STOCHASTIC MOMENTUM METHODS FOR OPTIMAL CONTROL PROBLEMS GOVERNED BY CONVECTION-DIFFUSION EQUATIONS WITH UNCERTAIN COEFFICIENTS

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

BY

SITKI CAN TORAMAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
SCIENTIFIC COMPUTING

JANUARY 2022

Approval of the thesis:

STOCHASTIC MOMENTUM METHODS FOR OPTIMAL CONTROL PROBLEMS GOVERNED BY CONVECTION-DIFFUSION EQUATIONS WITH UNCERTAIN COEFFICIENTS

submitted by **SITKI CAN TORAMAN** in partial fulfillment of the requirements for the degree of **Master of Science in Scientific Computing Department, Middle East Technical University** by,

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

Assoc. Prof. Dr. Hamdullah Yücel _____
Head of Department, **Scientific Computing**

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

Examining Committee Members:

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

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

Assoc. Prof. Dr. Murat Uzunca _____
Department of Mathematics, Sinop Univ.

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: SITKI CAN TORAMAN

Signature :

ABSTRACT

STOCHASTIC MOMENTUM METHODS FOR OPTIMAL CONTROL PROBLEMS GOVERNED BY CONVECTION-DIFFUSION EQUATIONS WITH UNCERTAIN COEFFICIENTS

Toraman, Sıtkı Can

M.S., Department of Scientific Computing

Supervisor : Assoc. Prof. Dr. Hamdullah Yücel

January 2022, 73 pages

Many physical phenomena such as the flow of an aircraft, or heating process, or wave propagation are modeled mathematically by differential equations, in particular partial differential equations (PDEs). Analytical solutions to PDEs are often unknown or very hard to obtain. Because of that, we simulate such systems by numerical methods such as finite difference, finite volume, or finite element, etc. When we want to control the behavior of certain system components, such as the shape of a wing of an aircraft or an applied heat distribution, it becomes equivalent to optimizing certain parameters of the underlying PDEs. Optimization of real-world systems in this way is called PDE-constrained optimization or optimal control problems. To have a more accurate mathematical model, we employ uncertain coefficients in PDEs since nature has different sources of intrinsic randomness. In this thesis, we study a numerical investigation of a strongly convex and smooth tracking-type functional subject to a convection-diffusion equation with random coefficients. In spatial dimension, we use the Finite Element Method (FEM), in probability dimension, we use the Monte Carlo (MC) method, and as an optimization method, we use the stochastic gradient (SG) method, where the true gradient is replaced by a stochastic one to minimize the expected value over a random function. To accelerate the convergence of the stochastic approach, momentum terms, i.e., Polyak's and Nesterov's momentums,

are added. A full error analysis including Monte Carlo, finite element, and stochastic momentum gradient iteration errors are done. Numerical examples are presented to illustrate the performance of the proposed stochastic approximations in the PDE-constrained optimization setting.

Keywords: Optimal Control, Stochastic PDEs, Convection-Diffusion, Stochastic Gradient with Momentum

ÖZ

BELİRSİZ KATSAYILI KONVEKSİYON-DİFÜZYON DENKLEMLERİNİN YÖNETTİĞİ OPTİMAL KONTROL PROBLEMLERİ İÇİN STOKASTİK MOMENTUM YÖNTEMLERİ

Toraman, Sıtkı Can
Yüksek Lisans, Bilimsel Hesaplama Bölümü
Tez Yöneticisi : Doç. Dr. Hamdullah Yücel

Ocak 2022, 73 sayfa

Bir uçağın uçuşu, ısıtma işlemi veya dalga yayılımı gibi birçok fiziksel olay, matematiksel olarak diferansiyel denklemler, özellikle kısmi türevli diferansiyel denklemler ile modellenir. Kısmi türevli diferansiyel denklemlere yönelik analitik çözümler genellikle bilinmemektedir veya elde edilmesi çok zordur. Bu nedenle, bu tür sistemler sonlu fark, sonlu hacim veya sonlu eleman ve benzeri gibi sayısal yöntemlerle çözülebilir. Bir uçağın kanadının şekli veya uygulanan bir ısı dağıtımı gibi belirli sistem bileşenlerinin davranışını kontrol etmek istediğimizde bu, kısmi türevli diferansiyel denklemin belirli parametrelerini optimize etmeye karşılık gelir. Gerçek dünya sistemlerinin bu optimizasyonuna kısmi türevli diferansiyel denklemler tarafından kısıtlı en iyileme problemleri veya optimal kontrol problemleri denir. Daha doğru bir matematiksel modele sahip olmak için, doğa farklı içsel rastgelelik kaynaklarına sahip olduğundan, kısmı türevli diferansiyel denklemler rastgelelik içeren parametrelerle ifade edilebilir. Bu tezde, rastgele katsayılı bir konveksiyon-difüzyon denklemine tabi olan güçlü dışbükey ve düzgün izleme tipi bir fonksiyonelin sayısal bir araştırması ele alınmıştır. Fiziksel boyutta sonlu elemanlar, olasılık boyutunda Monte Carlo ayrıklaştırma yöntemleri olarak kullanılırken, optimizasyon yöntemi olarak gerçek gradyanın stokastik varyantı ile değiştirildiği stokastik gradyan yöntemini kullanılmaktadır. Stokastik yaklaşımın yakınsamasını hızlandırmak için momentum terimleri, yani Polyak

ve Nesterov'un momentumları eklenmiştir. Monte Carlo, sonlu eleman ve stokastik momentum gradyan yineleme hatalarını içeren tam bir hata analizi yapılmaktadır, Son olarak kısmi türevli diferansiyel denklemler tarafından kısıtlı optimizasyon düzeneğinde önerilen stokastik yaklaşımların performansını göstermek için sayısal örnekler sunulmaktadır.

Anahtar Kelimeler: Optimal Kontrol, Stokastik Kısmi Türevli Diferansiyel Denklemler, Konveksiyon-Difüzyon, Momentumlu Stokastik Gradyan

To my mother

ACKNOWLEDGMENTS

I would like to express my very great appreciation to my thesis supervisor Assoc. Prof. Dr. Hamdullah Yücel for posing a very high ideal to strive for, and patiently, masterfully guiding my thesis process and education. His energy and dedication will inspire me for a lifetime.

Also, I give many thanks to our research group members for providing a proactive and friendly scientific environment, and I especially thank my friends Mehmet Alp Üreten and Pelin Çiloğlu for their contributions to this study.

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 program 1001 with project number 119F022.

To God, all praise and glory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Science is the most powerful tool we have to understand and manipulate the world we live in, and arguably it is founded solely on pragmatic reasons; to control our environment. In this pursuit, we model the physical phenomenon mathematically so that we can predict and change the material world over some mathematical variables. An important class of phenomena arising in science and engineering is modelled by partial differential equations (PDEs). To control a real-world system, one may desire to optimize specific parameters of a system described by a PDE in order to increase the quality of the outcome, e.g., optimal shapes of airplane wings or the temperature control of a melting process. Such problems can be formulated as optimal control problems or optimization problems with PDE constraints, called PDE-constrained optimization problems. In real-life applications, the PDE models of these systems contain uncertainties which we can categorize into two classes; epistemic/systematic uncertainties and intrinsic/aleatoric uncertainties; see, e.g., [52] for more details. The epistemic uncertainties can be caused by loose mathematical modeling, imprecise measurements, or by insufficient knowledge of the system's parameters, i.e., data errors. The latter class of uncertainties is intrinsic to the nature of the system, such as uncertainty in quantum systems. While solving problems with such types of uncertainties, the uncertainties are characterized by parameters that are not known, and hence the solutions are modeled as random variables. We may not know the parameters associated with, for example, coefficients, boundary data, initial conditions, or source terms. Such kinds of problems are called stochastic PDEs or PDEs with uncertainty. Optimal control problems involving such PDEs with uncertainty is a new research area and have become very active in the last decade. They have been stud-

ied in various formulations such as mean–based control [11, 12], pathwise control [2, 57], average control [44, 74], robust deterministic control [26, 33, 35, 40, 46, 63], and stochastic control [10, 18, 42, 69]. Here, we are interested in the robust control problem, which includes an appropriate statistical measure of the objective function to be minimized. Such kinds of problems are also called risk-averse optimal control. We refer to [1, 40, 41] for various forms of the risk measure such as expectation, expectation plus variance, a quantile, or a conditional expectation above a quantile. The reason of interest in the robust control problem is that the robust control problem is insensitive to parameter uncertainties and the optimum is valid for broad range of parameters as discussed in [9]. In this work, we mainly focus on the robust deterministic control problem constrained by the convection-diffusion equation with uncertain coefficients. Our choice of the constraint equation, the convection-diffusion equation, arises in many different applications such as; pollutant dispersal in a river estuary, atmospheric pollution, Fokker-Planck equation (which originates from Boltzmann equation of kinetic theory), semi-conductor equations, groundwater transport equation, viscous compressible flow past an aerofoil, and turbulence transport [56]. We are interested in the steady case, which most famously arises in vorticity transport in the incompressible Navier-Stokes equation when the Reynolds number is sufficiently small. Convection-diffusion equations containing randomness are studied with various solution methods in the literature; in [18, 48] stochastic collocation method combined with finite elements, in [73] stochastic Galerkin method combined with lattice Boltzmann method, and in [71] generalized polynomial chaos combined with spectral elements.

