4.4.2.2: Results of adaptive algorithm with the viscosity parameter $\nu = 10^{-2}$ for varying marking parameter θ_h.

$\theta_q = 0.5$	$\theta_h = 0.3$	$\theta_h = 0.5$	$\theta_h = 0.8$
# iter	75	64	55
# Total DOFs	10,265,400	10,240,608	11,253,450
# $\mathfrak{B}$	900	784	650
N_d	11406	13062	17313
η_T	6.6956e-02	6.1645e-02	5.2375e-02
$\eta_h + \eta_{\theta,h}$	4.6923e-02	4.1618e-02	3.5066e-02
$\eta_q + \eta_{\theta,q}$	4.7763e-02	4.5476e-02	3.8905e-02

terior penalty Galerkin (SIPG) method with upwinding for the convection term in the spatial domain. We have showed the reliability of the proposed residual-based error estimator in the energy norm contributed by the error due to the SIPG discretization, data oscillations, and the error due to stochastic Galerkin discretization. The findings in Chapter 3 shows that the numerical solutions of convection dominated problems with uncertainty are affected by some challenges such as boundary/interior layers in the solution, and adding enough number of random data in the system. In several benchmark examples, it has been shown that the optimal remedy to these problems is the combination of the adaptivity in physical space and the adaptivity in random space based on the proposed estimator.

CHAPTER 5

ROBUST DETERMINISTIC CONTROL OF CONVECTION DIFFUSION EQUATIONS WITH RANDOM COEFFICIENTS

This chapter investigates a numerical behaviour of the following robust deterministic optimal control problem

$$\min_{u\in\mathcal{U}^{ad}} \mathcal{J}(y,u) := \frac{1}{2}\|y-y^d\|_{\mathcal{X}}^2 + \frac{\gamma}{2}\|\mathbb{S}(y)\|_{\mathcal{W}}^2 + \frac{\mu}{2}\|u\|_{\mathcal{U}}^2 \tag{5.1}$$

governed by

$$\mathcal{H}\big(y(\boldsymbol{x},\omega)\big) = f(\boldsymbol{x}) + u(\boldsymbol{x}) \quad \text{in } \mathcal{D}\times\Omega, \tag{5.2a}$$

$$y(\boldsymbol{x},\omega) = y_{DB}(\boldsymbol{x}) \quad \text{on } \partial\mathcal{D}\times\Omega, \tag{5.2b}$$

where $\mathcal{H}:\mathcal{Y}\to\mathcal{Y}'$ is a linear operator that contains uncertain parameters, $\mathcal{D}\subset\mathbb{R}^2$ is a convex bounded polygonal set with a Lipschitz boundary $\partial\mathcal{D}$, and Ω is a sample space of events. In the light of the findings in Chapter 3, the stochastic Galerkin approach (SG), turning the original optimization problem containing uncertainties into a large system of deterministic problems, is applied to discretize the stochastic domain, while a discontinuous Galerkin method, namely, symmetric interior penalty Galerkin (SIPG), is preferred for the spatial discretization due to its better convergence behaviour for optimization problems governed by convection dominated PDEs; see, e.g., [111].

In (5.1), the cost functional including a risk penalization via the standard deviation $\mathbb{S}(y)$ is denoted by $\mathcal{J}(y,u)$, whose first term is a measure of the distance between the state variable y and the desired state y^d in terms of expectation of $y-y^d$. Without loss of generality, it is assumed that the state $y\in\mathcal{Y}$ is a random field, whereas the

desired state $y^d \in \mathcal{Y}$ is modeled deterministically. The second term measures the standard deviation of y, which is added since it is desirable to have a control for which the state is more accurately known, leading to a risk averse optimum. The last term corresponds to distributive deterministic control. The constant $\mu > 0$ is a regularization parameter of the control u, whereas $\gamma \geq 0$ is a risk–aversion parameter. Deterministic source function and Dirichlet boundary conditions are denoted by f and y_{DB}, respectively. It is noted that the cost functional $\mathcal{J}$ is a deterministic quantity, although it contains uncertain inputs. Further, the closed convex admissible set in the control space $\mathcal{U}$ is defined by

$$\mathcal{U}^{ad} := \{u \in \mathcal{U} : \ u_a \leq u(\boldsymbol{x}) \leq u_b, \ \ \forall \boldsymbol{x} \in \mathcal{D}\}, \tag{5.3}$$

where constants $u_a, u_b \in \mathbb{R}$ with $u_a \leq u_b$.

This chapter is organized by first discussing the existence of the solution in the next section. In Section 5.2, the problem is reduced into a finite dimensional setting via Karhunen–Loève (KL) expansion, stochastic Galerkin method, and symmetric interior penalty Galerkin method. Error analyses are done in Section 5.3. In Section 5.4, we construct the matrix formulation of the underlying optimization problem by proceeding with the optimize-then-discretize approach, and then discuss the implementation of the low–rank GMRES solver. Results of the numerical experiments are provided in Section 5.5 to illustrate the efficiency of the proposed methodology. Finally, this chapter is ended with some conclusions and discussions in Section 5.6.

5.1 Existence and Uniqueness of the Solution

For a generic random field z on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ denoted by $z(\boldsymbol{x}, \omega) : \mathcal{D} \times \Omega \to \mathbb{R}$, the mean $\mathbb{E}[z]$, the standard deviation $\mathbb{S}(z)$, and the corresponding variance $\mathbb{V}(z)$ are defined in (2.4). For simplicity and readability, the mathematical analysis will be done throughout this chapter provided that the equation of state has homogeneous boundary condition, i.e., $y_{DB} = 0$. By following the standard arguments in the deterministic setting, it can be extended to the nonhomogeneous Dirichlet boundary conditions. Recalling the tensor–product space $H^k(\mathcal{D}) \otimes L^2(\Omega)$ in (2.6),

the state and control spaces are defined as follows, respectively,

$$\mathcal{Y} := H_0^1(\mathcal{D}) \otimes L^2(\Omega) \quad \text{and} \quad \mathcal{U} := L^2(\mathcal{D}).$$

We also set $\mathcal{X} := L^2(\mathcal{D}) \otimes L^2(\Omega)$ and $\mathcal{W} = L^2(\mathcal{D})$.

In order to show the existence of the solution, it is assumed that the operator $\mathcal{H}$ satisfies the following conditions:

a) $\mathcal{H}$ is coercive such that $\mathbb{P}$-a.s., $(\mathcal{H}v, v) \geq c\|v\|_{\mathcal{X}}, \ \forall v \in \mathcal{X}$, where c is a positive constant.

b) $(\mathcal{H}u, v) = (u, \mathcal{H}^* v), \ \forall u, v \in \mathcal{X}$, where $\mathcal{H}^*$ is the adjoint of $\mathcal{H}$.

To discuss the existence and uniqueness of the optimization problem (5.1)–(5.2), we first give the following optimality definition, and existence and uniqueness theorem for the quadratic Hilbert space optimization problem.

Definition 5.1.1. *Let a function $\bar{u} \in \mathcal{U}$ be an optimal control and $\bar{y} = \bar{y}(\bar{u})$ the associated optimal state corresponding to the optimization problem* (5.1)–(5.2)*. Then, it holds, $\mathbb{P}$-a.s.,*

$$\mathcal{J}(\bar{y}, \bar{u}) \leq \mathcal{J}(y(u), u), \quad \forall u \in \mathcal{U}.$$

Theorem 5.1.2. *[142, Theorem 2.14], [22, Theorem 3.2] Assume that $\{\mathcal{K}_1, \|\cdot\|\}$ and $\{\mathcal{K}_2, \|\cdot\|\}$ are Hilbert spaces, and $\widetilde{\mathcal{K}_1} \subset \mathcal{K}_1$ is a non-empty, closed, and convex set. Let $y^d \in \mathcal{K}_1$ and the constants $\gamma, \mu \geq 0$ be given and $L : \mathcal{K}_1 \to \mathcal{K}_2$ be a continuous linear operator. Then, the quadratic Hilbert space optimization problem*

$$\min_{u \in \widetilde{\mathcal{K}_1}} \ f(u) := \frac{1}{2}\|Lu - y^d\|_{\mathcal{K}_2}^2 + \frac{\gamma}{2}\|\mathbb{S}(Lu)\|_{\mathcal{K}_2}^2 + \frac{\mu}{2}\|u\|_{\mathcal{K}_1}^2$$

admits, $\mathbb{P}$-a.s., an optimal solution $\bar{u} \in \widetilde{\mathcal{K}_1}$. If $\mu > 0$, then $\bar{u}$ is uniquely determined.

Proof. Note first that the function $f(u) \geq 0$ is continuous and convex. Therefore, the proof follows analogously to that of [142, Theorem 2.14]. □

If setting $\widetilde{\mathcal{K}_1} = \mathcal{U}^{ad}$, $\mathcal{K}_1 = \mathcal{U}$, and $\mathcal{K}_2 = \mathcal{Y}$, denoting continuous linear operator $L : u \to y(u)$, and substituting L into the cost functional $\mathcal{J}(y, u)$ in (5.1), then the

following quadratic minimization problem is obtained in the Hilbert space $\mathcal{U}$:

$$\min\ \mathcal{J}(u) := \frac{1}{2}\mathbb{E}\int_{\mathcal{D}}\big(Lu - y^d\big)^2\ dx + \frac{\gamma}{2}\mathbb{E}\int_{\mathcal{D}}[Lu - (\mathbb{E}[Lu])]^2\ dx + \frac{\mu}{2}\mathbb{E}\int_{\mathcal{D}}u^2\,dx. \tag{5.4}$$

With the definitions above, $\mathcal{Y}$ and $\mathcal{U}$ are Hilbert spaces, the functional $\mathcal{J}$ is strictly convex, and the admissible set $\mathcal{U}^{ad}$ is a closed and convex set. Then, by Theorem 5.1.2, the optimization problem (5.1)–(5.2) has a unique solution. Next, we also state the relation between the optimal solution and the variational inequality.

Theorem 5.1.3. *([113, Theorem 1.3: Lion's Lemma]) Let the cost functional $\mathcal{J}(v)$ be strictly convex and differentiable. Then, a unique optimal control $\bar{u} \in \mathcal{U}$ exists if and only if the variational inequality holds*

$$\mathcal{J}'(\bar{u}) \cdot (v - \bar{u}) \geq 0, \quad \forall v \in \mathcal{U}^{ad}, \tag{5.5}$$

where

$$\mathcal{J}'(\bar{u}) \cdot w := \lim_{h \to 0} \frac{\mathcal{J}(\bar{u} + hw) - \mathcal{J}(\bar{u})}{h} \tag{5.6}$$

is the directional derivative of $\mathcal{J}$ with respect to u in the direction of w.

Now, the first order optimality system of the optimization problem containing uncertain coefficients (5.1)–(5.2) can be derived.

Theorem 5.1.4. *A pair (y, u) is a unique solution of the optimization problem* (5.1)–(5.2) *if and only if there exists an adjoint $p \in \mathcal{Y}$ such that the optimality system holds, $\mathbb{P}$-a.s., for the triplet $(y(u), u, p(u)) \in \mathcal{Y} \times \mathcal{U}^{ad} \times \mathcal{Y}$*

$$\mathcal{H}\big(y(u)\big) = f(\boldsymbol{x}) + u(\boldsymbol{x}), \tag{5.7a}$$

$$\mathcal{H}^*\big(p(u)\big) = y(u) - y^d + \gamma\big(y(u) - \mathbb{E}[y(u)]\big), \tag{5.7b}$$

$$\big(\mathbb{E}[p(u)] + \mu u, v - u\big) \geq 0, \qquad v \in \mathcal{U}^{ad}. \tag{5.7c}$$

Proof. Rewrite the objective functional $\mathcal{J}$ as

$$\begin{aligned}\mathcal{J}(u) \;=\; & \underbrace{\frac{1}{2}\mathbb{E}\left[\int_{\mathcal{D}}\big(y(u) - y^d\big)^2\ d\boldsymbol{x}\right]}_{\mathcal{J}_1(u)} + \underbrace{\frac{\gamma}{2}\mathbb{E}\left[\int_{\mathcal{D}} y(u)^2\,d\boldsymbol{x}\right]}_{\mathcal{J}_2(u)} \\ & \underbrace{-\frac{\gamma}{2}\int_{\mathcal{D}}\left(\mathbb{E}\left[y(u)\right]\right)^2\ d\boldsymbol{x}}_{\mathcal{J}_3(u)} + \underbrace{\frac{\mu}{2}\int_{\mathcal{D}} u^2\,d\boldsymbol{x}}_{\mathcal{J}_4(u)}.\end{aligned}$$

By the definition of directional derivative (5.6), it is obtained that

$$
\begin{aligned}
\mathcal{J}_1'(u)\cdot(v-u) &= \lim_{h\to 0^+} \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(y(u+h(v-u))-y^d\right)^2 \, d\boldsymbol{x}\right]}{2h} \\
&\quad - \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(y(u)-y^d\right)^2 \, d\boldsymbol{x}\right]}{2h} \\
&= \lim_{h\to 0^+} \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(y(u+h(v-u))^2-(y^d)^2\right) \, d\boldsymbol{x}\right]}{2h} \\
&\quad - \frac{\mathbb{E}\left[\int_{\mathcal{D}} 2\left(y(u+h(v-u))-y(u)\right) y^d \, d\boldsymbol{x}\right]}{2h} \\
&= \mathbb{E}\left[\int_{\mathcal{D}} \big(y(u)-y^d\big) y'(u)\cdot(v-u)\, d\boldsymbol{x}\right], \\
\mathcal{J}_2'(u)\cdot(v-u) &= \gamma \lim_{h\to 0^+} \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(y(u+h(v-u))\right)^2 d\boldsymbol{x}\right] - \mathbb{E}\left[\int_{\mathcal{D}} y(u)^2 \, d\boldsymbol{x}\right]}{2h} \\
&= \gamma\mathbb{E}\left[\int_{\mathcal{D}} y(u) y'(u)\cdot(v-u)\, d\boldsymbol{x}\right], \\
\mathcal{J}_3'(u)\cdot(v-u) &= \gamma \lim_{h\to 0^+} \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(\mathbb{E}\left[y(u+h(v-u))\right]\right)^2 d\boldsymbol{x}\right]}{2h} \\
&\quad - \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(\mathbb{E}\left[y(u)\right]\right)^2 d\boldsymbol{x}\right]}{2h} \\
&= \gamma\mathbb{E}\left[\int_{\mathcal{D}} \mathbb{E}[y(u)] y'(u)\cdot(v-u)\, d\boldsymbol{x}\right], \\
\mathcal{J}_4'(u)\cdot(v-u) &= \mu \lim_{h\to 0^+} \frac{\mathbb{E}\left[\int_{\mathcal{D}} (u+h(v-u))^2 \, d\boldsymbol{x}\right] - \mathbb{E}\left[\int_{\mathcal{D}} u^2 \, d\boldsymbol{x}\right]}{2h} \\
&= \mu \lim_{h\to 0^+} \frac{\mathbb{E}\left[\int_{\mathcal{D}} \left(h^2(v-u)^2 + 2hu(v-u)\right) d\boldsymbol{x}\right]}{2h} \\
&= \mu\mathbb{E}\left[\int_{\mathcal{D}} u\cdot(v-u)\, d\boldsymbol{x}\right].
\end{aligned}
$$

Hence, by combining all terms, it holds

$$
\begin{aligned}
&\mathcal{J}'(u)\cdot(v-u) \\
&\quad = \mathbb{E}\left[\int_{\mathcal{D}} \big(y(u)-y^d\big) y'(u)\cdot(v-u)\, d\boldsymbol{x}\right] + \gamma\mathbb{E}\left[\int_{\mathcal{D}} y(u) y'(u)\cdot(v-u)\, d\boldsymbol{x}\right] \\
&\qquad - \gamma\int_{\mathcal{D}} \mathbb{E}[y(u)] y'(u)\cdot(v-u)\, d\boldsymbol{x} + \mu\int_{\mathcal{D}} u\cdot(v-u)\, d\boldsymbol{x}. \qquad (5.8)
\end{aligned}
$$

By well–posedness of the state equation (5.2) followed from the Lax–Milgram lemma, one can easily show that the operator $\mathcal{H}$ is invertible so that, by taking directional

derivative, one gets

$$\begin{aligned} y'(u)\cdot(v-u) &= \lim_{h\to 0}\frac{y\left(u+h(v-u)\right)-y(u)}{h} \\ &= \lim_{h\to 0}\frac{\mathcal{H}^{-1}\left(f+u+h(v-u)\right)-\mathcal{H}^{-1}\left(f+u\right)}{h} \\ &= \lim_{h\to 0}\frac{h\mathcal{H}^{-1}\left(v-u\right)}{h} = \mathcal{H}^{-1}(v-u) = \mathcal{H}^{-1}(v+f-f-u) \\ &= \mathcal{H}^{-1}(f+v)-\mathcal{H}^{-1}(f+u) = y(v)-y(u). \end{aligned}$$

Thus, (5.8) gives us

$$\mathcal{J}'(u)\cdot(v-u) = \Psi(\gamma)+\mu\int_{\mathcal{D}} u\cdot(v-u)\,d\boldsymbol{x}, \tag{5.9}$$

where

$$\begin{aligned} \Psi(\gamma) = (1+\gamma)\mathbb{E}\left[\int_{\mathcal{D}} y(u)\cdot\big(y(v)-y(u)\big)\,d\boldsymbol{x}\right] - \gamma\int_{\mathcal{D}}\mathbb{E}[y(u)]\cdot\big(y(v)-y(u)\big)\,d\boldsymbol{x} \\ -\mathbb{E}\left[\int_{\mathcal{D}} y^d\cdot\big(y(v)-y(u)\big)\,d\boldsymbol{x}\right]. \end{aligned}$$

To guarantee the existence and uniqueness of the solution from Theorem 5.1.3, the following requirement is needed

$$\mathcal{J}'(u)\cdot(v-u)\geq 0. \tag{5.10}$$

We note that the adjoint state $p(u)\in\mathcal{Y}$ is introduced by

$$\mathcal{H}^*\big(p(u)\big) = y(u)-y^d+\gamma\big(y(u)-\mathbb{E}[y(u)]\big). \tag{5.11}$$

Multiplying both sides of (5.11) by $\big(y(v)-y(u)\big)$, integrating over $\mathcal{D}$, and taking the expectation of the resulting system, we obtain

$$\begin{aligned} \mathbb{E}\left[\int_{\mathcal{D}}\mathcal{H}^*\big(p(u)\big)\cdot\big(y(v)-y(u)\big)\,d\boldsymbol{x}\right] &= \mathbb{E}\left[\int_{\mathcal{D}} p(u)\cdot\big(\mathcal{H}\big(y(v)\big)-\mathcal{H}\big(y(u)\big)\big)\,d\boldsymbol{x}\right] \\ &= \mathbb{E}\left[\int_{\mathcal{D}} p(u)\cdot\big(v-u\big)\,d\boldsymbol{x}\right] \\ &= \Psi(\gamma). \end{aligned} \tag{5.12}$$

Inserting (5.12) into (5.9) and combining with (5.10) yield

$$\mathcal{J}'(u)\cdot(v-u) = \big(\mathbb{E}[p(u)]+\mu u, v-u\big)\geq 0, \tag{5.13}$$

which is the desired result. □

In this chapter, $\mathcal{H}$ is considered as the convection–diffusion operator

$$\mathcal{H} := -\nabla \cdot \big(a(\boldsymbol{x},\omega)\nabla\big) + \mathbf{b}(\boldsymbol{x},\omega)\cdot\nabla, \tag{5.14}$$

which turns the state equation (5.2) into

$$-\nabla \cdot \big(a(\boldsymbol{x},\omega)\nabla y\big) + \mathbf{b}(\boldsymbol{x},\omega)\cdot\nabla y = f + u \qquad \text{in } \mathcal{D}\times\Omega, \tag{5.15a}$$

$$y = y_{DB} \qquad \text{on } \partial\mathcal{D}\times\Omega, \tag{5.15b}$$

where $a : (\mathcal{D}\times\Omega) \to \mathbb{R}$ and $\mathbf{b} : (\mathcal{D}\times\Omega) \to \mathbb{R}^2$ are random diffusivity and velocity coefficients, respectively, which is assumed to have continuous and bounded covariance functions. In addition, we recall the following assumptions on the uncertain coefficients defined in Section 3.1:

i) $\exists\ a_{\min}, a_{\max}$ such that for almost every (a.e.) $(\boldsymbol{x},\omega)\in\mathcal{D}\times\Omega$,

$$0 < a_{\min} \le a(\boldsymbol{x},\omega) \le a_{\max} < \infty.$$

In addition, $a(\boldsymbol{x},\omega)$ has a uniformly bounded and continuous first derivatives.

ii) The velocity coefficient $\mathbf{b}$ satisfies $\mathbf{b}(\cdot,\omega) \in \big(L^\infty(\overline{\mathcal{D}})\big)^2$ for a.e. $\omega\in\Omega$ and $\nabla\cdot\mathbf{b}(\boldsymbol{x},\omega) = 0$.