The solutions of PDEs with uncertainty are stochastic fields, which are infinite dimensional. To represent them in a finite setting, we parameterize the random coefficients by using the Karhunen-Loève (KL) expansion [51]. KL expansion is a spectral decomposition method, which aims to represent random variables in orthogonal decompositions. KL does that by using the eigenpairs of the covariance function of the random variable it represents. Compared to its alternatives, such as spectral representation, which expands a stochastic process in a sum of trigonometric functions with random phases and amplitudes, KL is cheaper [66]. With the stochastic variables parameterized, we are ready to apply numerical methods. The standard numerical

techniques for solving PDEs with uncertain data can be separated into two groups; intrusive and non-intrusive methods [67]. The former, the intrusive methods are non-sampling methods that need reformulation of the system of equations for all uncertain model variables as a whole. A good example of the intrusive methods for our case is the stochastic Galerkin (SG) method, which is a non-sampling method aiming to re-model the problem into a large system of deterministic problems such that the residue is orthogonal to the space of polynomials [67]. The latter, the non-intrusive methods, are based on the generation of independent and identically distributed ($i.i.d.$) samples of the random variable with respect to its probability distribution, and we focus on a widely used class of them; Monte Carlo (MC) sampling methods [67]. The MC methods are constructed by generating pseudo-random realizations of the random variable and by solving the corresponding deterministic problem at each realization. A set of such solutions enables us to derive statistical information on the random variable. Although they are easy to implement and naturally parallelizable, they require a large number of samples to obtain the desired accuracy. Still, as they don't suffer from the curse of dimensionality like Stochastic Collocation (SC) methods [67, 69], they are still a reasonable choice. The aforementioned Stochastic Collocation is also a non-intrusive method that exploits the regularity of the solution; thus, it is more efficient provided that the stochastic dimension is low. Note that, in the MC method, the number of samples taken does not depend on the stochastic dimension, unlike the SC method, which is a major plus. In all of those, we choose to use the MC method in this study. There are also superior variants in terms of computational cost, such as Multi-Level Monte Carlo (MLMC) methods [9, 31, 68] or Quasi-Monte Carlo methods [55], yet they are left as possible improvements to be done in future work.

On the other hand, on the physical space, we use the standard continuous finite element method (FEM). The FEM, advantageous with its high accuracy, complex-mesh handling abilities [60] is favorable to alternatives as Finite Difference Methods or Finite Volume Methods. In the literature, there are a lot of studies where the physical domain of the stochastic PDEs is discretized by the FEM. It is combined with the Monte Carlo method in [7, 21, 53] and with the Stochastic Collocation method in [6], whereas it is used together with the stochastic Galerkin method in [8, 25]. In [45] and [38] we see an overview of FEM methods for solution of PDEs with random coeffi-

cients. In terms of PDE-constrained setting we refer to [4, 17, 43, 53] and references therein.

In terms of optimization, PDE-constrained optimization problems with uncertainties are solved by using deterministic optimization methods in combination with a sampling or discretization scheme for the stochastic space; see, e.g., [13, 26, 40, 69]. These approaches, such as; full gradient methods, sequential quadratic programming (SQP) methods, trust-region methods, or quasi-Newton methods have limitations when the stochastic dimension increases. For example, a gradient descent algorithm that requires first-order gradient information computed everywhere in the physical space for each random variable is a costly operation. On the other hand, recently, PDE-constrained optimization problems have been taken into account in the context of stochastic approximation methods, whose origin is back to the 1950s [39, 62]. Although stochastic gradient descent (SGD) methods are widely used in machine learning problems [14], there is limited work for the risk-averse PDE-constrained optimization problems. A comparison of the stochastic approximation approach with the sample average approximation method was considered in [34]. The projected stochastic gradient descent method was applied for control constraint optimization problems in [28, 29]. In [53], the SGD approach was used with the Monte Carlo method in the setting of robust optimal control problem governed by an elliptic PDE with uncertain coefficients and then was combined with multilevel Monte Carlo method in [54]. Further, the SGD was applied to solve semi-linear elliptic equations by reformulating the problem as a functional minimization problem in [72]. Due to the noisy nature of stochastic gradient iteration, naive use of the algorithm in many instances suffers from the complex tuning of parameters and prolonged convergence rate [58]. To fasten the convergence of the SGD we can use momentum methods originally introduced in [61]. In these methods, we essentially add a "momentum" term to reduce the effect of the noise present in SGD. In deterministic setting, the literature is clear that the momentum methods may still suffer from noise accumulation; they may stall in some neighborhood of the optimum, have sub-optimal convergence rates, or may even not converge at all [22, 24, 30]. However, when they are used with SGD we see that these drawbacks are not present; further, they are more robust than SGD at the same solution accuracy [20, 22, 24, 64]. Moreover, SGD with Polyak's

momentum is shown to have an accelerated linear rate in a more recent work [50]. Although momentum methods such as Polyak's momentum (or heavy ball momentum) [61] and Nesterov's momentum [59] are popular methods to accelerate the convergence in the deterministic or stochastic optimization settings, they are not applied to PDE–constrained optimization problems with uncertainty according to the best of our knowledge.

We begin, in the next chapter, by giving the preliminary knowledge and laying mathematical foundations. Then, in Chapter 3, we introduce the robust deterministic optimal control problem constrained by a convection-diffusion equation containing uncertain coefficients, and discuss existence and uniqueness results together with the optimality conditions. In Chapter 4, we discuss the numerical methods used in details. Section 4.1 is devoted to the semi-discretization of the optimization problem in probability space by applying a Monte Carlo type approximation. Then we move on to the approximation in physical space by FEM in Section 4.2, which brings us to the fully discretized problem. The fully discrete optimal control problem is approximated by using the stochastic momentum methods, i.e., Polyak's and Nesterov's momentum methods, in Section 4.3. We conclude that chapter by introducing optimization method we used and by giving convergence results. In Chapter 5, the numerical results are presented to illustrate the efficiency of the proposed methodology. Conclusions and discussions are provided in the last chapter.

CHAPTER 2

PRELIMINARIES

In this thesis, we try to control the behavior of a system that is represented by a PDE with uncertain input data. To construct the mathematical representation of this system as such, we need to study the relevant fundamental theories. In this chapter, for finding the optimal values of a cost functional constrained by a PDE, we first focus on studying the function spaces since we leave the space of sufficiently continuous functions behind by discretizing the PDE. To account for the uncertainty in our PDE, we will also state some basics from the probability theory and the necessary statistics. Moreover, the convergence of the MC method is dependent on the "Strong Law of Large Numbers", so we introduce it as well. Finally, the basics of Karhunen-Loève expansion will be given to represent stochastic processes as an expansion of the orthogonal functions.

2.1 Function Spaces

Here, we present the essential function spaces to write and solve partial differential equations containing uncertain terms. First, we introduce Banach spaces, and then move to Hilbert spaces. After that, Sobolev spaces will be introduced, finally enabling us to define our model problem.

2.1.1 Banach Spaces

To define Banach spaces, we need to review the notion of norm, convergence, and completeness. Let us note that we restrict ourselves to real fields.

Definition 2.1.1. (norm) A mapping $\|\cdot\| : X \mapsto \mathbb{R}$ for some real vector space X is called a norm if

- $\|u\| \geq 0$ and $\|u\| = 0$ if and only if $u = 0$.
- $\|\lambda u\| = |\lambda|\|u\|$ for all $u \in X$ and $\lambda \in \mathbb{R}$.
- $\|u + v\| \leq \|u\| + \|v\|$ for all $u, v \in X$ (triangle inequality).

A vector space X equipped with a norm is called a *normed vector space* and denoted by $(X, \|\cdot\|)$.

Definition 2.1.2. (uniform convergence) We say a sequence of functions $u_n \in C(\bar{D})$ converges uniformly to a limit u if $\lim_{n \mapsto \infty} \|u_n - u\| \mapsto 0$.

With the uniform convergence (or simply, convergence), we can now define what a Cauchy sequence is.

Definition 2.1.3. (Cauchy sequence, complete) Let $(X, \|\cdot\|)$ be a normed vector space. A sequence $u_n \in X$ satisfying

$$\|u_n - u_m\| < \epsilon, \qquad \forall n, m \geq N,$$

for all $\epsilon > 0$, where $n, m, N \in \mathbb{N}$ is called a *Cauchy sequence*. A normed vector space $(X, \|\cdot\|)$ is called *complete* if every Cauchy sequence in X converges to a limit $u \in X$.

After the discussion of completeness, we can state the definition of Banach spaces.

Definition 2.1.4. (Banach space) A *Banach space* is a complete normed vector space.

2.1.2 Lebesgue Integrals and Measurability

In our work, we use the Lebesgue integral notion, which is a generalization of the Riemann integral. Unlike the Riemann integral, which is defined by a limit on the

sum of piecewise constant functions, the Lebesgue integral is defined by a limit on the sum of constants on measurable sets. Lebesgue integral is necessary for the definition of L^p spaces and more.

Definition 2.1.5. ($\sigma-$algebra) A set $\mathcal{F}$ of subsets of a set X is a $\sigma-$algebra if

- The empty set $\{\} \in \mathcal{F}$.
- The complement $F^c := \{x \in X : x \notin F\} \in \mathcal{F}$ for all $F \in \mathcal{F}$.
- The union of $\bigcup F_j \in \mathcal{F}$ for $F_j \in \mathcal{F}$, $j \in \mathbb{N}$.

In plain language, we can say that a $\sigma-$algebra is a collection of subsets that is closed under a finite number of unions and complement operations while containing the empty set. The pair $(X, \mathcal{F})$ is known as a *measurable space*.

As it is needed in theoretical discussions later, we introduce the Borel $\sigma-$algebra next.

Definition 2.1.6. (Borel $\sigma-$algebra) For a topological space Y, the Borel $\sigma-$algebra $\mathcal{B}(Y)$ is the smallest $\sigma-$algebra containing all open subsets of Y.

In this study, the integral operator will be defined for the measurable functions with respect to the measure space, which will be introduced next.

Definition 2.1.7. (measurable) For a function $u : X \mapsto \mathbb{R}$, if $\{x \in X : u(x) \leq a\} \in \mathcal{F},\ \forall a \in \mathbb{R}$, we say it is $\mathcal{F}-$*measurable*. Equivalently, a function $u : X \mapsto \mathbb{R}$ is $\mathcal{F}-$*measurable* if the pullback set is in the $\mathcal{F}$, i.e., $u^{-1}(G) \in \mathcal{F},\ \forall G \in \mathcal{B}(Y)$.

Definition 2.1.8. (measure, measure space) Let $(X, \mathcal{F})$ be a measurable space. A *measure* μ is defined as a mapping $\mu : \mathcal{F} \mapsto \mathbb{R}^+ \cup \{\infty\}$ such that

- the measure of empty set is zero, i.e., $\mu(\{\}) = 0$,
- $\mu(\cup_{n\in\mathbb{N}} F_n) = \sum_{n\in\mathbb{N}} \mu(F_n)$ provided that $F_n \in \mathcal{F}$ are disjoint.

Together, $(X, \mathcal{F}, \mu)$ is known as a *measure space*.

Note that a measurable function u is equal to 0 *almost surely (a.s.)* if $\mu(\{x \in X : u(x) \neq 0\}) = 0$.

Definition 2.1.9. (integral) Let $(D, \mathcal{F}, \mu)$ be a measure space and $(Y, \|\cdot\|_Y)$ be a Banach space. A function $s : D \mapsto Y$ is called a *simple* function if $\exists\, s_j \in Y$, and $F_j \in \mathcal{F},\ j = 1, \ldots, N$ such that $\mu(F_j) < \infty$, and

$$s(x) = \sum_{j=1}^{N} s_j \mathbb{1}_{F_j}(x), \qquad x \in D,$$

where $\mathbb{1}_{F_j}$ is the indicator function. Then, the integral of a simple function s with respect to a measure space $(D, \mathcal{F}, \mu)$ is

$$\int_D s(x) d\mu(x) := \sum_{j=1}^{N} s_j \mu(F_j). \tag{2.1.1}$$

A function u is called *integrable with respect to* μ if there exists a sequence of simple functions $u_n(x) \mapsto u(x)$ as $n \mapsto \infty$ for almost all $x \in D$, and u_n is a Cauchy sequence for all $\epsilon > 0$, and sufficiently large n, m,

$$\int_D \|u_n - u_m\|_Y d\mu(x) < \epsilon.$$

Here, the key observation is that the integrand is a simple function. If u is integrable, we set

$$\int_D u(x) d\mu(x) := \lim_{n \mapsto \infty} \int_D u_n(x) d\mu(x).$$

If $F \in \mathcal{F}$, we have

$$\int_F u(x) d\mu(x) := \int_D u(x) \mathbb{1}_F(x) d\mu(x).$$

The Lebesgue integral, or integral for short, is a linear functional, and also satisfies

$$\left\| \int_D u(x) d\mu(x) \right\|_Y \leq \int_D \|u(x)\|_Y d\mu(x).$$

When $Y = \mathbb{R}^d$ and $D \subset \mathbb{R}^d$ the *Lebesgue integral* with respect to $(D, \mathcal{F}, Leb)$, where $\mathcal{F}$ is the σ-algebra defined on D, and Leb is the Lebesgue measure corresponding to the usual notion of volume in $\mathbb{R}^d$, is denoted by

$$\int_D u(x) dLeb(x) = \int_D u(x) d(x).$$

The reason we need to generalize from Riemann integral is to be able to define integral for the probability spaces.

Last, we will state the well-known Fubini's theorem, which is used in the rest of the thesis.

Theorem 2.1.1. ([51], Theorem 1.24) *Suppose* $(\Omega_1, \mathcal{F}_1, \mu_1)$ *and* $(\Omega_2, \mathcal{F}_2, \mu_2)$ *be* σ*-finite measure spaces. Let* u *be a measurable function such that* $u : \Omega_1 \times \Omega_2 \mapsto Y$. *If*

$$\int_{\Omega_2} \left(\int_{\Omega_1} \|u(x_1, x_2)\|_Y d\mu_1(x_1) \right) d\mu_2(x_2) < \infty, \tag{2.1.2}$$

then we say that u *is integrable with respect to the product of measure* $\mu_1 \times \mu_2$*, and we have*

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} u(x_1, x_2) d(\mu_1 \times \mu_2)(x_1, x_2) &= \int_{\Omega_2} \left(\int_{\Omega_1} u(x_1, x_2) d\mu_1(x_1) \right) d\mu_2(x_2) \\ &= \int_{\Omega_1} \left(\int_{\Omega_2} u(x_1, x_2) d\mu_2(x_2) \right) d\mu_1(x_1). \end{aligned}$$

2.1.3 L^p, Hilbert, Sobolev Spaces

After the discussion of the Lebesgue integral, we can introduce L^p spaces now. They are spaces of functions with finite integrals.

Definition 2.1.10. (L^p spaces) Let $(Y, \|\cdot\|_Y)$ be a Banach space, and let p be such that $1 \leq p < \infty$. For a domain D, the set of Borel measurable functions $u : D \mapsto \mathbb{R}$ is denoted by $L^p(D)$, and satisfies

$$\|u\|_{L^p(D)} := \left(\int_D |u(x)|^p dx \right)^{1/p}. \tag{2.1.3}$$

For a measure space $(\Omega, \mathcal{F}, P)$, the set of $\mathcal{F}$-measurable functions $u : \Omega \mapsto Y$ is denoted by $L^p(\Omega, Y)$, and we have

$$\|u\|_{L^p(\Omega,Y)} := \left(\int_\Omega \|u(x)\|_Y^p d\mathbb{P}(x) \right)^{1/p}.$$

$L^\infty(\Omega, Y)$ is the set of $\mathcal{F}$-measurable functions $u : \Omega \mapsto Y$ satisfying

$$\|u\|_{L^\infty(\Omega,Y)} := \operatorname*{ess\,sup}_{x \in \Omega} \|u(x)\|_Y \leq \infty.$$

Let us introduce the Hilbert spaces starting from the inner product. Here, we also confine ourselves to real numbers only.

Definition 2.1.11. (inner product) A function $(\cdot,\cdot) : X \times X$ on a vector space X is called an *inner product*, if it satisfies the following conditions:

- positive definiteness: $(u,u) \geq 0$ and $(u,u) = 0$ if and only if $u = 0$, $\forall u \in X$.
- symmetry: $(u,v) = (v,u), \qquad \forall u, v \in X$.
- linearity in the first argument: $(\alpha u + \beta v, w) = \alpha(u,w) + \beta(v,w), \qquad \forall u, v \in X, \quad \forall \alpha, \beta \in \mathbb{R}$.

Definition 2.1.12. (Hilbert space) For a vector space H with the inner product $(\cdot,\cdot)$, if H is complete with respect to the induced norm $\|u\| =: (u,u)^{1/2}$, then it is called a *Hilbert space*. Any Hilbert space is also a Banach space.

A crucial thing to notice is, for a function to belong in an L^p space, it does not need to be continuous. To describe their regularity, first, we need to introduce a weaker type of derivative that does not require continuity. By the degree of the "weak derivative" that they can hold, we classify these functions, and those classes are called Sobolev spaces.

Definition 2.1.13. (weak derivative) For a Banach space Y, α-th *weak derivative* of a measurable function $u : D \mapsto Y$ is defined as a measurable function $\mathcal{D}^\alpha u : D \mapsto \mathbb{R}$ if

$$\int_D \mathcal{D}^\alpha u(x)\phi(x)dx = (-1)^{|\alpha|} \int_D u(x)\mathcal{D}^\alpha \phi(x)dx, \qquad \forall \phi \in C_c^\infty(D),$$

where $C_c^\infty(D)$ is the space of infinitely many times differentiable functions with compact support on a subset of D.

Definition 2.1.14. (Sobolev spaces) For a Banach space Y, a domain D, and $p \geq 1$, $W^{r,p}(D,Y)$ is called the Sobolev space, that is, the space of functions whose weak-derivatives up to order $r \in \mathbb{N}$ are members of $L^p(D,Y)$, i.e.,

$$W^{r,p}(D,Y) := \{u \in L^p(D,Y) : \mathcal{D}^\alpha u \in L^p(D,Y) \text{ if } |\alpha| \leq r\}.$$

Moreover, Sobolev spaces are endowed with the following norms

$$\|u\|_{W^{r,p}(D,Y)} := \left(\sum_{0\leq|\alpha|\leq r} \|\mathcal{D}^\alpha u\|^p_{L^p(D,Y)} \right)^{1/p}, \qquad 1 \leq p < \infty,$$

$$\|u\|_{W^{r,p}(D,Y)} := \max_{0\leq|\alpha|\leq r} \|\mathcal{D}^\alpha u\|_{L^\infty(D,Y)}, \qquad p = \infty.$$

Let H' be a Hilbert space. When $p = 2$, $W^{r,2}(D, H')$ is denoted as $H^r(D, H')$ and it is a Hilbert space as well.