Then, the well–posedness of the state equation (5.15) can be shown by following the classical Lax–Milgram lemma; see, e.g., [11, 117].

Now, the corresponding weak formulation of the optimization problem containing uncertainty (5.1)–(5.2) is given as follows:

$$\begin{aligned}\min_{u\in\mathcal{U}^{ad}} \mathcal{J}(u) \;&=\; \frac{1}{2}\mathbb{E}\left[\int_{\mathcal{D}} \big(y(u)-y^d\big)^2\, d\boldsymbol{x}\right] + \frac{\gamma}{2}\mathbb{E}\left[\int_{\mathcal{D}} \big(y(u)-\mathbb{E}[y(u)]\big)^2\, d\boldsymbol{x}\right] \\ &\quad + \frac{\mu}{2}\int_{\mathcal{D}} u^2\, d\boldsymbol{x}\end{aligned} \tag{5.16}$$

governed by

$$a[y,v] + b[u,v] = [f,v], \qquad v\in\mathcal{Y}, \tag{5.17}$$

where

$$\begin{aligned} a[y,v] \;&=\; \mathbb{E}\left[\int_{\mathcal{D}} \big(a(\boldsymbol{x},\omega)\nabla y\cdot\nabla v + \mathbf{b}(\boldsymbol{x},\omega)\cdot\nabla y\, v\big)\, d\boldsymbol{x}\right], \quad \forall y, v\in\mathcal{Y}, \\ b[u,v] \;&=\; -\mathbb{E}\left[\int_{\mathcal{D}} uv\, d\boldsymbol{x}\right] \quad \text{and} \quad [f,v] = \mathbb{E}\left[\int_{\mathcal{D}} fv\, d\boldsymbol{x}\right], \quad \forall u\in\mathcal{U}, v\in\mathcal{Y}. \end{aligned}$$

Moreover, the optimality system in (5.7) can be stated in the weak formulation as follows:

$$a[y, \upsilon] + b[u, \upsilon] = [f, \upsilon], \qquad \upsilon \in \mathcal{Y}, \tag{5.19a}$$

$$a[q, p] = [y - y^d, q] + \gamma\big[y - \mathbb{E}[y], q\big], \qquad q \in \mathcal{Y}, \tag{5.19b}$$

$$\big(\mathbb{E}[p] + \mu u, w - u\big) \geq 0, \qquad w \in \mathcal{U}^{ad}, \tag{5.19c}$$

where the adjoint $p \in \mathcal{Y}$ solves the following convection diffusion equation having negative convection term containing uncertain inputs:

$$-\nabla \cdot \big(a(\boldsymbol{x}, \omega)\nabla p\big) - \mathbf{b}(\boldsymbol{x}, \omega) \cdot \nabla p = (y - y^d) + \gamma\big(y - \mathbb{E}[y]\big) \text{ in } \mathcal{D} \times \Omega, \tag{5.20a}$$

$$p = 0 \qquad \text{on } \partial\mathcal{D} \times \Omega. \tag{5.20b}$$

In the following, the techniques, that is, Karhunen–Loève (KL) expansion (2.16), stochastic Galerkin, and discontinuous Galerkin method, will be summarized to recast the infinite–dimensional model problem (5.16)–(5.17) into the finite dimensional problem.

5.2 Stochastic Galerkin Discretization

To solve (5.16)–(5.17) numerically, it is needed to reduce the stochastic process into finite mutually uncorrelated random variables. Therefore, the coefficients $a(\boldsymbol{x}, \omega)$ and $\mathbf{b}(\boldsymbol{x}, \omega)$ are approximated by the truncated KL expansion (2.16), which is a finite representation of the random field in the sense that the mean-square error of approximation is minimized; see, e.g., [8]. To guarantee the positivity of the truncated KL expansion (2.16) for the diffusivity coefficient $a(\boldsymbol{x}, \omega)$, it is also assumed that the mean of random coefficient exhibits a stronger dominance; see, e.g., [128].

By the assumption on the finite dimensional in Section 2.16 and Doob–Dynkin lemma [127], the solution of (5.17) can be expressed in the finite dimensional stochastic space, that means, $y(\boldsymbol{x}, \xi(\omega)) \in \mathcal{Y}_\rho = L^2(H_0^1(\mathcal{D}); \Gamma)$ with $\xi = \big(\xi_1(\omega), \ldots, \xi_N(\omega)\big)$.

Then, setting $\widetilde{\mathbb{E}}[y] = \int_\Gamma y\,\rho(\xi)d\xi$, the optimization problem (5.16)-(5.17) becomes

$$\begin{aligned}\min_{u\in\mathcal{U}^{ad}} \mathcal{J}(u) &= \frac{1}{2}\widetilde{\mathbb{E}}\left[\int_{\mathcal{D}}\left(y(u)-y^d\right)^2 d\boldsymbol{x}\right] + \frac{\gamma}{2}\widetilde{\mathbb{E}}\left[\int_{\mathcal{D}}\Big(y(u)-\widetilde{\mathbb{E}}[y(u)]\Big)^2 d\boldsymbol{x}\right] \\ &\quad + \frac{\mu}{2}\int_{\mathcal{D}} u^2\, d\boldsymbol{x} \end{aligned} \tag{5.21}$$

subject to

$$a[y,\upsilon]_\rho + b[u,\upsilon]_\rho = [f,\upsilon]_\rho, \quad \forall \upsilon \in \mathcal{Y}_\rho, \tag{5.22}$$

where

$$a[y,\upsilon]_\rho = \int_\Gamma\int_{\mathcal{D}}\big(a(\boldsymbol{x},\xi)\nabla y\cdot\nabla\upsilon + \mathbf{b}(\boldsymbol{x},\xi)\cdot\nabla y\,\upsilon\big)\,d\boldsymbol{x}\,\rho(\xi)d\xi,\ \forall y,\upsilon\in\mathcal{Y}_\rho, \tag{5.23a}$$

$$b[u,\upsilon]_\rho = -\int_\Gamma\int_{\mathcal{D}} u\upsilon\, d\boldsymbol{x}\,\rho(\xi)d\xi, \qquad \forall u\in\mathcal{U},\ \upsilon\in\mathcal{Y}_\rho, \tag{5.23b}$$

$$[f,\upsilon]_\rho = \int_\Gamma\int_{\mathcal{D}} f\upsilon\, d\boldsymbol{x}\,\rho(\xi)d\xi, \qquad \forall \upsilon\in\mathcal{Y}_\rho. \tag{5.23c}$$

Then, the optimization problem (5.21)-(5.22) has a unique solution pair $(y,u)\in\mathcal{Y}_\rho\times\mathcal{U}^{ad}$ if and only if there is an adjoint $p\in\mathcal{Y}_\rho$ such that the following optimality system holds for the triplet (y,u,p):

$$a[y,\upsilon]_\rho + b[u,\upsilon]_\rho = [f,\upsilon]_\rho, \qquad \upsilon\in\mathcal{Y}_\rho, \tag{5.24a}$$

$$a[q,p]_\rho = [y-y^d,q]_\rho + \gamma\big[y-\widetilde{\mathbb{E}}[y],q\big]_\rho, \qquad q\in\mathcal{Y}_\rho, \tag{5.24b}$$

$$\big(\widetilde{\mathbb{E}}[p]+\mu u, w-u\big)\geq 0, \qquad w\in\mathcal{U}^{ad}. \tag{5.24c}$$

Next, the state solution $y(\boldsymbol{x},\xi)\in L^2(\Gamma,\mathcal{F},\mathbb{P})$, as well as the adjoint solution $p(\boldsymbol{x},\xi)\in L^2(\Gamma,\mathcal{F},\mathbb{P})$, can be represented by a finite generalized polynomial chaos (gPC) approximation (2.23) as stated in Cameron–Martin theorem [40],

$$y(\boldsymbol{x},\omega)\approx y_J(\boldsymbol{x},\xi) = \sum_{i=0}^{J-1} y_i(\boldsymbol{x})\Psi_i(\xi), \tag{5.25a}$$

$$p(\boldsymbol{x},\omega)\approx p_J(\boldsymbol{x},\xi) = \sum_{i=0}^{J-1} p_i(\boldsymbol{x})\Psi_i(\xi), \tag{5.25b}$$

where $y_i(\boldsymbol{x})$ and $p_i(\boldsymbol{x})$ are the deterministic modes of the expansion and J is the total number of PC basis given in (2.24).

For simplicity, we only deal with the state equation since the procedure for the adjoint equation is similar to the state one. By inserting KL expansions (2.16) of the

diffusion $a(\boldsymbol{x},\omega)$ and the convection $\mathbf{b}(\boldsymbol{x},\omega)$ coefficients, and the solution expression (5.25) into the variational form of the state equation (5.22) and projecting onto the space of the PC basis functions, and applying the SIPG discretization defined in Section 3.1.1, the following (bi)–linear forms of the stochastic discontinuous Galerkin (SDG) method is obtained for the state equation correspond to

$$a_\xi[y,v] + b_\xi[u,v] = [f,v]_\xi,$$

where

$$\begin{aligned} a_\xi[y,v] &= \int_\Gamma a_h(y,v,\xi)\rho(\xi)\,d\xi, \quad [f,v]_\xi = \int_\Gamma l_h(f,v,\xi)\rho(\xi)\,d\xi, \\ b_\xi[u,v] &= \int_\Gamma b_h(u,v,\xi)\rho(\xi)\,d\xi, \quad \text{with} \quad b_h(u,v,\xi) = -\sum_{K\in\mathcal{T}_h}\int_K uv\,d\boldsymbol{x}. \end{aligned}$$

Here, $a_h(\cdot,\cdot,\xi)$ and $l_h(\cdot,\cdot,\xi)$ correspond to the bilinear form in (3.9) and linear form in (3.10). Now, the discrete optimal control problem is stated as

$$\begin{aligned} \min_{u_h\in\mathcal{U}_h^{ad}} \mathcal{J}(u_h) \;=\; & \frac{1}{2}\widetilde{\mathbb{E}}\left[\int_{\mathcal{D}}\left(y_h - y^d\right)^2 d\boldsymbol{x}\right] + \frac{\gamma}{2}\widetilde{\mathbb{E}}\left[\int_{\mathcal{D}}\left(y_h - \widetilde{\mathbb{E}}[y_h]\right)^2 d\boldsymbol{x}\right] \\ & + \frac{\mu}{2}\int_{\mathcal{D}} u_h^2\,d\boldsymbol{x} \end{aligned} \tag{5.26}$$

governed by

$$a_\xi[y_h,\upsilon_h] + b_\xi[u_h,\upsilon_h] = [f,\upsilon_h]_\xi, \quad \forall \upsilon_h \in \mathcal{Y}_h = V_h \otimes \mathcal{S}_k^q, \tag{5.27}$$

where the discrete admissible set (5.3) is defined by

$$\mathcal{U}_h^{ad} := \{u_h \in \mathcal{U}_h \;:\; u_a \le u_h(\boldsymbol{x}) \le u_b, \text{ a.e. } \boldsymbol{x} \in K \subset \mathcal{T}_h\}, \tag{5.28}$$

with $\mathcal{U}_h^{ad} = \mathcal{U}_h \cap \mathcal{U}^{ad}$ and $\mathcal{U}_h = \mathcal{V}_h$.

Analogously, a pair $(y_h,u_h) \in \mathcal{Y}_h \times \mathcal{U}_h^{ad}$ is a unique solution of the control problem (5.26)-(5.27) if and only if an adjoint $p_h \in \mathcal{Y}_h$ exists such that the optimality system holds for $(y_h,u_h,p_h) \in \mathcal{Y}_h \times \mathcal{U}_h^{ad} \times \mathcal{Y}_h$

$$a_\xi[y_h,\upsilon_h] + b_\xi[u_h,\upsilon_h] = [f,\upsilon_h]_\xi, \qquad \upsilon_h \in \mathcal{Y}_h, \tag{5.29a}$$

$$a_\xi[q_h,p_h] = [y_h - y^d, q_h]_\xi + \gamma\left[y_h - \widetilde{\mathbb{E}}[y_h], q_h\right]_\xi, \qquad q_h \in \mathcal{Y}_h, \tag{5.29b}$$

$$[p_h + \mu u_h, w_h - u_h]_\xi \ge 0, \qquad w_h \in \mathcal{U}_h^{ad}, \tag{5.29c}$$

where $[p_h + \mu u_h, w_h - u_h]_\xi = \big(\widetilde{\mathbb{E}}[p_h] + \mu u_h, w_h - u_h\big)$ since the discrete solution u_h is deterministic.

Further, by denoting

$$\mathcal{J}_h'(u_h) \cdot w_h = [p_h + \mu u_h, w_h]_\xi, \quad \forall w_h \in \mathcal{U}_h^{ad}, \tag{5.30}$$

one can easily obtain the following expression for the discrete directional derivative of functional $\mathcal{J}_h(u_h)$:

$$\mathcal{J}_h'(u_h) \cdot (w_h - u_h) \geq 0, \quad \forall w_h \in \mathcal{U}_h^{ad}. \tag{5.31}$$

5.3 Error Analysis

This section provides an a priori error analysis of the optimization problem (5.16)-(5.17), discretized by the stochastic discontinuous Galerkin method in the energy norm (3.12). First, we introduce L^2–projection operator on the finite dimensional probability domain Γ, and H^1–projection operator and L^2–projection operator on the spatial domain $\mathcal{D}$.

- L^2–projection operators $\Pi_q : L^2(\Gamma) \to \mathcal{S}_k^q$ and $\Pi_h : L^2(\mathcal{D}) \to V_h \cap L^2(\mathcal{D})$ are given by

$$\begin{aligned} (\Pi_q(\xi) - \xi, \zeta)_{L^2(\Gamma)} &= 0, \qquad \forall \zeta \in \mathcal{S}_k^q, \qquad \forall \xi \in L^2(\Gamma), \quad (5.32a) \\ (\Pi_h(\nu) - \nu, \chi)_{L^2(\mathcal{D})} &= 0, \qquad \forall \chi \in V_h, \qquad \forall \nu \in L^2(\mathcal{D}) \quad (5.32b) \end{aligned}$$

  with the following estimate

$$\|\nu - \Pi_h(\nu)\|_{L^2(L^2(\mathcal{D};\Gamma))} \leq Ch\|\nu\|_{L^2(H^1(\mathcal{D};\Gamma))}. \tag{5.33}$$

  In addition, taking $\zeta = \Pi_q(\xi)$ and $\chi = \Pi_h(\nu)$ in (5.32a) and (5.32b), respectively, it holds that

$$\begin{aligned} \|\Pi_q(\xi)\|_{L^2(\Gamma)} &\leq C\|\xi\|_{L^2(\Gamma)}, \qquad \forall \xi \in L^2(\Gamma), \quad (5.34a) \\ \|\Pi_h(\nu)\|_{L^2(\mathcal{D})} &\leq C\|\nu\|_{L^2(\mathcal{D})}, \qquad \forall \nu \in L^2(\mathcal{D}). \quad (5.34b) \end{aligned}$$

- H^1–projection operator $\mathcal{R}_h : H^1(\mathcal{D}) \to V_h \cap H^1(\mathcal{D})$ is stated by

$$(\mathcal{R}_h(\upsilon) - \upsilon, \vartheta)_{L^2(\mathcal{D})} = 0, \qquad \forall \vartheta \in V_h, \qquad \forall \upsilon \in H^1(\mathcal{D}), \tag{5.35a}$$

$$(\nabla(\mathcal{R}_h(\upsilon) - \upsilon), \nabla\vartheta)_{L^2(\mathcal{D})} = 0, \qquad \forall \vartheta \in V_h, \qquad \forall \upsilon \in H^1(\mathcal{D}). \tag{5.35b}$$

With the help of the H^1–projection operator in (5.35a), the Cauchy–Schwarz inequality (2.11), the L^2–projection operator in (5.32a), and the approximation in (3.22), we have the following approximation property ([41, Theorem 3.2]): for all $v \in L^2(H^2(\mathcal{D});\Gamma) \cap H^{q+1}(H^1(\mathcal{D});\Gamma)$, and $\widetilde{v} \in V_h \times \mathcal{S}_k^q$

$$\begin{aligned} \|v - \widetilde{v}\|_{L^2(H^1(\mathcal{D});\Gamma)} \;\leq\; & Ch\|v\|_{L^2(H^2(\mathcal{D});\Gamma)} \\ & + \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{\xi_n}^{q_n+1} v\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}, \end{aligned} \tag{5.36}$$

where the constant C does not depend on v, h, and k_n.

To recognize error contributions emerging from the spatial domain $\mathcal{D}$ and the probability domain Γ, separately, a projection operator $\mathcal{P}_{hq}$ mapping onto the tensor product space $\mathcal{Y}_h$ is given by

$$\mathcal{P}_{hq}\Upsilon = \Pi_h\Pi_q\Upsilon = \Pi_q\Pi_h\Upsilon, \quad \forall \Upsilon \in L^2(L^2(\mathcal{D});\Gamma), \tag{5.37}$$

and we decompose as follows:

$$\Upsilon - \mathcal{P}_{hq}\Upsilon = (\Upsilon - \Pi_h\Upsilon) + \Pi_h(I - \Pi_q)\Upsilon, \quad \forall \Upsilon \in L^2(L^2(\mathcal{D});\Gamma). \tag{5.38}$$

Then, it follows from (5.34a), (5.34b), and (5.37) that

$$\|\mathcal{P}_{hq}\Upsilon\|_{L^2(L^2(\mathcal{D});\Gamma)} \leq C\|\Upsilon\|_{L^2(L^2(\mathcal{D});\Gamma)}, \quad \forall \Upsilon \in L^2(L^2(\mathcal{D});\Gamma). \tag{5.39}$$

Before the derivation of a priori error estimate, the following auxiliary problem is stated as

$$\mathcal{J}_h'(u) \cdot (w - u) = [p_h(u) + \mu u, w - u]_\xi \geq 0, \quad \forall w \in \mathcal{U}^{ad}, \tag{5.40}$$

where $p_h(u) \in \mathcal{Y}_h$ solves the following auxiliary system:

$$a_\xi[y_h(u), v_h] + b_\xi[u, v_h] = [f, v_h]_\xi, \qquad v_h \in \mathcal{Y}_h, \tag{5.41a}$$

$$a_\xi[q_h, p_h(u)] = [y_h(u) - y^d, q_h]_\xi + \gamma\big[y_h(u) - \widetilde{\mathbb{E}}[y_h(u)], q_h\big]_\xi, \; q_h \in \mathcal{Y}_h. \tag{5.41b}$$

It is also noted that we prefer to use $\|u\|_{L^2(L^2(\mathcal{D});\Gamma)}$ in the derivation of error estimates instead of $\|u\|_{L^2(\mathcal{D})}$ for better readability in terms of notation.

Lemma 5.3.1. *With the definition in* (5.40)*, the following estimate holds:*

$$(\mathcal{J}_h'(w) - \mathcal{J}_h'(u)) \cdot (w - u) \geq \mu \|w - u\|^2_{L^2(L^2(\mathcal{D});\Gamma)}. \tag{5.42}$$

Proof. By (5.40), we have

$$\big(\mathcal{J}_h'(w) - \mathcal{J}_h'(u)\big) \cdot (w - u) = [p_h(w) - p_h(u), w - u]_\xi + \mu[w - u, w - u]_\xi. \tag{5.43}$$

Now, it follows from (5.41) that

$$\begin{aligned}
[p_h(w) - p_h(u), w - u]_\xi &= a_\xi[y_h(w) - y_h(u), p_h(w) - p_h(u)] \\
&= (1 + \gamma)[y_h(w) - y_h(u), y_h(w) - y_h(u)]_\xi \\
&\quad - \gamma\big[\widetilde{\mathbb{E}}[y_h(w) - y_h(u)], y_h(w) - y_h(u)\big]_\xi.
\end{aligned} \tag{5.44}$$

The usage of Cauchy-Schwarz (2.11) and Young's inequalities (2.12) yields

$$\begin{aligned}
&- \gamma\big[\widetilde{\mathbb{E}}[y_h(w) - y_h(u)], y_h(w) - y_h(u)\big]_\xi \\
&\quad \geq -\frac{\gamma}{2}\|\widetilde{\mathbb{E}}[y_h(w) - y_h(u)]\|^2_{L^2(L^2(\mathcal{D});\Gamma)} - \frac{\gamma}{2}\|y_h(w) - y_h(u)\|^2_{L^2(L^2(\mathcal{D});\Gamma)}.
\end{aligned}$$

Since all norms are convex functions, Jensen's inequality (2.14) and $\widetilde{\mathbb{E}}[\widetilde{\mathbb{E}}[u]] = \widetilde{\mathbb{E}}[u]$ give us

$$-\gamma\big[\widetilde{\mathbb{E}}[y_h(w) - y_h(u)], y_h(w) - y_h(u)\big]_\xi \geq -\gamma\|y_h(w) - y_h(u)\|^2_{L^2(L^2(\mathcal{D});\Gamma)}. \tag{5.45}$$

Thus, inserting (5.45) into (5.44), it is obtained that

$$[p_h(w) - p_h(u), w - u]_\xi \geq \underbrace{\|y_h(w) - y_h(u)\|^2_{L^2(L^2(\mathcal{D});\Gamma)}}_{\geq 0}. \tag{5.46}$$

Hence, (5.43) and (5.46) imply that (5.42) holds. □

Next step is to derive an upper bound for the error between the discrete solutions (y_h, p_h) and the auxiliary solutions $(y_h(u), p_h(u))$.

Lemma 5.3.2. *Let* (y_h, p_h) *and* $(y_h(u), p_h(u))$ *be the solutions of* (5.29) *and* (5.41)*, respectively. Then, there exist the following estimates for positive constants* C_1 *and* C_2 *independent of* h

$$\|y_h - y_h(u)\|_\xi \leq C_1\|u - u_h\|_{L^2(L^2(\mathcal{D});\Gamma)}, \tag{5.47a}$$

$$\|p_h - p_h(u)\|_\xi \leq C_2\|u - u_h\|_{L^2(L^2(\mathcal{D});\Gamma)}. \tag{5.47b}$$

Proof. By subtracting (5.41a) from (5.29a) and taking $v_h = y_h - y_h(u)$, it holds that

$$a_\xi[y_h - y_h(u), y_h - y_h(u)] = [u_h - u, y_h - y_h(u)]_\xi.$$

With the help of the coercivity of a_ξ (3.13) and the Cauchy-Schwarz inequality (2.11), one can obtain

$$\begin{aligned} c_{cv}\|y_h - y_h(u)\|_\xi^2 &\le a_\xi[y_h - y_h(u), y_h - y_h(u)] \\ &\le \|u_h - u\|_{L^2(L^2(\mathcal{D});\Gamma)}\|y_h - y_h(u)\|_\xi, \end{aligned}$$

which yields the desired result (5.47a).

Analogously, by subtracting (5.41b) from (5.29b) and taking $v_h = p_h - p_h(u)$, we have that

$$\begin{aligned} &a_\xi[p_h - p_h(u), p_h - p_h(u)] \\ &\quad = (1+\gamma)[y_h - y_h(u), p_h - p_h(u)]_\xi - \gamma\left[\widetilde{\mathbb{E}}[y_h - y_h(u)], p_h - p_h(u)\right]_\xi \\ &\quad = (1+\gamma)[y_h - y_h(u), p_h - p_h(u)]_\xi + \gamma\left[\widetilde{\mathbb{E}}[y_h(u) - y_h], p_h - p_h(u)\right]_\xi. \end{aligned}$$

It follows from the coercivity of a_ξ (3.13), Cauchy-Schwarz inequality (2.11), and Jensen's inequality (2.14) that

$$\begin{aligned} c_{cv}\|p_h - p_h(u)\|_\xi^2 &\le a_\xi[p_h - p_h(u), p_h - p_h(u)] \\ &\le (1+2\gamma)\|p_h - p_h(u)\|_{L^2(L^2(\mathcal{D});\Gamma)}\|y_h - y_h(u)\|_\xi. \end{aligned} \tag{5.48}$$

It is noted that the procedure applied in (5.45) is also used in the derivation of (5.48). Hence, by (5.48) and (5.47a), the desired result (5.47b) is deduced. □

To obtain an upper bound for the control, the domain $\mathcal{D}$ is divided into pieces by considering the active and inactive parts of the control u as done in [4, 112]:

$$\mathcal{D}^+ = \left\{\bigcup_K : K \subset \mathcal{D},\ u_a < u|_K < u_b\right\}, \tag{5.49a}$$

$$\mathcal{D}^\partial = \left\{\bigcup_K : K \subset \mathcal{D},\ u|_K = u_a \text{ or } u|_K = u_b\right\}, \tag{5.49b}$$

$$\mathcal{D}^- = \mathcal{D} \setminus (\mathcal{D}^+ \cup \mathcal{D}^\partial). \tag{5.49c}$$

It is assumed that these sets are disjoint, $\mathcal{D} = \mathcal{D}^+ \cup \mathcal{D}^\partial \cup \mathcal{D}^-$, and $\mathcal{D}^-$ satisfies the following inequality related to the regularity of u and $\mathcal{T}_h$

$$\text{meas}(\mathcal{D}^-) \le Ch, \tag{5.50}$$

which is valid if the boundary of the $\mathcal{D}^{\partial}$ is represented by finite rectifiable curves [123]. Further, a set is defined as $\mathcal{D}^{+} \subset \mathcal{D}^{*} = \{\boldsymbol{x} \in \mathcal{D} \ : \ u_a < u(\boldsymbol{x}) < u_b\}$ [164].

Lemma 5.3.3. *Suppose that* (y, u, p) *and* (y_h, u_h, p_h) *are the solutions of* (5.19) *and* (5.29)*, respectively. Let* $u \in L^2(W^{1,\infty}(\mathcal{D}); \Gamma)$ *with* $u|_{\mathcal{D}^+} \in L^2(H^2(\mathcal{D}^+); \Gamma)$ *be given. Then, it holds that*

$$\begin{aligned}
&\|u - u_h\|_{L^2(L^2(\mathcal{D});\Gamma)} \\
&\quad \leq C\|p - p_h(u)\|_{L^2(L^2(\mathcal{D};\Gamma))} + Ch^{3/2}\|u\|_{L^2(W^{1,\infty}(\mathcal{D});\Gamma)} \\
&\qquad + C\Big(h\|p\|_{L^2(H^1(\mathcal{D});\Gamma)} + \sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}p\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\Big). \qquad (5.51)
\end{aligned}$$

Proof. With the help of Lemma 5.3.1, the expression (5.40), the standard Lagrangian interpolation Πu, the assumption $\mathcal{D}^{+} \subset \mathcal{D}^{*}$, and the notation $p_h = p_h(u_h)$, it is obtained that

$$\begin{aligned}
\mu\|u - u_h\|^2_{L^2(L^2(\mathcal{D});\Gamma)} &\leq \mathcal{J}_h'(u)\cdot(u - u_h) - \mathcal{J}_h'(u_h)\cdot(u - u_h) \\
&= [\mu u + p_h(u), u - u_h]_{\xi} - [\mu u_h + p_h, u - u_h]_{\xi} \\
&= \underbrace{[\mu u + p, u - u_h]_{\xi}}_{-\mathcal{J}'(u)\cdot(u_h - u)\leq 0} - [p - p_h(u), u - u_h]_{\xi} \\
&\quad + \underbrace{[\mu u_h + p_h, u_h - \Pi u]_{\xi}}_{-\mathcal{J}_h'(u_h)\cdot(\Pi u - u_h)\leq 0} + [\mu u_h + p_h, \Pi u - u]_{\xi} \\
&\leq [\mu u_h + p_h, \Pi u - u]_{\xi} + [p_h(u) - p, u - u_h]_{\xi}. \qquad (5.52)
\end{aligned}$$

The first term in (5.52) can be rewritten as follows

$$\begin{aligned}
[\mu u_h + p_h, \Pi u - u]_{\xi} &= [\mu u_h + p_h - \mu u - p, \Pi u - u]_{\xi} + [\mu u + p, \Pi u - u]_{\xi} \\
&= [\mu u_h - \mu u, \Pi u - u]_{\xi} + [\mu u + p, \Pi u - u]_{\xi} \\
&\quad + [p_h - p_h(u), \Pi u - u]_{\xi} + [p_h(u) - p, \Pi u - u]_{\xi}. \qquad (5.53)
\end{aligned}$$

Then, inserting (5.53) into (5.52) and applying Cauchy-Schwarz (2.11) and Young's (2.12) inequalities and Lemma 5.3.2 produce

$$\begin{aligned}
\mu\|u - u_h\|^2_{L^2(L^2(\mathcal{D});\Gamma)} \ \leq \ & c_1\|p_h(u) - p\|^2_{L^2(L^2(\mathcal{D});\Gamma)} + c_2\|u - u_h\|^2_{L^2(L^2(\mathcal{D});\Gamma)} \\
& + c_3\|u - \Pi u\|^2_{L^2(L^2(\mathcal{D});\Gamma)} + [\mu u + p, \Pi u - u]_{\xi}. \qquad (5.54)
\end{aligned}$$

Since $\Pi u(\boldsymbol{x}) = u(\boldsymbol{x})$ for any vertex $\boldsymbol{x}$, $\Pi u \in \mathcal{U}_h^{ad}$ and the following estimates hold

$$\|u - \Pi u\|_{L^2(L^2(\mathcal{D}^+);\Gamma)} \leq Ch^2 \|u\|_{L^2(H^2(\mathcal{D}^+);\Gamma)}, \tag{5.55a}$$

$$\|u - \Pi u\|_{L^2(W^{0,\infty}(\mathcal{D}^-);\Gamma)} \leq Ch \|u\|_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)} \tag{5.55b}$$

for $u \in L^2(W^{1,\infty}(\mathcal{D});\Gamma)$ and $u|_{\mathcal{D}^*} \subset L^2(H^2(\mathcal{D}^*);\Gamma)$. Hence

$$\begin{aligned}
&\|u - \Pi u\|^2_{L^2(L^2(\mathcal{D});\Gamma)} \\
&\quad = \|u - \Pi u\|^2_{L^2(L^2(\mathcal{D}^+);\Gamma)} + \underbrace{\|u - \Pi u\|^2_{L^2(L^2(\mathcal{D}^\partial);\Gamma)}}_{=0} + \|u - \Pi u\|^2_{L^2(L^2(\mathcal{D}^-);\Gamma)} \\
&\quad \leq \|u - \Pi u\|^2_{L^2(L^2(\mathcal{D}^+);\Gamma)} + C\|u - \Pi u\|^2_{L^2(W^{0,\infty}(\mathcal{D}^-);\Gamma)} \text{meas}(\mathcal{D}^-) \\
&\quad \leq Ch^4 \|u\|^2_{L^2(H^2(\mathcal{D}^+);\Gamma)} + Ch^3 \|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)} \\
&\quad \leq Ch^3 \left(h\|u\|^2_{L^2(H^2(\mathcal{D}^+);\Gamma)} + \|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)} \right) \\
&\quad \leq Ch^3 \left(\|u\|^2_{L^2(H^2(\mathcal{D}^+);\Gamma)} + \|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)} \right).
\end{aligned} \tag{5.56}$$

By the variational inequality (5.29c) and the definitions of domains (5.49), it can be seen that

$$\mu u + p = 0 \text{ on } \mathcal{D}^+ \quad \text{and} \quad \Pi u - u = 0 \text{ on } \mathcal{D}^\partial.$$

Then,

$$\begin{aligned}
[\mu u + p, \Pi u - u]_\xi = &\underbrace{[\mu u - \Pi_h(\mu u) + \Pi_h(\mu u), \Pi u - u]_{\mathcal{D}^-}}_{T_1} \\
&+ \underbrace{[p - \mathcal{P}_{hq}(p) + \mathcal{P}_{hq}(p), \Pi u - u]_{\mathcal{D}^-}}_{T_2}.
\end{aligned} \tag{5.57}$$

It follows from the inequalities (5.33), (5.34b), and (5.55b), Sobolev embedding theorem, see, e.g., [2], and Young's inequality (2.12) that

$$\begin{aligned}
T_1 &= [\mu u - \Pi_h(\mu u), \Pi u - u]_{\mathcal{D}^-} + [\Pi_h(\mu u), \Pi u - u]_{\mathcal{D}^-} \\
&\leq \mu \Big(\|u - \Pi_h u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} + \|\Pi_h u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} \Big) \|u - \Pi u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} \\
&\leq \mu \Big(\|u - \Pi_h u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} + C\|u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} \Big) \|u - \Pi u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} \\
&\leq \mu \Big(\|u - \Pi_h u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} + C\|u\|_{L^2(W^{0,\infty}(\mathcal{D}^-);\Gamma)} \text{meas}(\mathcal{D}^-) \Big) \\
&\qquad \times \|u - \Pi u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)} \\
&\leq Ch \|u\|_{L^2(H^1(\mathcal{D}^-);\Gamma)} \|u - \Pi u\|_{L^2(W^{0,\infty}(\mathcal{D}^-);\Gamma)} \text{meas}(\mathcal{D}^-) \\
&\leq Ch^3 \|u\|_{L^2(H^1(\mathcal{D}^-);\Gamma)} \|u\|_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)} \\
&\leq Ch^3 \Big(\|u\|^2_{L^2(H^1(\mathcal{D}^-);\Gamma)} + \|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)} \Big).
\end{aligned} \tag{5.58}$$

Next, with the help of the projector operator in (5.37) and the bounds in (3.22), (5.33), (5.39), and (5.55b), Sobolev embedding theorem, and Cauchy and Young's inequalities (2.11) and (2.12), a bound for the second term T_2 in (5.57) is obtained

$$\begin{aligned}
T_2 &= [p-\Pi_h(p),\Pi u-u]_{\mathcal{D}^-}+[\mathcal{P}_{hq}(p),\Pi u-u]_{\mathcal{D}^-}+[\Pi_h(I-\Pi_q)(p),\Pi u-u]_{\mathcal{D}^-}\\
&\leq \Big(\|p-\Pi_h(p)\|_{L^2(L^2(\mathcal{D}^-);\Gamma)}+\|\mathcal{P}_{hq}(p)\|_{L^2(L^2(\mathcal{D}^-);\Gamma)}\Big)\|\Pi u-u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)}\\
&\quad+\|\Pi_h(I-\Pi_q)(p)\|_{L^2(L^2(\mathcal{D}^-);\Gamma)}\|\Pi u-u\|_{L^2(L^2(\mathcal{D}^-);\Gamma)}\\
&\leq C_1\Big(h\|p\|_{L^2(H^1(\mathcal{D}^-);\Gamma)}+\|p\|_{L^2(W^{0,\infty}(\mathcal{D}^-);\Gamma)}\mathrm{meas}(\mathcal{D}^-)\Big)h^2\|u\|_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)}\\
&\quad+C_2\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}p\|_{L^2(H^1(\mathcal{D}^-);\Gamma)}}{(q_n+1)!}h^2\|u\|_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)}\\
&\leq C_1\Big(\frac{h^2}{2}\|p\|^2_{L^2(H^1(\mathcal{D}^-);\Gamma)}+\frac{h^4}{2}\|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)}\Big)\\
&\quad+C_2\left(\frac{1}{2}\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{2q_n+2}\frac{\|\partial_{\xi_n}^{q_n+1}p\|^2_{L^2(H^1(\mathcal{D}^-);\Gamma)}}{((q_n+1)!)^2}+\frac{h^4}{2}\|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)}\right).
\end{aligned}\tag{5.59}$$

Combination of (5.58) and (5.59) yields

$$\begin{aligned}
&[\mu u+p,\Pi u-u]_\xi\\
&\quad\leq Ch^3\Big(\|u\|^2_{L^2(H^1(\mathcal{D}^-);\Gamma)}+\|u\|^2_{L^2(W^{1,\infty}(\mathcal{D}^-);\Gamma)}\Big)\\
&\qquad+Ch^2\|p\|^2_{L^2(H^1(\mathcal{D}^-);\Gamma)}+C\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{2q_n+2}\frac{\|\partial_{\xi_n}^{q_n+1}p\|^2_{L^2(H^1(\mathcal{D}^-);\Gamma)}}{((q_n+1)!)^2}.
\end{aligned}\tag{5.60}$$