In general, any $W^{r,p}(D, Y)$ with $p \neq 2$ is a Banach space with the corresponding norm, whereas $H^r(D, H')$ is a Hilbert space with the inner product

$$(u, v)_{H^r(D,H')} := \sum_{0\leq|\alpha|\leq r} (\mathcal{D}^\alpha u, \mathcal{D}^\alpha v)_{L^2(D,H')}.$$

When $H' = \mathbb{R}$, we abbreviate $H^r(D, \mathbb{R})$ as $H^r(D)$ and it is equipped with the following norm

$$\|u\|_{H^r(D)} := \left(\sum_{0\leq|\alpha|\leq r} \|\mathcal{D}^\alpha u\|^2_{L^2(D)} \right)^{1/2}.$$

Moreover, the semi-norm on the Hilbert space $H^r(D)$ is given by

$$|u|_{H^r(D)} := \left(\sum_{|\alpha|=r} \|\mathcal{D}^\alpha u\|^2_{L^2(D)} \right)^{1/2}.$$

Further, $H^1_0(D)$ is a Hilbert space vanishing on the boundary with the $H^1(D)$ inner product and is expressed mathematically as follows

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

2.2 Basic Concepts of Probability Theory

The uncertainty quantification of a system such as a physical system modeled as a partial differential equation (PDE) with random data has one of its main roots in the probability theory. The solution is represented by a random field, which is a generalization of random variables. In this section, we discuss the basics of probability theory briefly.

We begin with probability measure and probability space to study random variables.

Definition 2.2.1. (probability measure) A measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a probability measure if it satisfies $\mathbb{P}(\Omega) = 1$. We say an event F happens *almost surely* $(a.s.)$ if $\mathbb{P}(F) = 1$.

Definition 2.2.2. (probability space) A probability space $(\Omega, \mathcal{F}, P)$ is constituted of a sample space Ω, a $\sigma-$algebra $\mathcal{F}$ of Ω whose elements F are called $events$, and a probability measure $\mathbb{P}$. For a probability space, $\mathbb{P}$ assigns a probability between 0 and 1 to every measurable set of events in $\mathcal{F}$.

With a $(\Omega, \mathcal{F}, P)$ at hand, we can say that a *random variable* is a function assigning values to outcomes of an experiment modeled as $\omega \in \Omega$. More formally, it is defined as follows.

Definition 2.2.3. (random variables, realization) Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\Psi, \mathcal{G})$ be a measurable space. Then, a function X is called a $\Psi-$*valued random variable* if X is measurable from $(\Omega, \mathcal{F}, P)$ to $(\Psi, \mathcal{G})$. An outcome of an event, $X(\omega)$ for some $\omega \in \Omega$ is called a *realisation* and belongs to Ψ.

Every random variable X has an associated probability distribution $\mathbb{P}_X$ which is an integral concept to understand and work on the probability theory. Hence, let us move on to distributions and statistics.

2.2.1 Distributions and Statistics

Solutions of PDEs with uncertain coefficients are represented as random variables. It is only possible to obtain certain properties of these random variables, such as their expectation value, their variance, or, more generally, moments of these random variables. The reason is that we do not have access to the underlying probability distribution. We only have samples of solutions from which only moments of a random variable are attainable.

Definition 2.2.4. (probability distribution) Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\Psi, \mathcal{G})$ be a measurable space. Then, the *probability distribution* of a random variable

X is the probability measure on $(\Psi, \mathcal{G})$, and defined as $\mathbb{P}_X := \mathbb{P}(X^{-1}(G))$, where $X^{-1}(G) := \{\omega \in \Omega : X(\omega) \in G\}$ for $G \in \mathcal{G}$ is called a *pullback set*.

The definition of a probability distribution shows that $\mathbb{P}_X(\Psi) = \mathbb{P}_X(X^{-1}(\Psi)) = \mathbb{P}(\Omega) = 1$, i.e., $\mathbb{P}_X$ is also a probability measure, and $(\Psi, \mathcal{G}, \mathbb{P}_X)$ is a probability space.

Since we work with the statistical quantities of random variables in this thesis, we next give proper definitions of them. Statistical quantities, such as the mean, variance, standard deviation, and higher moments, are described by expectations of the random variables.

Definition 2.2.5. (expectation) For a Banach space-valued random variable X defined on $(\Omega, \mathcal{F}, P)$, the expectation of X is defined as

$$\mathbb{E}[X] := \int_{\Omega} X(\omega) d\mathbb{P}(\omega),$$

provided that X is integrable. For an integer-valued random variable X defined on $(\Omega, \mathcal{F}, P)$, the expectation of X corresponds to

$$\mathbb{E}[X] := \sum_{j \in \mathbb{Z}} j\mathbb{P}(x = j).$$

The expectation of X is also called the *mean* of X.

Definition 2.2.6. (moments) The expectation given as $\mathbb{E}[X^k]$ is called the kth order *moment* of a random variable X. The expression $\mathbb{E}[(X - \mu)^k]$, where μ is the mean of X, is called a *centered moment*. The second order centered moment $\mathbb{V}(X) := \mathbb{E}[(X - \mu)^2]$ is called the *variance* of X. It is the quantification of the spread of X around the mean. Moreover, the quantity $\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}$ is called the *standard deviation*.

At this point, we need to address that the integral defined in the definition (2.2.5) is not easily computable. We need to apply a change of variables. First, we go from an integral Ω to D, then use a so-called probability density function (pdf). Next, we state it formally.

Definition 2.2.7. (pdf) For a probability measure $\mathbb{P}$ on $(D, \mathcal{F})$ where $D \subset \mathbb{R}$, if there exists a function $p : D \mapsto [0, \infty)$ such that $\mathbb{P}(B) = \int_B p(x)dx$ for any $B \in \mathcal{B}(D)$, we say that p is the *probability density function* of probability measure $\mathbb{P}$.

We use a pdf as

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega) = \int_{D} x d\mathbb{P}_X(x) = \int_{D} x p(x) dx.$$

It is evident that if the pdf is known, it is straightforward to calculate the probability of some event. For example,

$$\mathbb{P}(X \in (a,b)) = \mathbb{P}(\{\omega \in \Omega : a < X(\omega) < b\}) = \mathbb{P}_X((a,b)) = \int_a^b p(x) dx.$$

2.2.2 Correlation and Independence

Now that we have the essentials about a random variable, we need to be able to relate the two of them together. We introduce the concept of covariance for that reason.

Definition 2.2.8. (covariance) Let X, Y be two real-valued random variables on $(\Omega, \mathcal{F}, P)$, and X, Y are measurable from $(\Omega, \mathcal{F})$ to $(D_X, \mathcal{B}(D_X)), (D_Y, \mathcal{B}(D_Y))$ for $D_X, D_Y \subset \mathbb{R}$, respectively. Then, the *covariance* of X, Y is

$$\mathcal{C}(X,Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X \mu_Y,$$

where μ_X, μ_Y are the means of X and Y, respectively. Note that $\mathcal{C}(X,X) = \mathbb{V}(X)$, and it is called the *variance* of X.

The expectation $\mathbb{E}[XY]$ in the covariance is calculated as

$$\mathbb{E}[XY] = \int_{\Omega} X(\omega) Y(\omega) d\mathbb{P}(\omega) = \int_{D_X \times D_Y} xy \; d\mathbb{P}_{X,Y}(x,y),$$

where $\mathbb{P}_{X,Y}(x,y)$ is the joint probability distribution of X, and Y. Often we scale the covariance and call it the *correlation coefficient* $\rho(X,Y) := \mathcal{C}(X,Y)/\mathbb{S}_X \mathbb{S}_Y$.

Definition 2.2.9. (joint probability distribution) The *joint probability distribution* $\mathbb{P}_{X,Y}$ of X and Y can be said as the probability distribution of the bivariate random variable $\mathbf{X} = [X,Y]^\top$ such that $\mathbb{P}_{X,Y}(B) := \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \in B\})$, where $B \in \mathcal{B}(D_X \times D_Y)$. Also, the *joint pdf* $p_{X,Y}$ of $\mathbb{P}_{X,Y}$ is, provided it exists,

$$\mathbb{P}_{X,Y}(B) = \int_B p_{X,Y}(x,y) \; dxdy.$$

We say that two random variables X and Y are *uncorrelated* if their covariance $\mathcal{C}(X,Y) = 0$. One should note that it is different from being *independent*, which we define next.

Definition 2.2.10. (independence of random variables) Two random variables X, Y on $(\Omega, \mathcal{F}, P)$ are called *independent* if the $\sigma-$algebras $\sigma(X), \sigma(Y)$ generated by them are independent.

If X and Y are two independent random variables and their expectations are finite, then they are uncorrelated. The converse does not hold in general. For more details, we refer to [51].

Also, the real-valued random variables X and Y are independent if and only if their *joint pdf* $p_{X,Y}$ satisfies

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \qquad x, y \in \mathbb{R}.$$

Towards the solutions of partial differential equations containing uncertain terms, which are random fields, we discuss what a multivariate random variable is.

Definition 2.2.11. (multivariate random variables) The concatenation of d random variables $\mathbf{X} = [X_1, \ldots, X_d]^\top$ from $(\Omega, \mathcal{F}, P)$ to $(D, \mathcal{B}(D))$ where $D \subset \mathbb{R}^d,\ d > 1$ are called *multivariate random variables*.

The mean vector $\mu \in \mathbb{R}^d$ consists of $\mu_j = \mathbb{E}[X_j]$, i.e.,

$$\mu = \mathbb{E}[\mathbf{X}] = \int_\Omega \mathbf{X}(\omega) d\mathbb{P}(\omega) = \big[\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]\big]^\top.$$

The relation of two multivariate random variables $\mathbf{X}, \mathbf{Y}$ are described by a covariance matrix $\mathcal{C}(\mathbf{X}, \mathbf{Y})$ such that

$$\mathcal{C}(\mathbf{X}, \mathbf{Y}) := \mathbb{E}\big[(\mathbf{X} - \mu_X)(\mathbf{Y} - \mu_Y)^\top\big],$$

whose $(i, j)-$th entry is $c_{ij} = \mathcal{C}(X_i, Y_j)$. We say that $\mathbf{X}, \mathbf{Y}$ are uncorrelated if the covariance matrix is a $d \times d$ zero matrix.