Finally, inserting (5.56) and (5.60) into (5.54), the proof of Lemma 5.3.3 is completed. □

Lemma 5.3.4. *Suppose that* (y,p) *and* $(y_h(u),p_h(u))$ *are the solutions of* (5.19) *and* (5.41)*, respectively. Then, it holds*

$$\begin{aligned}
\|y-y_h(u)\|_\xi &\leq Ch\|y\|_{L^2(H^2(\mathcal{D});\Gamma)}\\
&\quad+\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!},
\end{aligned}\tag{5.61}$$

and

$$\begin{aligned}
&\|p - p_h(u)\|_\xi \\
&\quad \le Ch\Big(\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \|p\|_{L^2(H^2(\mathcal{D});\Gamma)}\Big) \\
&\qquad + \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\Big(\|\partial_{\xi_n}^{q_n+1} y\|_{L^2(H^1(\mathcal{D});\Gamma)} + \|\partial_{\xi_n}^{q_n+1} p\|_{L^2(H^1(\mathcal{D});\Gamma)}\Big)}{(q_n+1)!}.
\end{aligned} \tag{5.62}$$

Proof. An application of the coercivity and continuity of a_ξ in (3.13), $H^1(\mathcal{D})$–projection $\mathcal{R}_h$ in (5.35), $L^2(\mathcal{D})$–projection Π_q in (5.32a), and Galerkin orthogonality yields

$$\begin{aligned}
c_{cv}&\|y - y_h(u)\|_\xi^2 \\
&\le a_\xi[y - y_h(u), y - y_h(u)] \\
&\le a_\xi[y - y_h(u), y - \Pi_q\left(\mathcal{R}_h(y)\right)] + \underbrace{a_\xi[y - y_h(u), \Pi_q\left(\mathcal{R}_h(y)\right) - y_h(u)]}_{=0} \\
&\le c_{ct}\|y - y_h(u)\|_\xi \|y - \Pi_q\left(\mathcal{R}_h(y)\right)\|_\xi.
\end{aligned}$$

Then, by the approximation property (5.36), it is obvious to get

$$\begin{aligned}
\|y - y_h(u)\|_\xi &\le \frac{c_{ct}}{c_{cv}} \|y - \Pi_q\left(\mathcal{R}_h(y)\right)\|_\xi \\
&\le Ch\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \sum_{n=1}^{N} \left(\frac{k_n}{2}\right)^{q_n+1} \frac{\|\partial_{\xi_n}^{q_n+1} y\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!},
\end{aligned}$$

which is the desired result (5.61). Analogously, it is deduced that

$$\begin{aligned}
c_{cv}&\|p - p_h(u)\|_\xi^2 \\
&\le a_\xi[p - p_h(u), p - p_h(u)] \\
&\le a_\xi[p - \Pi_q\left(\mathcal{R}_h(y)\right), p - p_h(u)] + \underbrace{a_\xi[\Pi_q\left(\mathcal{R}_h(y)\right) - p_h(u), p - p_h(u)]}_{=0} \\
&= (1+\gamma)[y - y_h(u), p - \Pi_q\left(\mathcal{R}_h(y)\right)]_\xi + \gamma\big[\widetilde{\mathbb{E}}[y_h(u) - y], p - \Pi_q\left(\mathcal{R}_h(y)\right)\big]_\xi \\
&\le (1+2\gamma)\|y - y_h(u)\|_\xi \|p - \Pi_q\left(\mathcal{R}_h(y)\right)\|_{L^2(L^2(\mathcal{D});\Gamma)} \\
&\le \frac{(1+2\gamma)}{2}\|y - y_h(u)\|_\xi^2 + \frac{(1+2\gamma)}{2}\|p - \Pi_q\left(\mathcal{R}_h(y)\right)\|_{L^2(L^2(\mathcal{D});\Gamma)}^2,
\end{aligned}$$

where the definition of bilinear forms (3.9), the procedure applied in (5.45), and Young's inequality (2.12) are used. Then, using the approximation property (5.36) and (5.61), the proof of (5.62) is completed. □

Now, by combining the findings in Lemmas 5.3.3 and 5.3.4, the error analysis is finalized.

Theorem 5.3.5. *Assume that* (y, u, p) *and* (y_h, u_h, p_h)*, respectively, are the solutions of* (5.19) *and* (5.29)*. Then, it holds that*

$$\begin{aligned}
&\|u - u_h\|_{L^2(L^2(\mathcal{D});\Gamma)} + \|y - y_h\|_\xi + \|p - p_h\|_\xi \\
&\leq Ch^{3/2}\|u\|_{L^2(W^{1,\infty}(\mathcal{D});\Gamma)} + Ch\big(\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \|p\|_{L^2(H^2(\mathcal{D});\Gamma)}\big) \\
&\quad + C\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\Big(\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)} + \|\partial_{\xi_n}^{q_n+1}p\|_{L^2(H^1(\mathcal{D});\Gamma)}\Big)}{(q_n+1)!}.
\end{aligned} \tag{5.63}$$

Proof. From (5.51) and (5.62), it is obtained that

$$\begin{aligned}
&\|u - u_h\|_{L^2(L^2(\mathcal{D});\Gamma)} \\
&\quad \leq Ch^{3/2}\|u\|_{L^2(W^{1,\infty}(\mathcal{D});\Gamma)} + Ch\big(\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \|p\|_{L^2(H^2(\mathcal{D});\Gamma)}\big) \\
&\qquad + C\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\frac{\Big(\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)} + \|\partial_{\xi_n}^{q_n+1}p\|_{L^2(H^1(\mathcal{D});\Gamma)}\Big)}{(q_n+1)!}.
\end{aligned} \tag{5.64}$$

Moreover, by Lemmas 5.3.2 and 5.3.4, and the bound (5.64), it follows

$$\begin{aligned}
&\|y - y_h\|_\xi + \|p - p_h\|_\xi \\
&\leq \|y - y_h(u)\|_\xi + \|y_h(u) - y_h\|_\xi + \|p - p_h(u)\|_\xi + \|p_h(u) - p_h\|_\xi \\
&\leq C\|u - u_h\|_{L^2(L^2(\mathcal{D});\Gamma)} + Ch\big(\|y\|_{L^2(H^2(\mathcal{D});\Gamma)} + \|p\|_{L^2(H^2(\mathcal{D});\Gamma)}\big) \\
&\quad + C\sum_{n=1}^{N}\left(\frac{k_n}{2}\right)^{q_n+1}\left(\frac{\|\partial_{\xi_n}^{q_n+1}y\|_{L^2(H^1(\mathcal{D});\Gamma)} + \|\partial_{\xi_n}^{q_n+1}q\|_{L^2(H^1(\mathcal{D});\Gamma)}}{(q_n+1)!}\right).
\end{aligned} \tag{5.65}$$

Thus, by combining (5.64) and (5.65), the desired result (5.63) is achieved. □

5.4 Matrix Formulation

This section first constructs the matrix formulation of the underlying problem (5.16)–(5.17) by employing the "optimize-then-discretize"approach; see, e.g., [142]. In this methodology, one obtains the optimality system (5.19) of the infinite-dimensional optimization problem, and then discretizes the optimality system by a stochastic discontinuous Galerkin method discussed in Section 5.2. Later, we propose a low–rank

variant of the generalized minimal residual (GMRES) method with a suitable preconditioner to solve the corresponding linear system based on the numerical findings in Chapter 3.

5.4.1 State System

After an application of the discretization techniques discussed in Section 5.2, one gets the following linear system for the state part of the optimality system (5.19):

$$\underbrace{\left(\sum_{i=0}^{N} \mathcal{G}_i \otimes \mathcal{K}_i\right)}_{\mathcal{A}} \mathbf{y} - \underbrace{(\mathcal{G}_0 \otimes M)}_{\mathcal{M}} \mathbf{u} = \underbrace{\left(\sum_{i=0}^{N} \mathbf{g}_i \otimes \mathbf{f}_i\right)}_{\mathcal{F}}, \tag{5.66}$$

where $\mathbf{y} = (y_0, \ldots, y_{J-1})^T$ and $\mathbf{u} = (u_0, \ldots, u_{J-1})^T$ with $y_i, u_i \in \mathbb{R}^{N_d}, i = 0, 1, \ldots, J-1$ and N_d corresponds to the degree of freedoms for the spatial discretization. The mass matrix $M \in \mathbb{R}^{N_d \times N_d}$ is given by

$$M(r,s) = \sum_{K \in \mathcal{T}_h} \int_K \varphi_r \varphi_s \, d\boldsymbol{x}.$$

Here, $\mathcal{K}_i \in \mathbb{R}^{N_d \times N_d}$ and $\mathcal{G}_i \in \mathbb{R}^{J \times J}$ represent the stiffness and stochastic matrices, respectively, whereas $\mathbf{f}_i \in \mathbb{R}^{N_d}$ and $\mathbf{g}_i \in \mathbb{R}^J$ are the right–hand side and stochastic vectors, respectively; see Section 3.1.2 for the constructions of the matrices and vectors in details.

5.4.2 Matrix Formulation of the Optimality System

The discrete optimality system in (5.29) can be represented as a block matrix system including the state, adjoint, and variational equations in the finite dimensional setting. To solve the underlying block linear system, "the primal–dual active set (PDAS) methodology as a semi-smooth Newton step" is applied; see, e.g., [24] for more details. After a definition of the active sets

$$\begin{aligned} A^- &= \bigcup \{\boldsymbol{x} \in K \ : \ -\mathbf{p} - \mu \mathbf{u}_a < 0, \ \forall K \in \mathcal{T}_h\}, \\ A^+ &= \bigcup \{\boldsymbol{x} \in K \ : \ -\mathbf{p} - \mu \mathbf{u}_b > 0, \ \forall K \in \mathcal{T}_h\}, \end{aligned}$$

and the inactive set

$$\mathcal{I} \;=\; \mathcal{T}_h \setminus \left(A^- \cup A^+\right),$$

the block formulation becomes

$$\mathcal{A}\mathbf{y} - \mathcal{M}_I\mathbf{u} = \mathcal{F}, \tag{5.67a}$$

$$\mathcal{A}^*\mathbf{p} - \mathcal{M}_\gamma\mathbf{y} = -\mathcal{F}^d, \tag{5.67b}$$

$$\left(\mathcal{G}_0 \otimes \mathrm{diag}(\mathbb{1}_\mathcal{I})\right)\mathbf{p} + \mu\left(\mathcal{G}_0 \otimes I\right)\mathbf{u} = \left(\mathbf{g}_0 \otimes \mathbb{1}_{A^-}\right)\mu\,\mathbf{u}_a + \left(\mathbf{g}_0 \otimes \mathbb{1}_{A^+}\right)\mu\,\mathbf{u}_b, \tag{5.67c}$$

where

$$\mathcal{M}_I := I \otimes M,$$

$$\mathcal{F}^d := \mathbf{g}_0 \otimes \mathbf{y}^d \text{ with } \mathbf{y}^d(s) = \sum_{K\in\mathcal{T}_h}\int_K y^d\varphi_s\, d\boldsymbol{x},$$

$$\mathcal{M}_\gamma := \left(\mathcal{G}_0 \otimes M\right) + \gamma\left(\mathcal{M}_0 \otimes M\right) \text{ with } \mathcal{M}_0 = \mathrm{diag}\left(0, \langle\Psi_1\rangle^2, \ldots, \langle\Psi_{J-1}\rangle^2\right),$$

and $\mathbb{1}_{A^-}$, $\mathbb{1}_{A^+}$, and $\mathbb{1}_\mathcal{I}$ correspond to the characteristic functions of A^-, A^+, and $\mathcal{I}$, respectively. Equivalently, $\mathcal{M}_\gamma$ can be rewritten as

$$\mathcal{M}_\gamma := \mathcal{G}_\gamma \otimes M, \quad \text{with} \quad \mathcal{G}_\gamma := \mathcal{G}_0 + \gamma\mathcal{M}_0,$$

where

$$\mathcal{G}_\gamma(r,s) = \begin{cases} \langle\Psi_0\rangle^2, & \text{if } r = s = 0, \\ (1+\gamma)\langle\Psi_r\rangle^2, & \text{if } r = s = 1,\ldots,J-1, \\ 0, & \text{otherwise.} \end{cases} \tag{5.68}$$

Rearranging (5.67) gives us the following linear matrix system

$$\begin{bmatrix} \mathcal{M}_\gamma & 0 & -\mathcal{A}^* \\ 0 & \mu\left(\mathcal{G}_0 \otimes I\right) & \mathcal{G}_0 \otimes \mathrm{diag}(\mathbb{1}_\mathcal{I}) \\ -\mathcal{A} & \mathcal{M}_I & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathcal{F}^d \\ \mu\big(\left(\mathbf{g}_0 \otimes \mathbb{1}_{A^-}\right)\mathbf{u}_a + \left(\mathbf{g}_0 \otimes \mathbb{1}_{A^+}\right)\mathbf{u}_b\big) \\ -\mathcal{F} \end{bmatrix}, \tag{5.69}$$

which is a saddle point system. It is noted that since Legendre polynomials are used, $\mathcal{G}_0 = I$, and hence, $\mathcal{M}_I = \mathcal{M}$.

The saddle point system (5.69) is generally very large in practical implementations, depending on the length of the random vector ξ and the number of refinements in the

spatial discretization. To overcome the curse of dimensionality, it is useful to employ a low–rank approximation that decreases both computational complexity and memory needs by using a Kronecker–product structure of the matrices provided in (5.69).

5.4.3 Low–Rank Approach

In this section, low–rank approximation techniques are explained, and how they can be used with iterative solvers in order to solve the saddle point system (5.69). Iterative techniques, such as Krylov subspace approaches, are particularly appealing when the optimality system (5.70) is large and sparse since their storage needs often depend primarily on the number of nonzero entries in the coefficient matrix. The concept of low–rank approximation for the state equation, namely convection diffusion equation with random inputs (3.1), is discussed in Section 3.3. As we have already noted in Chapter 3, an optimal Krylov subspace solver for the convection-dominated problems is the GMRES algorithm [133]. Thus, the low-rank version of the GMRES method, which uses the Kronecker product structure of the obtained linear system to reduce computational complexity and memory requirements, is used to address large matrix systems emerging from stochastic Galerkin methods in this chapter.

Following the properties in Section 2.1.1 and the construction in Section 3.3, the system (5.69) can be interpreted as

$$\underbrace{\begin{bmatrix} \mathcal{G}_\gamma \otimes M & 0 & -\sum_{i=0}^{N} \mathcal{G}_i \otimes \mathcal{K}_i^* \\ 0 & \mu\left(\mathcal{G}_0 \otimes I\right) & \mathcal{G}_0 \otimes \text{diag}(\mathbb{1}_{\mathcal{I}}) \\ -\sum_{i=0}^{N} \mathcal{G}_i \otimes \mathcal{K}_i & \mathcal{G}_0 \otimes M & 0 \end{bmatrix}}_{\mathcal{L}} \underbrace{\begin{bmatrix} \texttt{vec}(Y) \\ \texttt{vec}(U) \\ \texttt{vec}(P) \end{bmatrix}}_{\Theta} = \underbrace{\begin{bmatrix} \texttt{vec}(B_1) \\ \texttt{vec}(B_2) \\ \texttt{vec}(B_3) \end{bmatrix}}_{\mathcal{B}}, \quad (5.70)$$

where

$$Y = (y_0, \ldots, y_{J-1}),\ U = (u_0, \ldots, u_{J-1}),\ P = (p_0, \ldots, p_{J-1}),\ B_1 = \texttt{mat}\left(\mathcal{F}^d\right),$$
$$B_2 = \texttt{mat}\left(\mu\big(\left(\mathbf{g}_0 \otimes \mathbb{1}_{A^-}\right)\mathbf{u}_a + \left(\mathbf{g}_0 \otimes \mathbb{1}_{A^+}\right)\mathbf{u}_b\big)\right),\ B_3 = \texttt{mat}\left(-\mathcal{F}\right).$$

By the identity (2.1.1), we have

$$\mathcal{L}\Theta = \texttt{vec}\left(\begin{bmatrix} MY\mathcal{G}_\gamma^T - \sum_{i=0}^{N}\mathcal{K}_i^* P\mathcal{G}_i^T \\ \mu IU\mathcal{G}_0^T + \text{diag}(\mathbb{1}_\mathcal{I})P\mathcal{G}_0^T \\ -\sum_{i=0}^{N}\mathcal{K}_i Y\mathcal{G}_i^T + MU\mathcal{G}_0^T \end{bmatrix}\right) = \texttt{vec}\left(\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}\right). \tag{5.71}$$

Assuming that the matrices Θ and $\mathcal{B}$ have the following low–rank representations, see, e.g., [139, 16, 69],

$$\begin{aligned} Y &= W_Y V_Y^T, && W_Y \in \mathbb{R}^{N_d \times r_Y},\ V_Y \in \mathbb{R}^{J \times r_Y}, \\ U &= W_U V_U^T, && W_U \in \mathbb{R}^{N_d \times r_U},\ V_U \in \mathbb{R}^{J \times r_U}, \\ P &= W_P V_P^T, && W_P \in \mathbb{R}^{N_d \times r_P},\ V_P \in \mathbb{R}^{J \times r_P}, \\ B_1 &= B_{11} B_{12}^T, && B_{11} \in \mathbb{R}^{N_d \times r_{B_1}},\ B_{12} \in \mathbb{R}^{J \times r_{B_1}}, \\ B_2 &= B_{21} B_{22}^T, && B_{21} \in \mathbb{R}^{N_d \times r_{B_2}},\ B_{22} \in \mathbb{R}^{J \times r_{B_2}}, \\ B_3 &= B_{31} B_{32}^T, && B_{31} \in \mathbb{R}^{N_d \times r_{B_3}},\ B_{32} \in \mathbb{R}^{J \times r_{B_3}}, \end{aligned} \tag{5.72}$$

with $r_Y, r_U, r_P, r_{B_1}, r_{B_2}, r_{B_3} \ll N_d, J$, (5.71) can be stated as follows:

$$\begin{bmatrix} MW_Y V_Y^T \mathcal{G}_\gamma^T - \sum_{i=0}^{N}\mathcal{K}_i^* W_P V_P^T \mathcal{G}_i^T \\ \mu IW_U V_U^T \mathcal{G}_0^T + \text{diag}(\mathbb{1}_\mathcal{I}) W_P V_P^T \mathcal{G}_0^T \\ -\sum_{i=0}^{N}\mathcal{K}_i W_Y V_Y^T \mathcal{G}_i^T + MW_U V_U^T \mathcal{G}_0^T \end{bmatrix} = \begin{bmatrix} B_{11}B_{12}^T \\ B_{21}B_{22}^T \\ B_{31}B_{32}^T \end{bmatrix}, \tag{5.73}$$

where vec operator is ignored for the simplicity and readability. Moreover, the three block rows in (5.73) can be written as

$$\underbrace{\begin{bmatrix} MW_Y & -\sum_{i=0}^{N}\mathcal{K}_i^* W_P \end{bmatrix}}_{\widehat{W}_1} \underbrace{\begin{bmatrix} \mathcal{G}_\gamma V_Y & \mathcal{G}_i V_P \end{bmatrix}^T}_{\widehat{V}_1^T}, \tag{5.74a}$$

$$\underbrace{\begin{bmatrix} \mu IW_U & \text{diag}(\mathbb{1}_\mathcal{I}) W_P \end{bmatrix}}_{\widehat{W}_2} \underbrace{\begin{bmatrix} \mathcal{G}_0 V_U & \mathcal{G}_0 V_P \end{bmatrix}^T}_{\widehat{V}_2^T}, \tag{5.74b}$$

$$\underbrace{\begin{bmatrix} -\sum_{i=0}^{N}\mathcal{K}_i W_Y & MW_U \end{bmatrix}}_{\widehat{W}_3} \underbrace{\begin{bmatrix} \mathcal{G}_i V_Y & \mathcal{G}_0 V_U \end{bmatrix}^T}_{\widehat{V}_3^T}, \tag{5.74c}$$

in low–rank formats $\widehat{W}_i \widehat{V}_i^T$ for $i = 1, 2, 3$. By the usage of (5.74), the low–rank approximate solutions to (5.70) can be obtained; see Algorithm 7 modified from [69] for

details of the low–rank implementation of the GMRES. Moreover, the inner product computation in Algorithm 7 denoted by

$$\texttt{trprod}(A_{11}, A_{12}, A_{21}, A_{22}, A_{31}, A_{32}, B_{11}, B_{12}, B_{21}, B_{22}, B_{31}, B_{32})$$

can be computed as following:

$$\begin{aligned}
\langle A, B\rangle_F &= \text{trace}\left(\left(A_{11}A_{12}^T\right)^T\left(B_{11}B_{12}^T\right)^T\right) + \text{trace}\left(\left(A_{21}A_{22}^T\right)^T\left(B_{21}B_{22}^T\right)^T\right) \\
&\quad + \text{trace}\left(\left(A_{31}A_{32}^T\right)^T\left(B_{31}B_{32}^T\right)^T\right) \\
&= \text{trace}\left(A_{11}^T B_{11} A_{12}^T B_{12}\right) + \text{trace}\left(A_{21}^T B_{21} A_{22}^T B_{22}\right) \\
&\quad + \text{trace}\left(A_{31}^T B_{31} A_{32}^T B_{32}\right),
\end{aligned}$$

where

$$A = \texttt{vec}\left(\begin{bmatrix} A_{11}A_{12}^T \\ A_{21}A_{22}^T \\ A_{31}A_{32}^T \end{bmatrix}\right), \qquad B = \texttt{vec}\left(\begin{bmatrix} B_{11}B_{12}^T \\ B_{21}B_{22}^T \\ B_{31}B_{32}^T \end{bmatrix}\right).$$

Low-rank factors can raise their rank throughout the iteration process using either matrix vector products or vector (matrix) additions. Thus, by utilizing truncation based on singular values [102] or truncation based on coarse–grid rank reduction [109], the expense of rank-reduction approaches is kept under control. Our strategy, which is explicitly discussed in Section 3.3 and inspired by the discussion in [139, 20], in which a truncated SVD of $U = W^T V \approx B\,\text{diag}(\sigma_1, \ldots, \sigma_r)C^T$ is built for the largest singular values that are bigger than the specified truncation tolerance. This operation is carried out by the truncation operator in Algorithm 7. Further, to represent accurately the full-rank solution in the numerical simulations, a rather small truncation tolerance is utilized.

When employed with a proper preconditioner, iterative techniques, such as GMRES, are known to show better convergence in terms of the number of iterations. The low–rank variants also display the same behaviour so that we use a block diagonal mean-based preconditioner of the form

$$\mathcal{P}_0 = \begin{bmatrix} \mathcal{M}_\gamma & 0 & 0 \\ 0 & \mu\left(\mathcal{G}_0 \otimes I\right) & 0 \\ 0 & 0 & \widetilde{S} \end{bmatrix},$$

Algorithm 7 Low–rank preconditioned GMRES (LRPGMRES)

Input: Coefficient matrix $\mathcal{L} : \mathbb{R}^{3N_d \times J} \to \mathbb{R}^{3N_d \times J}$, inverse of the preconditioner matrix $\mathcal{P}_0^{-1} : \mathbb{R}^{3N_d \times J} \to \mathbb{R}^{3N_d \times J}$, and right–hand side matrix $\mathcal{B}$ in the low–rank formats, truncation operator $\mathcal{T}$ with given tolerance ϵ_{trunc}.

Output: Matrix $\Theta \in \mathbb{R}^{3N_d \times J}$ satisfying $\|\mathcal{L}(\Theta) - \mathcal{B}\|_F / \|\mathcal{B}\|_F \leq \epsilon_{tol}$.

Choose initial guess $\Theta_{11}^{(0)}, \Theta_{12}^{(0)}, \Theta_{21}^{(0)}, \Theta_{22}^{(0)}, \Theta_{31}^{(0)}, \Theta_{33}^{(0)}$.

$(\widetilde{\Theta}_{11}, \widetilde{\Theta}_{12}, \widetilde{\Theta}_{21}, \widetilde{\Theta}_{22}, \widetilde{\Theta}_{31}, \widetilde{\Theta}_{32}) = \mathcal{L}(\Theta_{11}^{(0)}, \Theta_{12}^{(0)}, \Theta_{21}^{(0)}, \Theta_{22}^{(0)}, \Theta_{31}^{(0)}, \Theta_{32}^{(0)})$. $\quad \widetilde{\Theta}_{ij} \leftarrow \mathcal{T}(\widetilde{\Theta}_{ij})$

$R_{11}^{(0)} = \{B_{11}, \ -\Theta_{11}^{(0)}\}, \quad R_{12}^{(0)} = \{B_{12}, \ \Theta_{12}^{(0)}\}.$

$R_{21}^{(0)} = \{B_{21}, \ -\Theta_{21}^{(0)}\}, \quad R_{22}^{(0)} = \{B_{22}, \ \Theta_{22}^{(0)}\}.$ $\quad R_{ij}^{(0)} \leftarrow \mathcal{T}(R_{ij}^{(0)})$

$R_{31}^{(0)} = \{B_{31}, \ -\Theta_{31}^{(0)}\}, \quad R_{32}^{(0)} = \{B_{32}, \ \Theta_{32}^{(0)}\}.$

$\|R^0\| = \sqrt{\text{trprod}(R_{11}^{(0)}, \ldots, R_{11}^{(0)}, \ldots)}.$

$V_{11}^{(0)} = R_{11}^{(0)} / \|R^0\|_F, \quad V_{12}^{(0)} = R_{12}^{(0)}.$

$V_{21}^{(0)} = R_{21}^{(0)} / \|R^0\|_F, \quad V_{22}^{(0)} = R_{22}^{(0)}.$ $\quad V_{ij}^{(0)} \leftarrow \mathcal{T}(V_{ij}^{(0)})$

$V_{31}^{(0)} = R_{31}^{(0)} / \|R^0\|_F, \quad V_{32}^{(0)} = R_{32}^{(0)}.$

$\gamma = [\gamma_1, 0, \ldots, 0], \qquad \gamma_1 = \sqrt{\text{trprod}(V_{11}^{(0)}, \ldots, V_{11}^{(0)}, \ldots)}.$

while $i \leq$ maxit **do**

$(Z_{11}^{(i)}, Z_{12}^{(i)}, Z_{21}^{(i)}, Z_{22}^{(i)}, Z_{31}^{(i)}, Z_{32}^{(i)}) = \mathcal{P}_0^{-1}(V_{11}^{(i)}, V_{12}^{(i)}, V_{21}^{(i)}, V_{22}^{(i)}, V_{31}^{(i)}, V_{32}^{(i)})$, $\quad Z_{ij}^{(i)} \leftarrow \mathcal{T}(Z_{ij}^{(i)})$

$(W_{11}, W_{12}, W_{21}, W_{22}, W_{31}, W_{32}) = \mathcal{L}(Z_{11}^{(i)}, Z_{12}^{(i)}, Z_{21}^{(i)}, Z_{22}^{(i)}, Z_{31}^{(i)}, Z_{32}^{(i)})$. $\quad W_{ij} \leftarrow \mathcal{T}(W_{ij})$

for $j = 1, \ldots, i$ **do**

$m_{j,i} = \sqrt{\text{trprod}(W_{11}, \ldots, V_{11}^{(j)}, \ldots)}$

$W_{11} = \{W_{11}, \ -m_{j,i} V_{11}^{(i)}\}, \quad W_{12} = \{W_{12}, \ V_{12}^{(j)}\}.$

$W_{21} = \{W_{21}, \ -m_{j,i} V_{21}^{(i)}\}, \quad W_{22} = \{W_{22}, \ V_{22}^{(j)}\}.$ $\quad W_{ij} \leftarrow \mathcal{T}(W_{ij})$

$W_{31} = \{W_{31}, \ -m_{j,i} V_{31}^{(i)}\}, \quad W_{32} = \{W_{32}, \ V_{32}^{(j)}\}.$

end for

$m_{i+1,i} = \sqrt{\text{trprod}(W_{11}, \ldots, W_{11}, \ldots)}$

$V_{11}^{(i+1)} = W_{11} / m_{i+1,k}, \quad V_{12}^{(k+1)} = W_{12}.$

$V_{21}^{(i+1)} = W_{21} / m_{i+1,k}, \quad V_{22}^{(k+1)} = W_{22}.$ $\quad V_{ij}^{(i+1)} \leftarrow \mathcal{T}(V_{ij}^{(i+1)})$

$V_{31}^{(i+1)} = W_{31} / m_{i+1,k}, \quad V_{32}^{(k+1)} = W_{32}.$

Perform Givens rotations for the ith column of m:

for $j = 1, \ldots, i-1$ **do**

$$\begin{bmatrix} m_{j,i} \\ m_{j+1,i} \end{bmatrix} = \begin{bmatrix} c_j & s_j \\ -s_j & c_j \end{bmatrix} \begin{bmatrix} m_{j,i} \\ m_{j+1,i} \end{bmatrix}$$

end for

Compute ith Givens rotation, and perform for γ and last column of m.

$$\begin{bmatrix} \gamma_i \\ \gamma_{i+1} \end{bmatrix} = \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} \gamma_i \\ 0 \end{bmatrix}$$

$m_{i,i} = c_i m_{i,i} + s_i m_{i+1,i}, \qquad m_{i+1,i} = 0.$

if $|\gamma_{i+1}| \leq \epsilon_{tol}$ **then**

Compute y from $My = \xi$, where $(M)_{j,i} = m_{j,i}$.

$Y_{11} = \{y_1 V_{11}^{(1)}, \ldots, y_k V_{11}^{(i)}\}, \quad Y_{12} = \{V_{12}^{(1)}, \ldots, V_{12}^{(i)}\}.$

$Y_{21} = \{y_1 V_{21}^{(1)}, \ldots, y_k V_{21}^{(i)}\}, \quad Y_{22} = \{V_{22}^{(1)}, \ldots, V_{22}^{(i)}\}.$ $\quad Y_{ij} \leftarrow \mathcal{T}(Y_{ij})$

$Y_{31} = \{y_1 V_{31}^{(1)}, \ldots, y_k V_{31}^{(i)}\}, \quad Y_{32} = \{V_{32}^{(1)}, \ldots, V_{32}^{(i)}\}.$

$(\widetilde{Y}_{11}, \widetilde{Y}_{12}, \widetilde{Y}_{21}, \widetilde{Y}_{22}, \widetilde{Y}_{31}, \widetilde{Y}_{32}) = \mathcal{P}_0^{-1}(Y_{11}, Y_{12}, Y_{21}, Y_{22}, Y_{31}, Y_{32})$. $\quad \widetilde{Y}_{ij} \leftarrow \mathcal{T}(\widetilde{Y}_{ij})$

$\Theta_{11} = \{\Theta_{11}^{(0)}, \ \widetilde{Y}_{11}\}, \quad \Theta_{12} = \{\Theta_{12}^{(0)}, \ \widetilde{Y}_{12}\}.$

$\Theta_{21} = \{\Theta_{21}^{(0)}, \ \widetilde{Y}_{21}\}, \quad \Theta_{22} = \{\Theta_{22}^{(0)}, \ \widetilde{Y}_{22}\}.$ $\quad \Theta_{ij} \leftarrow \mathcal{T}(\Theta_{ij})$

$\Theta_{31} = \{\Theta_{31}^{(0)}, \ \widetilde{Y}_{31}\}, \quad \Theta_{32} = \{\Theta_{32}^{(0)}, \ \widetilde{Y}_{32}\}.$

end if

end while

where $\widetilde{S} = \big(\mathcal{G}_0 \otimes \widetilde{\mathcal{K}}_0\big)\mathcal{M}_\gamma^{-1}\big(\mathcal{G}_0 \otimes \widetilde{\mathcal{K}}_0\big)^T$ corresponds to the approximated Schur complement with $\widetilde{\mathcal{K}}_0 = \mathcal{K}_0 + \sqrt{\frac{1+\gamma}{\mu}} M \operatorname{diag}(\mathbb{1}_{\mathcal{I}})$; see, e.g, [21, 128].

5.5 Numerical Results

The numerical experiments in this section demonstrate the performance of the proposed discretization approaches and a low–rank variation of the GMRES methodology. On a 32 GB RAM Ubuntu Linux computer, all numerical calculations are carried out in MATLAB R2021a. When the residual shrinks below the specified tolerance threshold ($\epsilon_{tol} = 5 \times 10^{-3}$) or when the maximum number of iterations ($\#iter_{max} = 250$) is reached, iterative techniques are terminated. In order to avoid iterating the noise during the low-rank process, the truncation tolerance $\epsilon_{trunc} = 10^{-8}$ is selected such that $\epsilon_{trunc} \leq \epsilon_{tol}$.

In the numerical experiments, the random coefficient z is described by the covariance function in (3.59) with the correlation length ℓ_n. Linear elements are used to generate the discontinuous Galerkin basis, whereas the Legendre polynomials are taken as the stochastic basis functions since the underlying random variables have uniform distribution over $[-\sqrt{3}, \sqrt{3}]$, that is, $\xi_j \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}],\ j = 1, \ldots, N$. Further, all parameters used in the simulations are described in Table 5.1.

Table 5.1: Descriptions of the parameters used in the simulations.

Parameter	Description
N_d	degree of freedoms for the spatial discretization
N	truncation number in KL expansion
Q	highest order of basis polynomials for the stochastic domain
μ	regularization parameter of the control u
γ	risk-aversion parameter
ν	viscosity parameter
ℓ	correlation length
κ_z	standard deviation

5.5.1 Unconstrained Problem with Random Diffusion Parameter

For the first example, we present an unconstrained optimal control problem, that is, $\mathcal{U}^{ad} = \mathcal{U}$, having a random diffusion coefficient defined on $\mathcal{D} = [-1, 1]^2$ with the source function $f(\boldsymbol{x}) = 0$, the convection parameter $\mathbf{b}(\boldsymbol{x}) = (0, 1)^T$, and the Dirichlet boundary condition

$$y_{DB}(\boldsymbol{x}) = \begin{cases} y_{DB}(x_1, -1) = x_1, & y_{DB}(x_1, 1) = 0, \\ y_{DB}(-1, x_2) = -1, & y_{DB}(1, x_2) = 1. \end{cases}$$

The random diffusion parameter is chosen as $a(\boldsymbol{x}, \omega) = \nu\, z(\boldsymbol{x}, \omega)$, where the random field $z(\boldsymbol{x}, \omega)$ has the unity mean with the corresponding covariance function (3.59) and ν is the viscosity parameter. The desired state (or target) y^d corresponds to the stochastic solution of the forward model by taking $u(\boldsymbol{x}) = 0$. It is noted that the desired state exhibits an exponential boundary layer near $x_2 = 1$, where the solution changes in a dramatic manner; see Figure 3.1.

Table 5.2: Example 5.5.1: Computational values of the cost functional $\mathcal{J}(u_h)$ and tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$ obtained by $\mathcal{L}\backslash\mathcal{B}$ with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.5$, and $\gamma = 1$ for varying values of the viscosity parameter ν and the regularization parameter μ.

		$\mu = 1$	$\mu = 10^{-2}$	$\mu = 10^{-4}$	$\mu = 10^{-6}$
$\nu = 1$	$\mathcal{J}(u_h)$	1.2393e-05	2.5034e-06	1.0257e-06	9.7533e-07
	$\|y_h - y^d\|^2_{\mathcal{X}}$	5.4679e-06	7.1725e-07	3.2113e-08	1.6931e-09
$\nu = 10^{-2}$	$\mathcal{J}(u_h)$	1.4349e-05	8.3285e-06	7.2067e-07	6.0683e-07
	$\|y_h - y^d\|^2_{\mathcal{X}}$	1.3120e-05	4.1452e-06	9.0731e-08	3.5983e-09
$\nu = 10^{-4}$	$\mathcal{J}(u_h)$	1.3675e-05	1.1798e-06	3.9380e-07	3.7211e-07
	$\|y_h - y^d\|^2_{\mathcal{X}}$	1.5285e-05	4.3924e-07	1.1422e-07	8.3896e-09

The tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$ and cost functional $\mathcal{J}(u_h)$ obtained by $\mathcal{L}\backslash\mathcal{B}$ are shown in Table 5.2 for various values of the viscosity parameter ν and the regularization parameter μ. As μ declines, it is noticed that both the tracking term and the objective functional get smaller. Moreover, Table 5.3 shows that the peak values of states' variance can be reduced by increasing the value of the parameter γ.

In Table 5.4, we next show the performance of $\mathcal{L}\backslash\mathcal{B}$ in terms of total CPU times (in seconds) and storage requirements (in KB). Nevertheless, due to the simulation termi-

Table 5.3: Example 5.5.1: Peak values of the states' variance obtained by $\mathcal{L}\backslash\mathcal{B}$ with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\nu = 1$, and $\mu = 1$ for varying values of the risk–aversion γ and the standard deviation κ_z.

	$\kappa_z = 0.05$	$\kappa_z = 0.25$	$\kappa_z = 0.5$
$\gamma = 0$	4.5406e-05	1.1980e-03	5.7995e-03
$\gamma = 1$	4.1995e-05	1.0984e-03	5.1327e-03
$\gamma = 2$	3.8944e-05	1.0110e-03	4.5807e-03
$\gamma = 3$	3.6207e-05	9.3377e-04	4.1243e-03
$\gamma = 4$	3.3731e-05	8.6520e-04	3.7409e-03

Table 5.4: Example 5.5.1: Total CPU times (in seconds) and memory (in KB) for $N_d = 6144$, $Q = 3$, $\ell = 1$, $\mu = 10^{-2}$, $\gamma = 1$, and $\kappa_z = 0.5$.

$\mathcal{L}\backslash\mathcal{B}$	$\nu = 10^0$	$\nu = 10^{-2}$	$\nu = 10^{-4}$
N	CPU (Memory)	CPU (Memory)	CPU (Memory)
2	116.0 (2880)	116.1 (2880)	117.5 (960)
3	779.6 (5760)	787.7 (5760)	813.0 (5760)
4	OoM	OoM	OoM

nating with "out of memory", which we have designated as "OoM", some numerical results could not been reported. Therefore, we require efficient numerical techniques or solvers, such as the low–rank variation of GMRES iteration with a mean based preconditioner $\mathcal{P}_0$, to address the curse of dimensionality and hence to raise the value of truncation number N.