2.2.3 Hilbert Space-Valued Random Variables

For now, we have been dealing with real-valued random variables. It is often the case that we consider random variables taking values from a Hilbert space H and the adequate Borel $\sigma-$algebra $\mathcal{B}(H)$. Families of this kind of random variables form Banach and Hilbert spaces, and their expectations and covariance are worth mentioning.

Definition 2.2.12. ($L^p(\Omega, H)$ spaces) Let $(\Omega, \mathcal{F}, P)$ be a probability space and H be a Hilbert space with norm $\|\cdot\|$. Then, for $1 \leq p < \infty$, $L^p(\Omega, H)$ is the space of $H-$valued $\mathcal{F}-$measurable random variables $X : \Omega \mapsto H$ and is a Banach space with the norm

$$\|X\|_{L^p(\Omega,H)} := \Big(\int_\Omega \|X(\omega)\|^p d\mathbb{P}(\omega)\Big)^{1/p} = \mathbb{E}\big[\|X\|^p\big]^{1/p},$$

and for $p = \infty$, the Banach space $L^\infty(\Omega, H)$ is

$$\|X\|_{L^\infty(\Omega,H)} := \operatorname*{ess\,sup}_{\omega \in \Omega} \|X(\omega)\| < \infty.$$

$L^2(\Omega, H)$, the most used one, is the space of *H-valued mean-square integrable random variables*, and has the following inner product

$$(X, Y)_{L^2(\Omega,H)} := \int_\Omega (X(\omega), Y(\omega)) d\mathbb{P}(\omega) = \mathbb{E}\big[(X, Y)\big].$$

Definition 2.2.13. (covariance operator) Let H be a Hilbert space. A linear operator $\mathcal{C} : H \mapsto H$ is the *covariance* of $H-$valued random variables X and Y if

$$(\mathcal{C}\phi, \psi) = \mathcal{C}((X, \phi)(Y, \psi)), \qquad \forall \phi, \psi \in H.$$

If $\mathcal{C}$ is a zero operator, then we say that X and Y are *uncorrelated*.

2.2.4 Conditional Expectation

In a problem, often we have some partial knowledge about the solution and want to find its expectation value, or we may be using a stochastic solution method that employs some set of *independent and identically distributed (i.i.d.)* random variables and the correct way to look for a solution is to find the expectation value solely on that set of *i.i.d.* random variables. These frequent cases require *conditional expectation*, and here we give the fundamentals.

Definition 2.2.14. (conditional expectation) Let $X \in L^2(\Omega, \mathcal{F}, H)(:= L^2(\Omega, H))$. If $\mathcal{A}$ is a sub $\sigma-$algebra of $\mathcal{F}$, the *conditional expectation* $\mathbb{E}[X|\mathcal{A}]$ is the orthogonal projection from $L^2(\Omega, \mathcal{F}, H)$ to $L^2(\Omega, \mathcal{A}, H)$, i.e., $\mathbb{E}[X|\mathcal{A}] = PX$, where P is the corresponding orthogonal projector.

Note that $\mathbb{E}[X|\mathcal{A}]$ is a $\mathcal{A}-$measurable random variable. The conditional expectation given by a random variable Y is calculated by letting $\mathcal{A} = \sigma(Y)$, where $\sigma(Y)$ is the σ-algebra generated by Y. Notice, then, $\mathbb{E}[X|\mathcal{A}]$ is a function of Y by Doob-Dynkin Theorem [51].

We generalize the definition of conditional expectation from square-integrable random variables by the following theorem.

Theorem 2.2.1. ([51],Theorem 4.52) *Let H be a separable Hilbert space and X be a $H-$valued random variable on the probability space $(\Omega, \mathcal{F}, P)$ with $\mathbb{E}[\|X\|] < \infty$. For any sub $\sigma-$algebra $\mathcal{A}$ of $\mathcal{F}$, there exists an $\mathcal{A}-$measurable random variable $\mathbb{E}[X|\mathcal{A}]$, unique almost surely, such that $\mathbb{E}\Big[\|\mathbb{E}[X|\mathcal{A}]\|\Big] < \infty$ and*

$$\int_G \mathbb{E}[X|\mathcal{A}](\omega)d\mathbb{P}(\omega) = \int_G X(\omega)d\mathbb{P}(\omega), \qquad \forall G \in \mathcal{A}.$$

Furthermore, the following properties hold:

- ***Linearity:*** *If X, Y are $\mathcal{F}-$measurable, then for $a, b \in \mathbb{R}$*

$$\mathbb{E}[aX + bY|\mathcal{A}] = a\,\mathbb{E}[X|\mathcal{A}] + b\,\mathbb{E}[Y|\mathcal{A}] \quad a.s.$$

- ***Independence:*** *$\mathbb{E}[X|\mathcal{A}] = \mathbb{E}[X]$ $a.s.$ if $\mathcal{A}$ and $\sigma(X)$ are independent $\sigma-$algebras.*
- ***Taking out what is known:*** *If Y is $\mathcal{A}-$measurable, then $\mathbb{E}[Y|\mathcal{A}] = Y$ and $\mathbb{E}[XY|\mathcal{A}] = Y\mathbb{E}[X|\mathcal{A}] \quad a.s.$*
- ***Tower property:*** *If $\mathcal{A}_1$ is a sub $\sigma-$algebra of $\mathcal{A}_2$, $\mathbb{E}\Big[\mathbb{E}[X|\mathcal{A}_2]|\mathcal{A}_1\Big] = \mathbb{E}[X|\mathcal{A}_1]$ $a.s.$*

2.2.5 Convergence of Random Variables

So far, we have stated a significant portion of the basic concepts of probability, and we can now discuss the convergence of approximations to stochastic differential equations. The approximate solution X_n and the exact solution X are random variables. We need to define the limits of a sequence of random variables to investigate convergence. The ways to do this are dependent on the number of different types of approximations, there are mainly two. First, we approximate each realization $X(\omega)$ as a function of the sample $\omega \in \Omega$. Second, we approximate averages $\mathbb{E}[\phi(X)]$ for a

test function ϕ. Next, we make these ways precise using the concepts of convergence almost surely, convergence in probability, convergence in p-th mean, and convergence in distribution.

Definition 2.2.15. (convergence of random variables) Let H be a Hilbert space and X_n be a sequence of H–valued random variables. We say X_n converges to a random variable $X \in H$ *almost surely* if $X_n(\omega) \mapsto X(\omega)$ for almost all $\omega \in \Omega$, i.e.,

$$\mathbb{P}\big(\|X_n - X\| \mapsto 0 \quad \text{as} \quad n \mapsto \infty\big) = 1.$$

We say X_n converges to a random variable $X \in H$ *in probability* if $\mathbb{P}(\|X_n - X\| > \epsilon) \mapsto 0$ as $n \mapsto \infty$ for any $\epsilon > 0$. We say X_n converges to a random variable $X \in H$ *in p-th mean, or in $L^p(\Omega, H)$* if $\mathbb{E}[\|X_n - X\|^p] \mapsto 0$ as $n \mapsto \infty$. We say that $p = 2$ is convergence in mean square. We say X_n converges to a random variable $X \in H$ *in distribution* if $\mathbb{E}[\phi(X_n)] \mapsto \mathbb{E}[\phi(X)]$ as $n \mapsto \infty$ for any bounded continuous function $\phi : H \mapsto \mathbb{R}$. This type of convergence is also known as *weak convergence* or *convergence of laws*.

2.2.6 Strong Law of Large Numbers

One of the main type of solution methods, which we also use in this thesis, is the Monte Carlo methods, and the Strong Law of Large Numbers is fundamental to this type of methods [36].

Theorem 2.2.2. ([36],Theorem 20.1) *Let X_n be sequence of $i.i.d.$ random variables. Assume that $\mu = \mathbb{E}[X_j]$ and $\mathbb{V}[X_j] < \infty$ for all $j \in 1, \ldots, n$, and $S_n = \sum_{j=1}^{n} X_j$. Then,*

$$\lim_{n \mapsto \infty} \frac{S_n}{n} = \lim_{n \mapsto \infty} \frac{1}{n} \sum_{j=1}^{n} X_j = \mu \quad a.s. \text{ and in } L^2.$$

2.3 Karhunen-Loève Expansion

We need to represent stochastic processes in a finite setting since they are continuous and do not fit in computers' discrete nature. A representation of a stochastic process with a known covariance function can be done by the Karhunen-Loève (KL)

expansion. It is to express a stochastic process as an infinite linear combination of the eigenvalues and eigenfunctions of its covariance function with uncorrelated random variables as coefficients. In this sense, it is a Fourier-like expansion. To examine the KL expansion, let us introduce spectral decomposition of a real-valued symmetric matrix since it will eventually lead to a discrete version of the Karhunen-Loève expansion, and it is easier to grasp in this way.

Let $\{X(t) : t \in \mathcal{T}\}$ be a real-valued Gaussian process, and $\mathcal{T} \subset \mathbb{R}$. Let $\mu(t) := \mathbb{E}[X(t)]$, and $\mathcal{C}(s,t)$ be the mean and covariance functions, respectively. For $t_1, \ldots, t_N \in \Omega$, let us define

$$\mathbf{X} = [X(t_1), \ldots, X(t_N)]^\top \sim N(\mu, \mathcal{C}_\mathbf{N}), \tag{2.3.1}$$

where $\mu = [\mu(t_1), \ldots, \mu(t_N)]^\top$, and $c_{ij} = \mathcal{C}(t_i, t_j)$ are the entries of covariance matrix $\mathcal{C}_\mathbf{N}$. We know that the samples of $\mathbf{X}$ can be generated by

$$\mathbf{X} = \mu + \mathbf{V}^\top \xi, \tag{2.3.2}$$

where $\xi := [\xi_1, \ldots, \xi_N]^\top$ with $i.i.d.$ components $\xi_j \sim N(0,1)$ and $\mathbf{V}$ comes from the decomposition of $\mathcal{C}_N$

$$\mathcal{C}_N = \mathbf{V}^\top \mathbf{V}, \tag{2.3.3}$$

where V is found by the spectral decomposition.