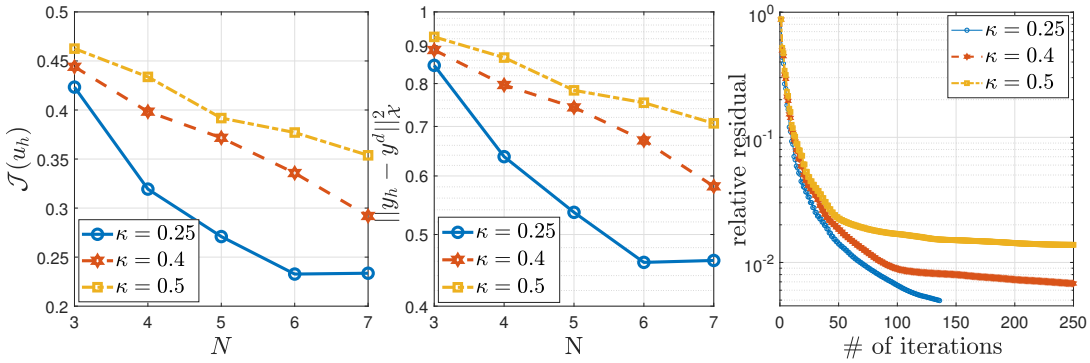

Figure 5.1: Example 5.5.1: Behaviours of the cost functional $\mathcal{J}(u_h)$ (left), the tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$ (middle), and the relative residual (right) with $N_d = 6144$, $Q = 3$, $\ell = 1$, $\nu = 1$, $\mu = 10^{-2}$, $\gamma = 0$, and the mean-based preconditioner $\mathcal{P}_0$ for varying values of κ_z.

Table 5.5 presents the simulation results by taking into account several data sets in the low–rank format. We provide findings for altering the truncation number N in the KL

Table 5.5: Example 5.5.1: Total number of iterations, total rank of the truncated solutions, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.5$, $\nu = 1$, $\gamma = 0$, and the mean-based preconditioner $\mathcal{P}_0$ for varying values of N and μ.

		$\mu = 1$	$\mu = 10^{-2}$	$\mu = 10^{-4}$
$N = 4$	#iter	250	250	250
	Rank	51	51	51
	CPU	40126.1	40017.9	39950.4
	Resi.	3.5759e-02	3.3521e-02	3.7375e-02
	Memory	2461.9	2461.9	2461.9
$N = 5$	#iter	250	250	250
	Rank	84	84	84
	CPU	91366.2	90544.0	90021.2
	Resi.	2.1672e-02	2.3056e-02	3.1080e-02
	Memory	4068.8	4068.8	4068.8
$N = 6$	#iter	250	250	250
	Rank	126	126	126
	CPU	208643.4	207964.0	207464.4
	Resi.	1.8357e-02	1.8064e-02	2.0494e-02
	Memory	6130.7	6130.7	6130.7
$N = 7$	#iter	250	250	250
	Rank	180	180	180
	CPU	355115.9	355167.5	355652.3
	Resi.	1.1208e-02	1.3833e-02	1.4914e-02
	Memory	8808.8	8808.8	8808.8

expansion and the regularization parameter μ for $\kappa_z = 0.5$ in Table 5.5 while holding the other parameters fixed. The difficulty of the task, as measured by the number of ranks, memory, and CPU time, grows as N increases. Another important finding is that when N increases, the relative residual declines regardless of the value of μ.

Next, the effect of the standard deviation parameter κ_z is investigated on the numerical simulations. The behaviours of the cost functional $\mathcal{J}(u_h)$, the tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$, and the relative residual for different values of κ_z are shown in Figure 5.1. It can be seen that when the value of κ_z rises, the values of $\mathcal{J}(u_h)$ and $\|y_h - y^d\|^2_{\mathcal{X}}$ monotonically decrease. Moreover, the low-rank version of the preconditioned GMRES algorithm produces convergence behaviour for all κ_z values. Last but not least, Figure 5.2 shows that the speed of convergence of relative residual decreases by in-

creasing the value of risk–aversion parameter γ in the beginning of the iteration.

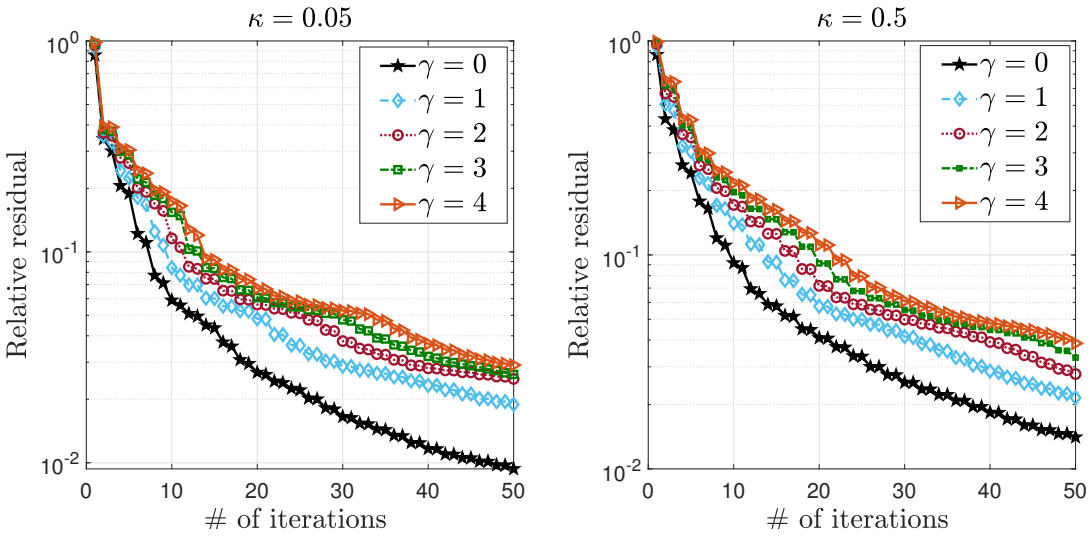

Figure 5.2: Example 5.5.1: Convergence of LRPGMRES with $N_d = 6144$, $N = 5$, $Q = 3$, $\ell = 1$, $\mu = 1$, and $\nu = 1$ for varying κ_z and γ.

5.5.2 Unconstrained Problem with Random Convection Parameter

Our second example is an unconstrained optimal control problem containing random velocity input parameter. To be precise, we set the deterministic diffusion parameter $a(\boldsymbol{x}, \omega) = \nu > 0$, the deterministic source function $f(\boldsymbol{x}) = 0$, and homogeneous Dirichlet boundary conditions on the spatial domain $\mathcal{D} = [-1, 1]^2$. On the other hand, the random velocity field $\mathbf{b}(\boldsymbol{x}, \omega)$ is defined as $\mathbf{b}(\boldsymbol{x}, \omega) = (z(\boldsymbol{x}, \omega), z(\boldsymbol{x}, \omega))^T$, where the random input $z(\boldsymbol{x}, \omega)$ has the unity mean, i.e., $\overline{z}(\boldsymbol{x}) = 1$. Further, the desired state y^d is given by

$$y^d(\boldsymbol{x}) = \exp\left[-64\left(\left(x_1 - \frac{1}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^2\right)\right].$$

Figure 5.3 and 5.4 display, respectively, the mean of state $\mathbb{E}[y_h]$ and the control u_h for varied values of the regularization parameter μ obtained by solving the full–rank system $\mathcal{L}\backslash\mathcal{B}$. As the previous example, we observe that the state y_h becomes closer to the target solution y^d while μ decreases.

Next, we compare the full–rank solutions obtained by solving the system $\mathcal{L}\backslash\mathcal{B}$ with the low–rank ones. Figure 5.5 exhibits the behaviours of the cost functional $\mathcal{J}(u_h)$

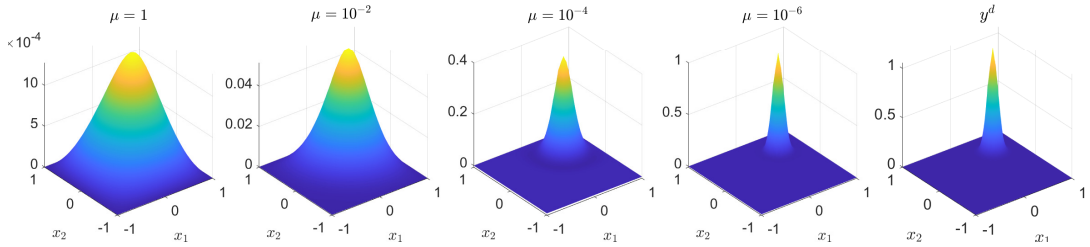

Figure 5.3: Example 5.5.2: Simulations of the mean of state $\mathbb{E}[y_h]$ obtained by $\mathcal{L}\backslash\mathcal{B}$ with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $\nu = 1$, and $\gamma = 0$ for varying $\mu = 1, 10^{-2}, 10^{-4}, 10^{-6}$ and the desired state y^d.

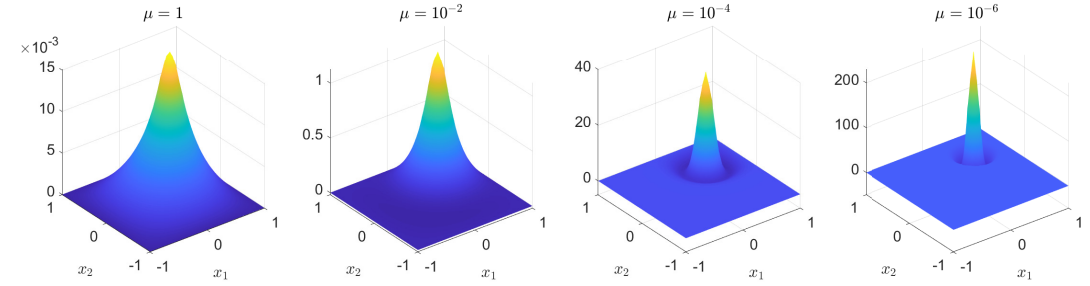

Figure 5.4: Example 5.5.2: Simulations of the control u_h obtained solving by $\mathcal{L}\backslash\mathcal{B}$ with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, $\nu = 1$, and $\gamma = 0$ for varying regularization parameter $\mu = 1, 10^{-2}, 10^{-4}, 10^{-6}$.

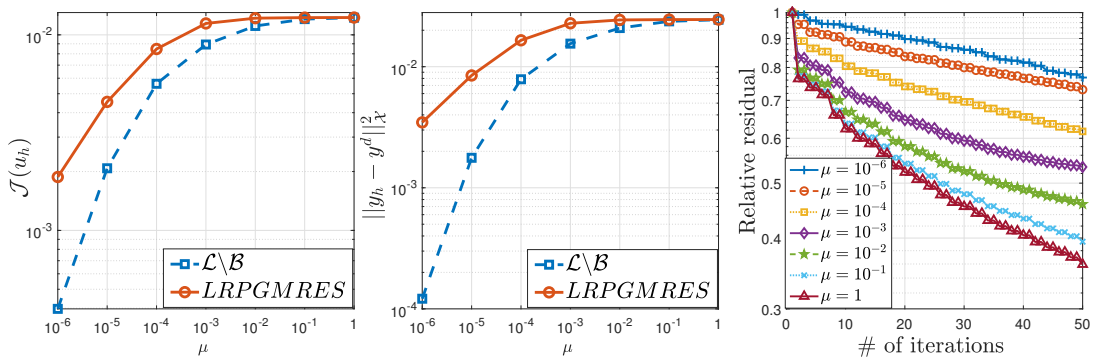

Figure 5.5: Example 5.5.2: Behaviours of the cost functional $\mathcal{J}(u_h)$ (left), the tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$ (middle), and the relative residual (right) with $N_d = 6144$, $N = 3$, $Q = 3$, $\kappa_z = 0.05$, $\ell = 1$, $\nu = 1$, $\gamma = 0$, and the mean-based preconditioner $\mathcal{P}_0$ for varying μ.

(left), the tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$ (middle), and the relative residual (right) for varying values of the regularization parameter μ. The key observation is that the low–rank solutions display the same pattern with the full–rank solutions as μ increases. Moreover, Table 5.6 reports the results of the simulations by considering various values of the risk–aversion parameter γ. As the previous example, the relative residual becomes smaller as decreasing the value of γ.

Table 5.6: Example 5.5.2: Simulation results showing total number of iterations, ranks of the truncated solutions, total CPU times (in seconds), relative residual, and memory demand of the solution (in KB) with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\nu = 1$, $\mu = 10^{-6}$, and the mean-based preconditioner $\mathcal{P}_0$ for varying γ.

	$\gamma = 0$	$\gamma = 10^{-6}$	$\gamma = 10^{-4}$	$\gamma = 10^{-2}$	$\gamma = 1$
#iter	250	250	250	250	250
Rank	29	30	30	30	21
CPU	24468.2	19383.2	17382.0	17422.8	17797.1
Resi.	2.1733e-01	2.6663e-01	4.0428e-01	6.9542e-01	9.1911e-01
Memory	1396.5	1444.7	1444.7	1444.7	963.2

Last, we investigate the effect of the mean of random input $z(\boldsymbol{x})$ on both full–rank and low–rank solutions. Denoting the full–rank solution and the low-rank solution by y_f and y_l, respectively, the behaviour of the differences $\|y_f - y^d\|^2_{\mathcal{X}}$, $\|y_l - y^d\|^2_{\mathcal{X}}$, and $\|y_f - y_l\|^2_{\mathcal{X}}$ computed by solving the full–rank and low–rank systems is displayed in Figure 5.6. As increasing the mean of random input $z(\boldsymbol{x})$, the difference between the full–rank and low–rank solutions becomes smaller.

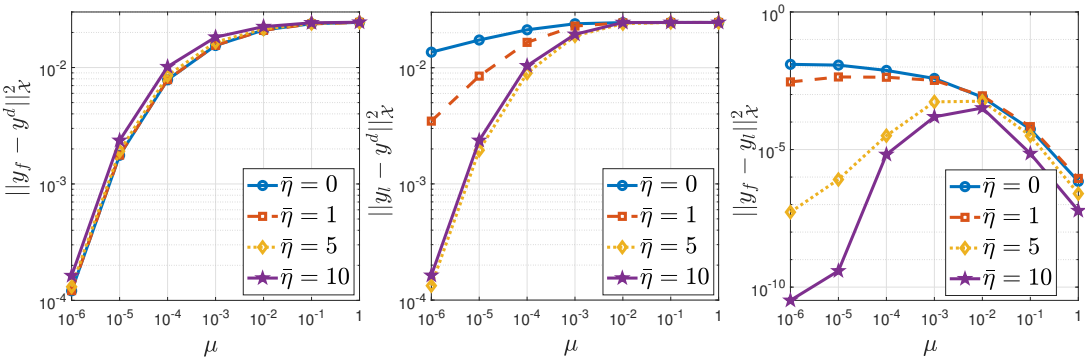

Figure 5.6: Example 5.5.2: Behaviour of the differences $\|y_f - y^d\|^2_{\mathcal{X}}$ (left), $\|y_l - y^d\|^2_{\mathcal{X}}$ (middle), and $\|y_f - y_l\|^2_{\mathcal{X}}$ (right), where the full–rank and low–rank solutions are denoted by y_f and y_l, respectively, computed by solving the full–rank and low–rank systems with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\mu = 10^{-6}$, $\gamma = 0$, $\nu = 1$, and $\kappa_z = 0.05$ for varying values of the mean of random input $z(x)$.

5.5.3 Constrained Problem with Random Convection Parameter

Last, a constrained optimal control problem containing a random velocity parameter is considered. With the exception of Example 5.5.2, there exists an upper bound for the control variable such as $u_b = 100$. The regularization and risk–averse parameters are selected as $\mu = 10^{-6}$ and $\gamma = 0$, respectively, taking the findings from the previous case into consideration.

The desired state y^d, the mean of the state $\mathbb{E}[y_h]$, and the control u_h obtained by $\mathcal{L}\backslash\mathcal{B}$ are shown in Figure 5.7. It is noted that the upper bound of the control constrained is satisfied. Table 5.7 compares the low–rank solutions with the full–rank ones. As increasing the truncation number N, we obtain better results as expected.

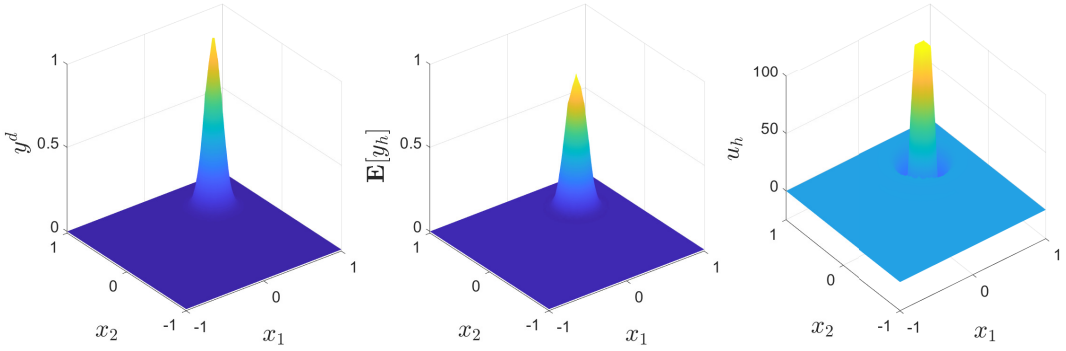

Figure 5.7: Example 5.5.3: Simulations of the desired state y^d, the mean of state $\mathbb{E}[y_h]$, and the control u_h (from left to right) obtained by $\mathcal{L}\backslash\mathcal{B}$ with $N_d = 6144$, $N = 3$, $Q = 3$, $\ell = 1$, $\kappa_z = 0.05$, and $\nu = 1$.

Table 5.7: Example 5.5.3: Simulation results showing the memory demand of the solution (in KB), the objective function $\mathcal{J}(u_h)$, the tracking term $\|y_h - y^d\|^2_{\mathcal{X}}$, the difference of the full–rank and low–rank $\|y_f - y_l\|^2_{\mathcal{X}}$, ranks of the truncated solutions, and the relative residual with $N_d = 6144$, $Q = 3$, $\ell = 1$, $\nu = 1$, and the mean-based preconditioner $\mathcal{P}_0$.