Every $N \times N$ real-valued symmetric matrix, say $\mathbf{A}$, can be written as $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, where $\mathbf{U}$ is an orthonormal matrix whose columns $\mathbf{u}_j$ are eigenvectors of $\mathbf{A}$ and $\mathbf{\Sigma}$ is the diagonal matrix of eigenvalues λ_j of $\mathbf{A}$. The form $\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$ is called as *spectral decomposition* of $\mathbf{A}$. The covariance matrix $\mathcal{C}_\mathbf{N}$ has such a decomposition, and since it is positive semi-definite we can order its eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ so that

$$\mathcal{C}_\mathbf{N} = \mathbf{V}^\top \mathbf{V}, \qquad \text{where} \quad \mathbf{V}^\top := \mathbf{U}\mathbf{\Sigma}^{1/2}. \tag{2.3.4}$$

With this choice, we have

$$\mathbf{X} = \mu + \sum_{j=1}^{N} \sqrt{\lambda_j}\mathbf{u}_j \xi_j, \qquad \xi_j \sim N(0,1) \quad i.i.d., \tag{2.3.5}$$

and we can generate samples from $N(\mu, \mathcal{C}_\mathbf{N})$.

It is often the case that the ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ are decaying rapidly, and we can truncate the sum (2.3.5) at the nth term for $n \ll N$ to have significant savings in computational cost. Let $\mathcal{C}_\mathbf{n}$ be such an approximation denoted as

$$\mathcal{C}_\mathbf{n} = \mathbf{U_n}\mathbf{\Sigma_n}\mathbf{U_n}^\top = \sum_{j=1}^{n} \lambda_j \mathbf{u_j}\mathbf{u_j}^\top, \tag{2.3.6}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the largest n eigenvalues, and $\mathbf{u_1}, \ldots, \mathbf{u_n}$ are the corresponding eigenvectors. Such a truncation would change (2.3.5) as

$$\widehat{\mathbf{X}} = \mu + \sum_{j=1}^{n} \sqrt{\lambda_j}\mathbf{u}_j\xi_j, \qquad \xi_j \sim N(0,1) \quad i.i.d. \tag{2.3.7}$$

The error of approximating $\mathcal{C}_\mathbf{N}$ by $\mathcal{C}_\mathbf{n}$ in matrix L^2 norm is

$$\|\mathcal{C}_\mathbf{N} - \mathcal{C}_\mathbf{n}\| = \Big\| \sum_{j=n+1}^{N} \lambda_j \mathbf{u_j}\mathbf{u_j}^\top \Big\| = \lambda_{n+1}.$$

Then, the mean-square error of approximating $\mathbf{X}$ by $\widehat{\mathbf{X}}$ is equivalent to

$$\mathbb{E}\Big[\|\mathbf{X} - \widehat{\mathbf{X}}\|_2^2\Big] = \mathbb{E}\Big[\sum_{j=n+1}^{N}\sum_{k=n+1}^{N} \sqrt{\lambda_j}\sqrt{\lambda_k}\mathbf{u}_j^\top\mathbf{u}_k\xi_j\xi_k\Big] = \sum_{j=n+1}^{N} \lambda_j.$$

Finally, the error of approximating averages of functions becomes, for some $K > 0$,

$$|\mathbb{E}[\phi(\mathbf{X})] - \mathbb{E}[\phi(\widehat{\mathbf{X}})]| \leq K\|\phi\|_2 \sum_{j=n+1}^{N} \lambda_j^2, \qquad \forall \phi \in L^2(\mathbb{R}^N).$$

All of the errors are minimized by choosing the n largest eigenvalues of $\mathcal{C}_N$. This truncated spectral decomposition provides a form of model order reduction; see [51] and references therein for more details.

As we have stated previously, Karhunen-Loève expansion is the generalisation of spectral decomposition (2.3.5) to stochastic processes. For a stochastic process $X(t)$, let $\mu(t) = \mathbb{E}[X(t)]$. We are interested in sampling paths $X(t,\omega)$ in terms of an orthonormal basis $\{\phi_j : j \in \mathbb{N}\}$ of $L^2(\mathcal{T})$, that is,

$$X(t,\omega) = \mu(t) + \sum_{j=1}^{\infty} \xi_j(\omega)\phi_j(t), \tag{2.3.8}$$

where the coefficients $\xi_j(t)$ are random variables given by a standard computation taking advantage of orthonormality of ϕ_js as

$$\xi_j(t) := \big(X(t,\omega) - \mu(t), \phi_j(t)\big)_{L^2(\mathcal{T})}.$$

Let $\mathcal{C}(s,t)$ be the covariance function of $X(t)$. Eigenvalues and eigenfunctions of $\mathcal{C}(s,t)$ are found by the Fredholm integral equation of the second kind

$$\int_{\mathcal{T}} \mathcal{C}(s,t)\phi_i(t)dt = \lambda_i \phi_i(t),$$

where the λ_i are the eigenvalues corresponding to the eigenvectors $\phi_i(t)$. The method of finding the orthonormal basis $\{\phi_j : j \in \mathbb{N}\}$ of $L^2(\mathcal{T})$ is what is called the Karhunen-Loève expansion. By inspection, it is clear the resemblence between (2.3.5) and (2.3.8). As mentioned before, (2.3.5) is often called the discrete Karhunen-Loève expansion. Now, we state the convergence result of the KL expansion.

Theorem 2.3.1. ([51],Theorem 5.28) *Let $\{X(t) : t \in \mathcal{T}\}$ be a stochastic process and supposes that $X \in L^2(\Omega, L^2(\mathcal{T}))$. Then,*

$$X(t,\omega) = \mu(t) + \sum_{j=1}^{\infty} \sqrt{\lambda_j}\phi_j(t)\xi_j(\omega), \tag{2.3.9}$$

where the sum convergeces in $L^2(\Omega, L^2(\mathcal{T}))$,

$$\xi_j(\omega) := \frac{1}{\sqrt{\lambda_j}}\big(X(t,\omega) - \mu(t), \phi_j(t)\big)_{L^2(\mathcal{T})},$$

and $\{\lambda_j, \phi_j\}$ denotes the eigenvalues and eigenfunctions of the covariance operator $\mathcal{C}(s,t)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. The random coefficients ξ_j are mean zero, unit variance, and pairwise uncorrelated. If the process is Gaussian, then $\xi_j \sim N(0,1)$ $i.i.d.$

The truncated version of KL expansion is then given by

$$X_j(t,\omega) := \mu(t) + \sum_{j=1}^{J} \sqrt{\lambda_j}\phi_j(t)\xi_j(\omega). \tag{2.3.10}$$

Ideally we try to choose the largest J eigenvalues and the corresponding eigenfunctions. The truncated KL expansion provides a good approximant $X_J(t)$ to the stochastic process $X(t)$. In fact, the error between $X(t)$ and $X_J(t)$ is

$$\|X - X_J\|^2_{L^2(\Omega,L^2(\mathcal{T}))} = \mathbb{E}\Big[\sum_{j=J+1}^{\infty} \lambda_j \|\phi_j\|^2_{L^2(\mathcal{T})}\xi_j^2\Big] = \sum_{j=J+1}^{\infty} \lambda_j. \tag{2.3.11}$$

Now, one may see that the Karhunen-Loève expansion is obtained, provided that the eigenvalues and eigenfunctions are available. This is generally not the case, and they need to be approximated by numerical methods as collocation, quadrature methods, or Galerkin approximation [51]. We finally note that efficient generalization of samples is out of context in this thesis; see, e.g., [23, 32] for other sampling techniques.

2.4 Inequalities

The following are some of the most well-known inequalities used frequently in the proofs we give, so hereby we state them together. First of all, here we introduce the very well-known Cauchy-Schwarz inequality.

Lemma 2.4.1. ([51]),Cauchy-Schwarz inequality) *For a Hilbert space H, we have*

$$|(u,v)_H| \leq \|u\|_H \|v\|_H \qquad \forall u, v \in H. \tag{2.4.1}$$

Young's inequality (2.4.2) is an important tool showing convergence results of optimization algorithms later.

Lemma 2.4.2. ([37],Young's inequality) *For some nonnegative real numbers $a \geq 0$, $b \geq 0$, and for some real numbers $p > 1$, $q > 1$ such that $1/p + 1/q = 1$, we have*

$$a\, b \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{2.4.2}$$

The particular case of $p = q = 2$ is of most importance to us;

$$a\, b \leq \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2} \qquad \epsilon > 0. \tag{2.4.3}$$

Last, we provide the Poincaré's inequality.

Theorem 2.4.3. ([51], Poincaré's inequality) *Let D be a bounded domain. Then, there exists a constant $K_p > 0$ satisfying*

$$\|u\|_{L^2(D)} \leq K_p |u|_{H^1(D)}, \qquad \forall u \in H_0^1(D). \tag{2.4.4}$$

CHAPTER 3

MODEL PROBLEM

The optimal control problems constrained by deterministic PDEs is a very well established research area. However, the case of constraint equation being random PDEs is an area that is newly emerging with works such as [18, 33, 35, 63]. In this thesis, we focus on robust deterministic optimal control problems constrained by convection-diffusion equations with random coefficients. As we will introduce shortly, our cost functional is a tracking-type cost functional similar to ones used in [9, 63]. Unlike the most of the literature, it contains a standard deviation term added which also minimizes the dispersion of the solution to have a better control.

In this chapter, we first present the existence and uniqueness of a solution to the underlying constraint equation, and our optimal control problem, called as the robust deterministic optimal control problem. Then, the derivation of the optimality system for the model problem will be discussed.