	Memory	$\mathcal{J}(u_h)$	$\|y_h - y^d\|^2_{\mathcal{X}}$	$\|y_f - y_l\|^2_{\mathcal{X}}$	Rank	Res.
$N = 3$	5744.0	5.508e-04	6.031e-04			
$N = 3$	1444.7	1.046e-02	2.091e-02	1.802e-02	30	9.232e-01
$N = 4$	2461.9	1.029e-02	2.056e-02	1.769e-02	51	9.161e-01
$N = 5$	4068.8	9.996e-03	1.996e-02	1.713e-02	84	9.042e-01
$N = 6$	6130.7	9.616e-03	1.919e-02	1.642e-02	126	8.895e-01

5.6 Discussion

In this chapter, the statistical moments of a robust deterministic optimal control problem subject to a convection diffusion equation having random coefficients have numerically been studied. Based on the finding in Chapter 3, the original problem is turned into a large system of deterministic optimal control problems for each realization of the random coefficients using the stochastic discontinuous Galerkin method. Nevertheless, some numerical results could not be reported when increasing the value of truncation number N. As a result, a low–rank variant of GMRES iteration (LRPGMRES) with a mean–based preconditioner, which reduces computational time and memory needs, has been used to break the curse of dimensionality problem. It has been shown in the numerical simulations that LRPGMRES can be an alternative to solve such large systems.

CHAPTER 6

CONCLUDING REMARKS

In this thesis, we have mainly investigated the numerical behaviour of a single partial differential equation, namely, convection diffusion, containing random coefficients and then have extended our findings to an optimal control problem constrained by the underlying PDE with uncertain terms.

The starting point has been the state equation which is the convection diffusion equation containing random inputs in Chapter 3. With the help of the stochastic Galerkin approach, we have transformed the original problem into a system consisting of deterministic convection diffusion equations for each realization of random coefficients. On the other hand, the symmetric interior penalty discontinuous Galerkin method has been applied to discretize the spatial domain due to its local mass conservativity, which is a crucial property for convection dominated problems. To solve the large–size and computationally challenging linear systems emerging from the stochastic Galerkin method, we have used the low–rank Krylov subspace methods, which reduce memory demand and computational costs. It has been seen in the numerical simulations that the low-rank GMRES algorithm is more efficient than CG, BiCGstab, and QMRCGstab for the convection dominated models.

Next, the focus of Chapter 4 has been on the efficient adaptive stochastic discontinuous Galerkin methods for the numerical solution of convection dominated equations with parameter dependent inputs. SG methods allow the separation of the spatial and stochastic variables, which provides a reuse of established numerical techniques such as a posterior error analysis, adaptive refinement in the physical, and adaptive enrichment probability domains. The boundary and/or interior layers, which cause

difficulties in computing numerical solutions to convection dominated problems, are resolved by applying adaptive stochastic discontinuous Galerkin methods driven by a residual–based posteriori error estimator. Promising numerical examples have been provided, opening the door to a variety of perspectives, such as combination with spatial mesh refinement and index enrichment of basis polynomials.

In the last part of the thesis, the findings in Chapter 3 have been extended to PDE-constrained optimization problems in Chapter 5. In addition to the single convection diffusion equations, the numerical solution of the optimization problems governed by convection diffusion PDEs suffers from additional memory requirements and computational complexity. In spite of the nice properties exhibited by the stochastic discontinuous Galerkin method, the dimension of the resulting linear system increases rapidly. As a remedy, we have applied a low–rank variant of the generalized minimal residual (GMRES) method [133] with a suitable preconditioner based on the results in Chapter 3. The numerical simulations of benchmark examples have shown that the low-rank iterative solver, especially the GMRES, is efficient in handling large-size problems.

As a future study, it will be interesting that randomness can be considered in different forms, for instance, in boundary conditions, desired state, or geometry; see, e.g., [85, 154]. Moreover, the adaptivity concepts can be extended to PDE–constrained optimization problems since the optimality conditions for the optimization problems governed by convection diffusion PDEs with uncertainty involve not only the original convection diffusion state equation, but also another convection diffusion PDE, the so-called adjoint PDE, so we refer to [157, 158, 160] and references therein. The diffusion part of the adjoint PDE is equal to that of the state PDE, but the convection in the adjoint PDE is equal to the negative of the convection in the state PDE. This has important implications for the behaviour of the solution, as well as for numerical methods. When convection dominates diffusion, the layers are generated in the state PDE as well as in the adjoint PDE, and are determined by the convection as well as by its negative; see, e.g., [70, 93]. This causes more difficulties than studying the solution of a single convection diffusion PDE. Recent studies in [23, 111] show that discontinuous Galerkin (DG) discretizations enjoy a better convergence behaviour for convection dominated optimal control problems.

REFERENCES

[1] R. Abgrall and S. Mishra, Uncertainty quantification for hyperbolic systems of conservation laws, in R. Abgrall and C.-W. Shu, editors, *Handbook of Numerical Methods for Hyperbolic Problems*, volume 18 of *Handbook of Numerical Analysis*, pp. 507–544, Elsevier, 2017.

[2] R. Adams, *Sobolev Spaces*, Academic Press, Orlando, San Diego, New-York, 1975.

[3] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics, John Wiley & Sons, New York, 2000.

[4] T. Akman, H. Yücel, and B. Karasözen, A priori error analysis of the upwind symmetric interior penalty Galerkin (SIPG) method for the optimal control problems governed by unsteady convection diffusion equations, Computational Optimization and Applications, 57, pp. 703–729, 2014.

[5] A. A. Ali, E. Ullmann, and M. Hinze, Multilevel Monte Carlo analysis for optimal control of elliptic PDEs with random coefficients, SIAM/ASA Journal on Uncertainty Quantification, 5, pp. 466–492, 2017.

[6] D. N. Arnold, An interior penalty finite element method with discontinuous elements, SIAM Journal on Numerical Analysis, 19(4), pp. 742–760, 1982.

[7] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM Journal on Numerical Analysis, 39(5), pp. 1749–1779, 2002.

[8] I. Babuška and P. Chatzipantelidis, On solving elliptic stochastic partial differential equations, Computer Methods in Applied Mechanics and Engineering, 191(37-38), pp. 4093–4122, 2002.

[9] I. Babuška, F. Nobile, and R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM Review, 52(2), pp. 317–355, 2010.

[10] I. Babuška and W. C. Rheinboldt, Error estimates for adaptive finite element computations, SIAM Journal on Numerical Analysis, 15, pp. 736–754, 1978.

[11] I. Babuška, R. Tempone, and G. E. Zouraris, Galerkin finite element approximations of stochastic elliptic partial differential equations, SIAM Journal on Numerical Analysis, 42(2), pp. 800–825, 2004.

[12] I. Babuška, R. Tempone, and G. E. Zouraris, Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation, Computer Methods in Applied Mechanics and Engineering, 194(12-16), pp. 1251–1294, 2005.

[13] J. Ballani and L. Grasedyck, A projection method to solve linear systems in tensor product, Numerical Linear Algebra with Applications, 20, pp. 27–43, 2013.

[14] A. V. Barel and S. Vandewalle, Robust optimization of PDEs with random coefficients using a multilevel Monte Carlo method, SIAM/ASA Journal on Uncertainty Quantification, 7(1), pp. 174–202, 2019.

[15] R. Becker and B. Vexler, Optimal control of the convection-diffusion equation using stabilized finite element methods, Numerische Mathematik, 106(3), pp. 349–367, 2007.

[16] P. Benner and T. Breiten, Low rank methods for a class of generalized Lyapunov equations and related issues, Numerische Mathematik, 124(3), pp. 441–470, 2013.

[17] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, Low–rank solvers for unsteady Stokes–Brinkman optimal control problem with random data, Computer Methods in Applied Mechanics and Engineering, 304, pp. 26–54, 2016.

[18] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, Low-rank solvers for unsteady Stokes–Brinkman optimal control problem with random data, Computer Methods in Applied Mechanics and Engineering, 304, pp. 26 – 54, 2016.

[19] P. Benner, S. Dolgov, A. Onwunta, and M. Stoll, Low-rank solution of an optimal control problem constrained by random Navier–Stokes equations, International Journal for Numerical Methods in Fluids, 92(11), pp. 1653–1678, 2020.

[20] P. Benner, A. Onwunta, and M. Stoll, Low-rank solution of unsteady diffusion equations with stochastic coefficients, SIAM/ASA Journal on Uncertainity Quantification, 3, pp. 622–649, 2015.

[21] P. Benner, A. Onwunta, and M. Stoll, Block-diagonal preconditioning for optimal control problems constrained by PDEs with uncertain inputs, SIAM Journal on Matrix Analysis and Applications, 37, pp. 491–518, 2016.

[22] P. Benner, A. Onwunta, and M. Stoll, On the existence and uniqueness of the solution of a parabolic optimal control problem with uncertain inputs, 2018, arXiv:1809.10645.

[23] P. Benner and H. Yücel, Adaptive symmetric interior penalty Galerkin method for boundary control problems, SIAM Journal on Numerical Analysis, 55(2), pp. 1101–1133, 2017.

[24] M. Bergounioux, K. Ito, and K. Kunisch, Primal-dual strategy for constrained optimal control problems, SIAM Journal on Control and Optimization, 37(4), pp. 1176–1194, 1999.

[25] A. Bespalov, C. E. Powell, and D. Silvester, Energy norm a posteriori error estimation for parametric operator equations, SIAM Journal on Scientific Computing, 36(2), pp. A339–A363, 2014.

[26] A. Bespalov, D. Praetorius, and M. Guggeri, Convergence and rate optimality of adaptive multilevel stochastic Galerkin FEM, IMA Journal of Numerical Analysis, 42, pp. 2190–2213, 2022.

[27] A. Bespalov, D. Praetorius, L. Rocchi, and M. Guggeri, Convergence of adaptive stochastic Galerkin FEM, SIAM Journal on Numerical Analysis, 57(5), pp. 2359–2382, 2019.

[28] A. Bespalov, D. Praetorius, L. Rocchi, and M. Guggeri, Goal–oriented error estimaton and adaptivity for ellpitic PDEs with parametric or uncertain inputs, Computer Methods in Applied Mechanics and Engineering, 345, pp. 951–982, 2019.

[29] A. Bespalov and L. Rocchi, Efficient adaptive algorithms for elliptic PDEs with random data, SIAM/ASA Journal on Uncertainty Quantification, 6(1), pp. 243–272, 2018.

[30] A. Bespalov and D. Silvester, Efficient adaptive stochastic Galerkin methods for parametric operator equations, SIAM Journal on Scientific Computing, 38(4), pp. A2118–A2140, 2016.

[31] L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders, editors, *Large-Scale PDE-Constrained Optimization*, Lecture Notes in Computational Science and Engineering, Vol. 30, Springer-Verlag, Heidelberg, 2003.

[32] A. Borzì, Multigrid and sparse-grid schemes for elliptic control problems with random coefficients, Computing and Visualization in Science, 13, pp. 153–160, 2010.

[33] A. Borzì, V. Schulz, C. Schillings, and G. von Winckel, On the treatment of distributed uncertainties in PDE constrained optimization, GAMM Mitteilungen, 33(2), pp. 230–246, 2010.

[34] A. Borzì and G. von Winckel, Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients, SIAM Journal on Scientific Computing, 31(3), pp. 2172–2192, 2009.

[35] A. Borzì and G. von Winckel, A POD framework to determine robust controls in PDE optimization, Computing and Visualization in Science, 14, pp. 91–103, 2011.

[36] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin, third edition, 2008.

[37] A. N. Brooks and T. J. R. Hughes, Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations, Computer Methods in Applied Mechanics and Engineering, 32, pp. 199–259, 1982.

[38] E. Burman and P. Hansbo, Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems, Computer Methods in Applied Mechanics and Engineering, 193(15), pp. 1437–1453, 2004.

[39] Z. Cai and S. Zhang, Recovery–based error estimator for interface problems: conforming linear elements, SIAM Journal on Numerical Analysis, 47, pp. 2132–2156, 2009.

[40] R. H. Cameron and W. T. Martin, The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals, Annals of Mathematics. Second Series, 48, pp. 385–392, 1947.

[41] P. Çiloğlu and H. Yücel, Stochastic discontinuous Galerkin methods with low–rank solvers for convection diffusion equations, Applied Numerical Mathematics, 172, pp. 157–185, 2022.

[42] P. Çiloğlu and H. Yücel, Stochastic discontinuous Galerkin methods for robust deterministic control of convection diffusion equations with uncertain coefficients, Advances in Computaional Mathematics, 49, 16, 2023.

[43] T. F. Chan, E. Gallopoulos, V. Simoncini, T. Szeto, and C. H. Tong, A quasi-minimal residual variant of the Bi-CGSTAB algorithm for nonsymmetric systems, SIAM Journal on Scientific Computing, 15(2), pp. 338–347, 1994.

[44] L. Chen, iFEM: an innovative finite element methods package in MATLAB, Technical report, Department of Mathematics, University of California, Irvine, CA 92697-3875, 2008.

[45] P. Chen, A. Quarteroni, and G. Rozza, Stochastic optimal Robin boundary control problems of advection-dominated elliptic equations, SIAM Journal on Numerical Analysis, 51, pp. 2700–2722, 2013.

[46] M. Christie, V. Demyanov, and D. Erbas, Uncertainty quantification for porous media flows, Journal of Computational Physics, 217(34), pp. 143–158, 2006.

[47] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics Appl. Math.*, SIAM, Philadelphia, PA, 2002.

[48] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup, Multilevel Monte Carlo methods and applications to elliptic PDEs with random coeffcients, Computing and Visualization in Science, 14, pp. 3–15, 2011.

[49] B. Cockburn and C.-W. Shu, The local discontinuous Galkerin method for time-dependent convection-diffusion systems, SIAM Journal on Numerical Analysis, 35, pp. 2440–2463, 1998.

[50] S. S. Collis and M. Heinkenschloss, Analysis of the streamline upwind/Petrov Galerkin method applied to the solution of optimal control problems, Technical Report TR02–01, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005–1892, 2002.

[51] A. Crowder, C. E. Powell, and A. Bespalov, Efficient adaptive multilevel stochastic Galerkin approximation using implicit a posteriori error estimation, SIAM Journal on Scientific Computing, 41, pp. A1681–A1705, 2019.

[52] C. Dawson, S. Sun, and M. Wheeler, Compatible algorithms for coupled flow and transport, Computer Methods in Applied Mechanics and Engineering, 193, pp. 2565–2580, 2004.

[53] J. P. Delhomme, Spatial variability and uncertainity in groundwater flow parameters: A geostatistical approach, Water Resources Research, pp. 269–280, 1979.

[54] D. A. Di Pietro and A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*, volume 69 of *Mathématiques et Applications*, Springer, New York, 2012.

[55] S. Dolgov, B. Khoromskij, A. Litvinenko, and H. G. Matthies, Polynomial chaos expansion of random coefficients and the solution of stochastic partial differential equations in the tensor train format, SIAM/ASA Journal on Uncertainty Quantification, 3, pp. 1109–1135, 2015.

[56] J. Dürrwächter, T. Kuhn, F. Meyer, L. Schlachter, and F. Schneider, A hyperbolicity-preserving discontinuous stochastic Galerkin scheme for uncertain hyperbolic systems of equations, Journal of Computational and Applied Mathematics, 370, p. 112602, 2020.

[57] C. Eckart and G. M. Young, The approximation of one matrix by another of lower rank, Psychometrika, 1, pp. 211–218, 1936.

[58] M. Eiermann, O. G. Ernst, and E. Ullmann, Computational aspects of the stochastic finite element method, Computing and Visualization in Science, 10(1), pp. 3–15, 2007.

[59] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, Adaptive stochastic Galerkin FEM, Computer Methods in Applied Mechanics and Engineering, 270, pp. 247–269, 2014.

[60] M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander, A convergent adaptive stochastic Galerkin finite element method with quasi–optimal spatial meshes,

ESAIM: Mathematical Modelling and Numerical Analysis, 49(5), pp. 1367–1398, 2015.

[61] M. Eigel, M. Marschall, M. Pfeffer, and R. Schneider, Adaptive stochastic Galerkin FEM for lognormal coefficients in hierarchical tensor representations, Numerische Mathematik, 145, pp. 655–692, 2020.

[62] M. Eigel and C. Merdon, Local equilibration error estimators for guaranteed error control in adaptive stochastic higher–order Galerkin finite elements, SIAM/ASA Journal on Uncertainty Quantification, 4, pp. 132–1397, 2016.

[63] H. C. Elman and T. Su, A low–rank multigrid method for the stochastic steady–state diffusion problems, SIAM Journal on Matrix Analysis and Applications, 39(1), pp. 492–509, 2018.

[64] H. C. Elman and T. Su, A low–rank solver for the stochastic unsteady Navier–Stokes problem, Computer Methods in Applied Mechanics and Engineering, 364, p. 112948, 2020.

[65] O. G. Ernst, C. E. Powell, D. J. 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), pp. 1424–1447, 2009.

[66] O. G. Ernst and E. Ullmann, Stochastic Galerkin matrices, SIAM Journal on Matrix Analysis and Applications, 31, pp. 1848–1872, 2010.

[67] G. S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, New York, 1996.

[68] P. Frauenfelder, C. Schwab, and R. Todor, Finite elements for elliptic problems with stochastic coefficients, Computer Methods in Applied Mechanics and Engineering, 194, pp. 205–228, 2005.

[69] M. A. Freitag and D. L. H. Green, A low–rank approach to the solution of weak constraint variational data assimilation problems, Journal of Computational Physics, 357, pp. 263–281, 2018.

[70] H. Fu and H. Rui, Adaptive characteristic finite element approximation of convection-diffusion optimal control problems, Numerical Methods for Partial Differential Equations, 29(3), pp. 979–998, 2012.

[71] S. Garreis and M. Ulbrich, Constrained optimization with low-rank tensors and applications to parametric problems with PDEs, SIAM Journal on Scientific Computing, 39, pp. A25–A54, 2017.

[72] L. Ge and T. Sun, A sparse grid stochastic collocation discontinuous Galerkin method for constrained optimal control problem governed by random convection dominated diffusion equations, Numerical Functional Analysis and Optimization, 40, pp. 763–797, 2019.

[73] R. Ghanem and S. Dham, Stochastic finite element analysis for multiphase flow in heterogeneous porous media, Transport in Porous Media, 32(3), pp. 239–262, 1998.

[74] R. Ghanem, D. Higdon, and H. Owhadi, *Handbook of Uncertainty Quantification*, Springer International Publishing, 2017.

[75] R. G. Ghanem and R. M. Kruger, Numerical solution of spectral stochastic finite element systems, Computer Methods in Applied Mechanics and Engineering, 129(3), pp. 289–303, 1996.

[76] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements:A Spectral Approach*, Springer-Verlag, 1991.

[77] S. Ghori, J. Heller, and A. Singh, An efficient method of generating random permeability fields by the source point method, Mathematical Geosciences, 25, pp. 559–572, 1993.

[78] M. B. Giles, Multilevel Monte Carlo path simulation, Operations Research, 56, pp. 607–617, 2008.

[79] C. J. Gittelson, An adaptive stochastic Galerkin method, Mathematics of Computation, 82, pp. 1515–1541, 2011.

[80] G. Golub, A. Hoffman, and G. Stewart, A generalization of the Eckart-Young-Mirsky matrix approximation theorem, Linear Algebra and its Applications, 88-89, pp. 317–327, 1987.

[81] G. H. Golub and C. F. V. Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[82] A. Graham, *Kronecker Products and Matrix Calculus: with Applications*, Ellis Horwood Series in Mathematics and Its Applications, Horwood, 1981.

[83] L. Grasedyck, Existence and computation of low kronecker-rank approximations for large linear systems of tensor product structure, Computing, 72, pp. 247–265, 2004.

[84] D. Guignard, F. Nobile, and M. Picasso, A posteriori error estimation for elliptic partial differential equations with small uncertainties, Numerical Methods for Partial Differential Equations, 32(1), pp. 175–212, 2015.

[85] D. Guignard, F. Nobile, and M. Picasso, A posteriori error estimation for the steady Navier–Stokes equations in random domains, Computer Methods in Applied Mechanics and Engineering, 313, pp. 483–511, 2017.

[86] M. D. Gunzburger, H.-C. Lee, and J. Lee, Error estimates of stochastic optimal Neumann boundary control problems, SIAM Journal on Numerical Analysis, 49(4), pp. 1532–1552, 2011.

[87] P. A. Guth, V. Kaarnioja, F. Y. Kuo, C. Schillings, and I. H. Sloan, A quasi-Monte Carlo method for optimal control under uncertainty, SIAM/ASA Journal on Uncertainty Quantification, 9(2), pp. 354–383, 2021.

[88] S. Heinrich, Multilevel Monte Carlo methods, in S. Margenov, J. Waśniewski, and P. Yalamov, editors, *Large-Scale Scientific Computing*, pp. 58–67, Springer Berlin Heidelberg, 2001.

[89] M. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, Journal of Research of the National Institute of Standards and Technology, 49, pp. 409–436, 1952.

[90] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods: Analysis, Algorithms, and Applications*, Springer, Berlin, 2008.

[91] J. G. Heywood and R. Rannacher, Finite-element approximation of the nonstationary Navier-Stokes problem. IV. Error analysis for second-order time discretization, SIAM Journal on Numerical Analysis, 27(2), pp. 353–384, 1990.

[92] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with Partial Differential Equations*, volume 23 of *Mathematical Modelling, Theory and Applications*, Springer, Heidelberg, 2009.

[93] M. Hinze, N. Yan, and Z. Zhou, Variational discretization for optimal control governed by convection dominated diffusion equations, Journal of Computational Mathematics, 27(2-3), pp. 237–253, 2009.

[94] L. S. Hou, J. Lee, and H. Manouzi, Finite element approximations of stochastic optimal control problems constrained by stochastic elliptic PDEs, Journal of Mathematical Analysis and Applications, 384, pp. 87–103, 2011.

[95] P. Houston, D. Schötzau, and T. P. Wihler, Energy norm a-posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems, Mathematical Models and Methods in Applied Sciences, 17, pp. 33–62, 2007.

[96] O. A. Karakashian and F. Pascal, Convergence of adaptive discontinuous Galerkin approximations of second-order elliptic problems, SIAM Journal on Numerical Analysis, 45(2), pp. 641–665, 2007.

[97] K. Karhunen, Über lineare Methoden in der Wahrscheinlichkeitsrechnung, Annales Academiae Scientiarum Fennicae. Series A. Mathematica - Physica, 1947(37), p. 79, 1947.

[98] G. E. Karniadakis, C.-H. Su, D. Xiu, D. Lucor, C. Schwab, and R. A. Todor, Generalized polynomial chaos solution for differential equations with random inputs, Technical Report 2005-01, Seminar for Applied Mathematics, ETH Zurich, Zurich, Switzerland, 2005.

[99] M. Kleiber and T. D. Hien, *The Stochastic Finite Element Method: Basic perturbation technique and computer implementation*, Wiley, 1992.

[100] R. Koekoek and P. A. Lesky, *Hypergeometric Orthogonal Polynomials and Their q–Analogues*, Springer-Verlag, 2010.

[101] D. P. Kouri, M. Heinkenschloss, D. Ridzal, and B. G. van Bloemen Waanders, A trust-region algorithm with adaptive stochastic collocation for PDE optimization under uncertainty, SIAM Journal on Scientific Computing, 35, pp. A1847–A1879, 2013.

[102] D. Kressner and C. Tobler, Low-rank tensor Krylov subspace methods for parametrized linear systems, SIAM Journal on Matrix Analysis and Applications, 32(4), pp. 1288–1316, 2011.

[103] M. Kubinova and I. Pultarova, Block preconditioning of stochastic Galerkin problems: New two-sided guaranteed spectral bounds, SIAM/ASA Journal on Uncertainty Quantification, 8(1), pp. 88–113, 2020.

[104] A. Kunoth and C. Schwab, Analytic regularity and gPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs, SIAM Journal on Control and Optimization, 51(3), pp. 2442–2471, 2013.

[105] A. Kunoth and C. Schwab, Sparse adaptive tensor Galerkin approximations of stochastic PDE-constrained control problems, SIAM/ASA Journal on Uncertainity Quantification, 4, pp. 1034–1059, 2016.

[106] M. Lazar and E. Zuazua, Averaged control and observation of parameter-depending wave equations, Comptes Rendus de l'Académie des Sciences, 352, pp. 497–502, 2014.

[107] O. P. Le Maitre and O. M. Knio, *Spectral Methods for Uncertainty Quantification With Applications to Computational Fluid Dynamics*, Scientific Computation, Springer-Verlag, Berlin, 2010.

[108] H.-C. Lee and J. Lee, A stochastic Galerkin method for stochastic control problems, Communications in Computational Physics, 14, pp. 77–106, 2013.

[109] K. Lee and H. C. Elman, A preconditioned low-rank projection method with a rank-reduction scheme for stochastic partial differential equations, SIAM Journal on Scientific Computing, 39(5), pp. S828–S850, 2017.

[110] K. Lee, H. C. Elman, and B. Sousedík, A low–rank solver for the Navier–Stokes equations with uncertainity viscosity, SIAM/ASA Journal on Uncertainty Quantification, 7(4), pp. 1275–1300, 2019.

[111] D. Leykekhman and M. Heinkenschloss, Local error analysis of discontinuous Galerkin methods for advection-dominated elliptic linear-quadratic optimal control problems, SIAM Journal on Numerical Analysis, 50(4), pp. 2012–2038, 2012.

[112] R. Li, W. Liu, H. Ma, and T. Tang, Adaptive finite element approximation for distributed elliptic optimal control problems, SIAM Journal on Control and Optimization, 41(5), pp. 1321–1349, 2002.

[113] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[114] K. Liu, Discontinuous Galerkin methods for parabolic partial differential equations with random input, 2013.

[115] K. Liu and B. Rivière, Discontinuous Galerkin methods for elliptic partial differential equations with random coefficients, International Journal of Computer Mathematics, 90(11), pp. 2477–2490, 2013.

[116] M. Loève, Fonctions aléatoires de second ordre, La Revue Scientifique, 84, pp. 195–206, 1946.

[117] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge University Press, New York, 2014.

[118] X. Ma and N. Zabaras, An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations, Journal of Computational Physics, 228(8), pp. 3084–3113, 2009.

[119] S. Markov, B. Cheng, and A. Asenov, Statistical variability in fully depleted SOI MOSFETs due to random dopant fluctuations in the source and drain extensions, IEEE Electron Device Letters, 33(3), pp. 315–317, 2012.

[120] L. Mathelin and O. L. Maître, Dual–based a posteriori error estimate for stochastic finite element methods, Communications in Applied Mathematics and Computational Science, 2, pp. 83–115, 2007.

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

[122] H. G. Matthies and E. Zander, Solving stochastic systems with low–rank tensor compression, Linear Algebra and its Applications, 436, pp. 3819–3838, 2012.

[123] D. Meidner and B. Vexler, A priori error estimates for space-time finite element discretization of parabolic optimal control problems. I. Problems without control constraints, SIAM Journal on Control and Optimization, 47(3), pp. 1150–1177, 2008.

[124] W. J. Morokoff and R. E. Caflisch, Quasi-Monte Carlo integration, Journal of Computational Physics, 122(2), pp. 218–230, 1995.

[125] F. Negri, A. Manzoni, and G. Rozza, Reduced basis approximation of parametrized optimal flow control problems for the Stokes equations, Computers & Mathematics with Applications, 69(4), pp. 319–336, 2015.

[126] F. Nobile, R. Tempone, and C. G. Webster, A sparse grid stochastic collocation method for partial differential equations with random input data, SIAM Journal on Numerical Analysis, 46(5), pp. 2309–2345, 2008.

[127] B. Øksendal, *Stochastic Differential Equations*, Springer-Verlag, Berlin, 2003.

[128] C. E. Powell and H. C. Elman, Block–diagonal preconditioning for spectral stochastic finite–element systems, IMA Journal of Numerical Analysis, 29(2), pp. 350–375, 2009.

[129] B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations. Theory and Implementation*, Frontiers in Applied Mathematics, SIAM, Philadelphia, 2008.

[130] B. Rivière, M. F. Wheeler, and V. Girault, Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems, Computational Geosciences, 8, pp. 231–244, 1999.

[131] E. Rosseel and G. N. Wells, Optimal control with stochastic PDE constraints and uncertain controls, Computer Methods in Applied Mechanics and Engineering, 213/216, pp. 152–167, 2012.

[132] J.-S. Ryu, M.-S. Kim, K.-J. Cha, T. H. Lee, and D.-H. Choi, Kriging interpolation methods in geostatistics and DACE model, KSME International Journal, 16, pp. 619–632, 2002.

[133] Y. Saad and M. H. Schultz, GMRES a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM Journal on Scientific Computing, 7, pp. 856–869, 1986.

[134] P. Sarma, L. Durlofsky, and K. Aziz, Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics, Mathematical Geosciences, 40(1), pp. 3–32, 2008.

[135] B. Schölkopf, A. J. Smola, and K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation, 10, pp. 1299–1319, 1998.

[136] D. Schötzau and L. Zhu, A robust a-posteriori error estimator for discontinuous Galerkin methods for convection-diffusion equations, Applied Numerical Mathematics, 59(9), pp. 2236–2255, 2009.

[137] L. R. Scott and S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions, Mathematics of Computation, 54, pp. 483–493, 1990.

[138] K. G. Siebert, Mathematically founded design of adaptive finite element software, in *Multiscale and Adaptivity: Modeling, Numerics and Applications*, volume 2040 of *Lecture Notes in Mathematics*, pp. 227–309, Springer Berlin / Heidelberg, 2012.

[139] M. Stoll and T. Breiten, A low-rank in time approach to PDE-constrained optimization, SIAM Journal on Scientific Computing, 37(1), pp. B1–B29, 2015.

[140] T. Sun, W. Shen, B. Gong, and W. Liu, A priori error estimate of stochastic Galerkin method for optimal control problem governed by stochastic elliptic PDE with constrained control, Journal of Scientific Computing, 67, pp. 405–431, 2016.

[141] H. Tiesler, R. M. Kirby, D. Xiu, and T. Preusser, Stochastic collocation for optimal control problems with stochastic PDE constraints, SIAM Journal on Control and Optimization, 50(5), pp. 2659–2682, 2012.

[142] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, volume 112 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, RI, 2010.

[143] E. Ullmann, A Kronecker product preconditioner for stochastic Galerkin finite element discretizations, SIAM Journal on Scientific Computing, 32(2), pp. 923–946, 2010.

[144] H. A. van der Vorst, Bi-CGSTAB: A fast and smoothly converging variant of bi-CG for the solution of nonsymmetric linear systems, SIAM Journal on Scientific and Statistical Computing, 13(2), pp. 631–644, 1992.

[145] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh–Refinement Techniques*, Wiley-Teubner Series: Advances in Numerical Mathematics, Wiley Teubner, Chicester, New York, Stuttgart, 1996.

[146] X. Wan and G. E. Karniadakis, Error control in multi–element generalized polynomial choas method for elliptic problems with random coefficients, Communications in Computational Physics, 5, pp. 793–820, 2009.

[147] M. F. Wheeler, An elliptic collocation-finite element method with interior penalties, SIAM Journal on Numerical Analysis, 15, pp. 152–161, 1978.

[148] N. Wiener, The homogeneous chaos, American Journal of Mathematics, 60, pp. 897–938, 1938.

[149] C. Winter and D. Tartakovsky, Mean flow in composite porous media, Geophysical Research Letters, 27, pp. 1759–1762, 2000.

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

[151] D. Xiu and J. S. Hesthaven, High-order collocation methods for differential equations with random inputs, SIAM Journal on Scientific Computing, 27(3), pp. 1118–1139, 2005.

[152] D. Xiu and G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM Journal of Scientific Computing, 24(2), p. 619–644, 2002.

[153] D. Xiu and J. Shen, Efficient stochastic Galerkin methods for random diffusion equations, Journal of Computational Physics, 228, pp. 266–281, 2009.

[154] D. Xiu and D. M. Tartakovsky, Numerical methods for differential equations in random domains, SIAM Journal on Scientific Computing, 28(3), pp. 1167–1185, 2006.

[155] N. Yan and Z. Zhou, A priori and a posteriori error analysis of edge stabilization Galerkin method for the optimal control problem governed by convection-dominated diffusion equation, Journal of Computational and Applied Mathematics, 223(1), pp. 198–217, 2009.

[156] R. M. Yao and L. J. Bo, Discontinuous Galerkin method for elliptic stochastic partial differential equations on two and three dimensional spaces, Science in China Series A: Mathematics, 50, pp. 1661–1672, 2007.

[157] H. Yücel and P. Benner, Adaptive discontinuous Galerkin methods for state constrained optimal control problems governed by convection diffusion equations, Computational Optimization and Applications, 62, pp. 291–321, 2015.

[158] H. Yücel and P. Benner, Distributed optimal control problems governed by coupled convection dominated pdes with control constraints, in A. Abdulle, S. Deparis, D. Kressner, F. Nobile, and M. Picasso, editors, *Numerical Mathematics and Advanced Applications - ENUMATH 2013*, volume 103 of *Lecture Notes in Computational Science and Engineering*, pp. 469–478, Springer International Publishing, 2015.

[159] H. Yücel, M. Heinkenschloss, and B. Karasözen, Distributed optimal control of diffusion-convection-reaction equations using discontinuous Galerkin methods, in *Numerical Mathematics and Advanced Applications 2011*, pp. 389–397, Springer, Berlin, 2013.

[160] H. Yücel and B. Karasözen, Adaptive symmetric interior penalty Galerkin (SIPG) method for optimal control of convection diffusion equations with control constraints, Optimization, 63, pp. 145–166, 2014.

[161] S. Zein, V. Rath, and C. Clauser, A multidimensional Markov chain model for simulating stochastic permeability conditioned by pressure measures, International Journal of Multiphysics, 4, pp. 359–374, 2010.

[162] D. Zhang, *Stochastic Methods for Flow in Porous Media: Coping with Uncertainties*, Academic Press, San Diego, 2002.

[163] D. Zhang and Q. Kang, Pore scale simulation of solute transport in fractured porous media, Geophysical Research Letters, 31, p. L12504, 2004.

[164] Z. Zhou and N. Yan, The local discontinuous Galerkin method for optimal control problem governed by convection diffusion equations, International Journal of Numerical Analysis & Modeling, 7(4), pp. 681–699, 2010.

[165] E. Zuazua, Averaged control, Automatica, 50(12), pp. 3077–3087, 2014.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Çiloğlu, Pelin

EDUCATION

Degree	**Institution**	**Year of Graduation**
Ph.D.	Institute of Applied Mathematics, METU	2023
B.S.	Department of Mathematics, METU	2015
High School	Etimesgut Anadolu Lisesi	2010

PROFESSIONAL EXPERIENCE

Year	**Place**	**Enrollment**
Dec 2017 -	Institute of Applied Mathematics, METU	Research Assistant
Sep 2014 - Dec 2016	Department of Mathematics, METU	Student Assistant

PROJECTS

1. Numerical Studies for Petrol and Gas Reservoir Problems (TUBITAK 1001 - 119F022), August 2019 – August 2021.

2. Numerical Studies of Korteweg-de Vries Equation with Random Input Data (YOP-705-2018-2820), May 2018 – May 2019.

PUBLICATIONS

Journal Publications

1. Çiloğlu, P., and Yücel, H., Stochastic Discontinuous Galerkin Methods for Robust Deterministic Control of Convection Diffusion Equations with Uncertain Coefficients, Advances in Computational Mathematics, 49, 16, 2023, doi: 10.1007/s10444-023-10015-5.

2. Çiloğlu, P., and Yücel, H., Stochastic Discontinuous Galerkin Methods with Low–Rank Solvers for Convection Diffusion Equations, Applied Numerical Mathematics 172, 157-185, 2022, doi:10.1016/j.apnum.2021.10.007.

International Conference Presentations

1. Çiloğlu, P., and Yücel, H., Adaptive Discontinuous Galerkin Methods for PDEs with Random Data, Workshop: Numerical Analysis of Stochastic Partial Differential Equations (NASPDE) - Eurandom, Eindhoven, Netherlands, May 15-17, 2023.

2. Çiloğlu, P., and Yücel, H., Stochastic Discontinuous Galerkin Methods for Robust Deterministic Control of Convection Diffusion Equations with Uncertain Coefficients, Hybrid: SIAM Conference on Uncertainty Quantification (UQ22), Atlanta, Georgia, U.S., April 12-15, 2022.

3. Çiloğlu, P., and Yücel, H., Solving Optimal Control Problems Containing Uncertain Coefficients with Stochastic Discontinuous Galerkin Methods, 17th Copper Mountain Conference On Iterative Methods (Virtual), April 4-8, 2022.

4. Çiloğlu, P., and Yücel, H., Stochastic Discontinuous Galerkin Methods for Robust Deterministic Optimal Control, The SFB 1294 Spring School 2022, Schorfheide (Brandenburg), Germany, March 21-25, 2022.

5. Çiloğlu, P., and Yücel, H., Stochastic Discontinuous Galerkin Methods with Low-Rank Solvers for Convection Diffusion Equations, Chemnitz FE Symposium 2021, Chemnitz, Germany, September 6-8, 2021.

6. Çiloğlu, P., and Yücel, H., Unsteady Convection Diffusion Equation with Random Input Data, Chemnitz FE Symposium 2018, Chemnitz, Germany, September 24-26, 2018.

National Conference Presentations

1. Çiloğlu, P., and Yücel, H., Discontinuous Galerkin Methods for Convection Diffusion Equation with Random Coefficients, BEYOND 2019: Computational Science and Engineering Conference, Ankara, Turkey, September 9-11, 2019.
2. Çiloğlu, P., and Yücel, H., Discontinuous Galerkin Methods for Unsteady Convection Diffusion Equation with Random Coefficients, BEYOND: Workshop on Computational Science and Engineering, Ankara, Turkey, October 20-21, 2018.

Participation in Scientific Meetings

1. Summer School – Uncertainty, Adaptivity, and Machine Learning, Augsburg, Germany, September 12-14, 2022.
2. The SFB 1294 Spring School 2022, Schorfheide (Brandenburg), Germany, March 21-25, 2022.
3. Chemnitz Finite Element Symposium 2020, Chemnitz, Germany, September 14-17, 2020 (Online).
4. Annual Meeting of the German Mathematical Society (DMV Jahrestagung 2020), Chemnitz, Germany, September 14-17, 2020 (Online).
5. Third Graduate Summer School of Association for Turkish Women in Maths, Middle East Technical University, Ankara, Turkey, June 18-27, 2018.