3.1 Robust Deterministic Optimal Control Problem

Let $D \subset \mathbb{R}^2$ be a convex bounded polygonal spatial domain with a Lipschitz boundary ∂D, and the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ denoting a 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 accompanying probability measure. Let X be a generic random field on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ denoted by $X(x, \omega) : D \times \Omega \to \mathbb{R}$. For a fixed $x \in D$, $X(x, \cdot)$ is a

real–valued square integrable random variable $X(x,\cdot) \in L^2(\Omega,\mathcal{F},\mathbb{P})$, i.e.,

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

where $L^2(\Omega) := L^2(\Omega,\mathcal{F},\mathbb{P})$. For any random variable X defined on $(\Omega,\mathcal{F},\mathbb{P})$, the mean $\mathbb{E}[X]$, the standard deviation $\mathbb{S}(X)$, and the corresponding variance $\mathbb{V}(X)$ are given, respectively, by

$$\mathbb{E}[X] = \int_\Omega X \, d\mathbb{P}(\omega), \qquad \mathbb{S}(X) = \left[\int_\Omega (X - \mathbb{E}[X])^2 \, d\mathbb{P}(\omega)\right]^{1/2},$$
$$\mathbb{V}(X) = [\mathbb{S}(X)]^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 .$$

We next introduce the tensor-product space $H^k(D) \otimes L^2(\Omega)$, which is equipped with the norm

$$\|X\|_{H^k(D)\otimes L^2(\Omega)} := \left(\int_\Omega \|X(\cdot,\omega)\|^2_{H^k(D)} \, d\mathbb{P}(\omega)\right)^{1/2} < \infty. \tag{3.1.1}$$

Moreover, the following isomorphism relation holds

$$H^k(D) \otimes L^2(\Omega) \simeq L^2(H^k(D);\Omega) \simeq H^k(D; L^2(\Omega)).$$

The state space and control space are, then, defined as

$$\mathcal{Y} := H^1_0(D) \otimes L^2(\Omega), \quad \text{and} \quad \mathcal{U} := L^2(D),$$

respectively. Also, setting $\mathcal{X} := L^2(D) \otimes L^2(\Omega)$, and $\mathcal{W} := L^2(D)$, we have $\mathcal{U}, \mathcal{Y} \subset \mathcal{X}$.

In this thesis, we consider the following robust deterministic optimal control problem governed by a convection-diffusion equation with uncertain coefficients

$$\min_{u\in\mathcal{U}} \; \mathcal{J}(y,u) := \frac{1}{2}\|y - y^d\|^2_{\mathcal{X}} + \frac{\gamma}{2}\|\mathbb{S}(y)\|^2_{\mathcal{W}} + \frac{\mu}{2}\|u\|^2_{\mathcal{U}} \tag{3.1.2}$$

subject to

$$-\nabla\cdot\big(a(x,\omega)\nabla y(x,\omega)\big) + \mathbf{b}(x,\omega)\cdot\nabla y(x,\omega) = f(x) + u(x) \;\text{ in } D\times\Omega, \tag{3.1.3a}$$
$$y(x,\omega) = 0 \qquad\qquad \text{on } \partial D\times\Omega, \tag{3.1.3b}$$

where the function $\mathcal{J}(y,u)$ is a cost functional of tracking–type, including a risk penalization by the standard deviation. The first term in (3.1.2) measures the difference

between the state function y and the desired function y^d in the expectation of $y - y^d$. We assume that the state function $y \in \mathcal{Y}$ is a random field, whereas the desired state $y^d \in \mathcal{Y}$ is modeled as deterministic without any loss of generality. The second term is a measure of the standard deviation of y, which is added to obtain more information about the state variable. The last term corresponds to the deterministic distributive control. Therefore, the underlying optimal control problem (3.1.2)–(3.1.3) is called a robust deterministic optimal control problem; see, e.g., [2]. The constant $\mu > 0$ is a regularization parameter that acts as a penalty on the control, whereas $\gamma \geq 0$ is the risk-aversion parameter. Note that, although the objective functional $\mathcal{J}$ contains uncertain terms it is a deterministic quantity.

Further, the coefficients $a : (D \times \Omega) \to \mathbb{R}$ and $\mathbf{b} : (D \times \Omega) \to \mathbb{R}^2$ are random diffusivity and velocity coefficients, respectively, which are assumed to have bounded and continuous covariance functions. To show the regularity of the solution $y(x, \omega)$ in (3.1.3), we need to make the following assumptions on the random coefficients.

Assumption 3.1.1. *i) The diffusivity coefficient $a(x, \omega)$ is $\mathbb{P}$–almost surely uniformly positive, i.e., there exist constants $a_{\min}, a_{\max}$ such that $0 < a_{\min} \leq a_{\max} < \infty$ with*

$$a_{\min} \leq a(x, \omega) \leq a_{\max}, \qquad \textit{a.e. in } D \times \Omega. \tag{3.1.4}$$

ii) The velocity coefficient $\mathbf{b}$ satisfies $\mathbf{b} \in \left(L^\infty(\overline{D})\right)^2$, and $\nabla \cdot \mathbf{b}(x, \omega) = 0$.

3.1.1 Existence and Uniqueness

By following the standard arguments, the weak formulation of (3.1.2)–(3.1.3) is given by

$$\begin{aligned} \min_{u \in \mathcal{U}} \mathcal{J}(y, u) \;\; = \;\; & \frac{1}{2}\mathbb{E}\left[\int_D \left(y(u) - y^d\right)^2 dx\right] + \frac{\gamma}{2}\mathbb{E}\left[\int_D \left(y(u) - \mathbb{E}[y(u)]\right)^2 dx\right] \\ & + \frac{\mu}{2}\mathbb{E}\left[\int_D u^2 \, dx\right] \end{aligned} \tag{3.1.5}$$

subject to

$$a[y, v] - [u, v] = [f, v], \qquad v \in \mathcal{Y}, \tag{3.1.6}$$

where

$$a[y,v] = \mathbb{E}\left[\int_D \left(a(x,\omega)\nabla y\cdot\nabla v + \mathbf{b}(x,\omega)\cdot\nabla y\, v\right) dx\right], \quad \forall y,v\in\mathcal{Y}, \text{(3.1.7a)}$$

$$[u,v] = \mathbb{E}\left[\int_D uv\, dx\right], \qquad \forall u\in\mathcal{U}, v\in\mathcal{Y}. \text{(3.1.7b)}$$

Before stating the existence and uniqueness of a solution to (3.1.5)-(3.1.6), we need to discuss that a solution to the bilinear form (3.1.6) for the fixed control u exists and is unique. Next, we state the well-known Lax-Milgram Lemma, which will be needed to show the existence of a solution to the bilinear form (3.1.6).

Theorem 3.1.2. (Lax-Milgram Lemma, [70, Lemma 2.2]) *Let H be a Hilbert space and $a : H\times H\mapsto\mathbb{R}$ be a bilinear form. Suppose that there exists constants $c_{cnt}, c_{cr}\in\mathbb{R}$ such that the bilinear form a is continuous and coercive, respectively, for all $u,v\in H$, i.e.,*

$$|a(u,v)| \le c_{cnt}\|u\|_H\|v\|_H,$$
$$a(v,v) \ge c_{cr}\|v\|_H^2.$$

Then, for every linear form, say, $\ell(v)\in H^*$ the variational problem (3.1.6) admits a unique solution $y\in H$. In addition, there exists a constant $c_a>0$ that is independent of ℓ, the following is satisfied

$$\|y\|_H \le c_a\|\ell\|_{H^*}.$$

Theorem 3.1.3. *Suppose that the conditions in Assumption 3.1.1 hold, and $f\in\mathcal{W}$. Then, the weak formulation (3.1.6) for a fixed control u has a unique solution $y\in\mathcal{Y}$.*

Proof. We now apply Lax-Milgram Lemma introduced in Theorem 3.1.2 for the weak formulation (3.1.6).

By the Assumption 3.1.1, we have

$$\int_D a(x,\omega)\nabla v\cdot\nabla v\, dx \ge a_{min}\|\nabla v\|^2_{L^2(D)}. \tag{3.1.8}$$

Then, with the help of the Poincaré inequality given in (2.4.4), we get

$$\|v\|^2_{H^1(D)} = \|v\|^2_{L^2(D)} + \|\nabla v\|^2_{L^2(D)} \le (1+K_p^2)\|\nabla v\|_{L^2(D)}, \tag{3.1.9}$$

where K_p is the Poincaré constant.

Combining (3.1.8) and (3.1.9), we obtain

$$\int_D a(x,\omega)\nabla v \cdot \nabla v \, dx \geq a_{min}\|\nabla v\|^2_{L^2(D)} \geq \frac{a_{min}}{1+K_p^2}\|v\|^2_{H^1(D)}. \tag{3.1.10}$$

An application of the identity $\frac{\partial v}{\partial x}\cdot v = \frac{1}{2}\frac{\partial (v^2)}{\partial x}$, integration by parts, and the Assumption 3.1.1 on the second term of (3.1.7a), yields

$$\begin{aligned}\int_D vb \cdot \nabla v \, dx &= \frac{1}{2}\int_D b \cdot \nabla v^2 \, dx = -\frac{1}{2}\int_D v^2 \nabla \cdot b \, dx + \frac{1}{2}\int_{\partial D} b \cdot \hat{\mathbf{n}} v^2 \, ds \\ &= -\frac{1}{2}\int_D v^2 \nabla \cdot b \, dx = 0. \end{aligned} \tag{3.1.11}$$

Adding (3.1.10) and (3.1.11) together, and taking the expectation gives us,

$$\mathbb{E}[a(v,v)] = a[v,v] \geq \mathbb{E}\Big[\underbrace{\Big(\frac{a_{min}}{1+K_p^2}\Big)}_{c_{cr}}\|v\|^2_{H^1(D)}\Big] = \underbrace{\Big(\frac{a_{min}}{1+K_p^2}\Big)}_{c_{cr}}\|v\|^2_{\mathcal{Y}},$$

where c_{cr} denotes the coercivity constant.

By using the following estimates obtained from the Cauchy-Schwarz and the Assumption 3.1.1,

$$\begin{aligned}\Big|\int_D a(x,\omega)\nabla u \cdot \nabla v \, dx\Big| &\leq a_{max}\|\nabla u\|_{L^2(D)}\|\nabla v\|_{L^2(D)} \leq a_{max}\|u\|_{H^1(D)}\|v\|_{H^1(D)}, \\ \Big|\int_{\partial D} vb \cdot \nabla u \, dx\Big| &\leq \|b\|_{L^\infty(D)}\|v\|_{L^2(D)}\|\nabla u\|_{L^2(D)} \leq \|b\|_{L^\infty(D)}\|v\|_{H^1(D)}\|u\|_{H^1(D)},\end{aligned}$$

we obtain the continuity of the bilinear form $a[\cdot,\cdot]$

$$a[u,v] \leq \underbrace{\big(a_{max} + \mathbb{E}[\|b\|_{L^\infty(D)}]\big)}_{c_{cnt}}\|u\|_{\mathcal{Y}}\|v\|_{\mathcal{Y}}. \tag{3.1.12}$$

Finally, hereby we give the continuity of the linear form $[f+u,\cdot]$. Evoking the Cauchy-Schwarz inequality on (f,v) we get

$$(f+u,v) \leq \|f+u\|_{L^2(D)}\|v\|_{L^2(D)} \leq \|f+u\|_{L^2(D)}\|v\|_{H^1(D)}.$$

Taking the expectation, we show that the linear form is continuous

$$(f+u,v) \leq \|f+u\|_{\mathcal{W}}\|v\|_{\mathcal{Y}}.$$

Hence, by the Lax-Milgram Lemma, (3.1.6) has a unique solution. □

Next, we introduce the reduced form of (3.1.2), such that $\mathcal{J}(u) := \mathcal{J}(y(u), u)$,

$$\begin{aligned}\min_{u\in\mathcal{U}} \mathcal{J}(u) &= \frac{1}{2}\mathbb{E}\left[\int_D \left(y(u) - y^d\right)^2 dx\right] + \frac{\gamma}{2}\mathbb{E}\left[\int_D \left(y(u) - \mathbb{E}[y(u)]\right)^2 dx\right] \\ &\quad + \frac{\mu}{2}\mathbb{E}\left[\int_D u^2\, dx\right]. \end{aligned} \tag{3.1.13}$$

Now, we examine the existence and uniqueness of the reduced objective function (3.1.13).

Theorem 3.1.4. *For the given Hilbert spaces $\mathcal{Y}, \mathcal{U}$, the desired state $y^d \in \mathcal{Y}$, and constants $\gamma \geq 0, \mu > 0$, the quadratic optimization problem given in (3.1.13) admits a unique optimal control $\bar{u}$.*

Proof. The convexity of (3.1.13) is obvious since any $p-$norm is convex, and a positive-weighted combination of convex functions is also convex. Moreover, since $\mu > 0$, the cost functional (3.1.13) is strictly convex. Then, the existence and uniqueness follows by the standard arguments in the optimal control theory, for more details see, e.g., [49, Theorem 1.3] and [70, Theorem 2.14]. □

3.1.2 Optimality Conditions

Now, as the existence and uniqueness of a solution of (3.1.2)–(3.1.3) is guaranteed; we can proceed to necessary optimality conditions. The optimality system, also known as the Karush-Kuhn-Tucker (KKT) system, is crucial to our understanding of optimal control problems, and is the system we will solve to find the solution of (3.1.2)–(3.1.3). For its derivation, it is necessary for us to introduce a generalization of the notion of derivatives since the elements of the spaces $\mathcal{Y}, \mathcal{U}, \mathcal{X}, \mathcal{W}$ need not to be continuous.

Definition 3.1.1. (directional derivative) Let V, Z be real Banach spaces. Let $\mathcal{V}$ be a nonempty, open subset of V and $F : \mathcal{V} \mapsto Z$. If the limit

$$\delta F(v, h) = \lim_{t\mapsto 0^+} \frac{F(v + th) - F(v)}{t}, \tag{3.1.14}$$

exists in Z, then δF is called the directional derivative of F at $v \in \mathcal{V}$ along the direction $h \in \mathcal{V}$.

Further, for a Hilbert space $\{H, (\cdot,\cdot)_H\}$ and $F : H \mapsto \mathbb{R}$ we have that

$$\delta F(v,h) = (\nabla F(v), h)\,, \tag{3.1.15}$$

where $\nabla F(v)$ is called the gradient of F at $v \in H$.

Theorem 3.1.5. *Let the Assumptions 3.1.1 hold, and let $\gamma \geq 0$, $\mu > 0$. Then, by the Theorem 3.1.4, the OCP (3.1.13) subject to (3.1.3) admits a unique control $u^* \in \mathcal{U}$. The optimal control u^* and the corresponding optimal state $y(u^*)$ are given by the optimality system*

$$-\nabla(a(x,\omega)\nabla y(x,\omega)) + \boldsymbol{b}(x,\omega)\cdot\nabla y(x,\omega) \;=\; f+u, \tag{3.1.16a}$$

$$-\nabla(a(x,\omega)\nabla p(x,\omega)) - \boldsymbol{b}(x,\omega)\cdot\nabla p(x,\omega) \;=\; y - y_d + \gamma(y-\mathbb{E}[y]), \tag{3.1.16b}$$

$$\nabla\mathcal{J}(u) \;=\; \mu u + \mathbb{E}[p], \tag{3.1.16c}$$

where p is called the adjoint variable.

Proof. We write the Langrangian as follows,

$$\mathcal{L}(y,u,p) = \mathcal{J}(y,u) + (p,c)_{\mathcal{X}}\,,$$

where $(\cdot,\cdot)_{\mathcal{X}}$ denotes the inner product in $\mathcal{X}$, $c(y,u) = \nabla(a\nabla y) - \mathbf{b}\cdot\nabla y + f + u = 0$ is the constraint equation, and $p \in \mathcal{Y}$ is a Langrange multiplier. The partial derivatives of $\mathcal{L}$ with respect to p, y, and u give the optimality conditions:

$$\begin{cases} 0 = \nabla_p\mathcal{L} = c(y,u), \\ 0 = \nabla_y\mathcal{L} = \nabla_y\mathcal{J} + \nabla_y(c,p)_{\mathcal{X}}, \\ 0 = \nabla_u\mathcal{L} = \nabla_u\mathcal{J} + \nabla_u(c,p)_{\mathcal{X}}. \end{cases}$$

Starting with the gradient with respect to y, $\nabla_y\mathcal{L}$, we compute the first term $\nabla_y\mathcal{J}$ as

$$\begin{aligned}
(\nabla_y\mathcal{J}, h)_{\mathcal{X}} &= \frac{\partial}{\partial y}\left(\frac{1}{2}\|y-y^d\|^2_{\mathcal{X}} + \frac{\gamma}{2}\|y-\mathbb{E}[y]\|^2_{\mathcal{X}} + \frac{\mu}{2}\|u\|^2_{\mathcal{U}}, h\right)_{\mathcal{X}} \\
&= \frac{1}{2}\frac{\partial}{\partial y}\left(\|y-y^d\|^2_{\mathcal{X}}, h\right)_{\mathcal{X}} + \frac{\gamma}{2}\frac{\partial}{\partial y}\left(\|y-\mathbb{E}[y]\|^2_{\mathcal{X}}, h\right)_{\mathcal{X}} \\
&= \left(y-y^d, h\right)_{\mathcal{X}} + \gamma\left(y-\mathbb{E}[y], \frac{d(y-\mathbb{E}[y])}{dy}[h]\right)_{\mathcal{X}} \\
&= \left(y-y^d, h\right)_{\mathcal{X}} + \gamma\left(y-\mathbb{E}[y], h-\mathbb{E}[h]\right)_{\mathcal{X}} \\
&= \left(y-y^d, h\right)_{\mathcal{X}} + \gamma\left(y-\mathbb{E}[y], h\right)_{\mathcal{X}}\,,
\end{aligned} \tag{3.1.17}$$

where the last equality results from $(y - \mathbb{E}[y], \mathbb{E}[h])_{\mathcal{X}} = 0$.

For the $(\nabla_y (p, c(y,u))_{\mathcal{X}}, h)_{\mathcal{X}}$ term in $\nabla_y \mathcal{L}$, we get

$$\begin{aligned}
&\frac{\partial}{\partial y} ((\nabla(a\nabla y) - \mathbf{b}\cdot\nabla y + f + u, p)_{\mathcal{X}}, h)_{\mathcal{X}} \\
&= \frac{\partial}{\partial y}\Big(-\int_{D\times\Omega} a\nabla y\cdot\nabla p \, dx d\mathbb{P} + \int_{\partial D\times\Omega} a\nabla y \, p \, ds d\mathbb{P} + \int_{D\times\Omega} \mathbf{b}\cdot\nabla p y \, dx d\mathbb{P} \\
&\qquad - \int_{\partial D\times\Omega} y\,\mathbf{b}\cdot\nabla p \, ds d\mathbb{P} + \int_{D\times\Omega} (f+u)p \, dx d\mathbb{P}, h\Big)_{\mathcal{X}},
\end{aligned}$$

where we apply integration by parts to both diffusion and convection terms. Then, using the homogenous boundary conditions $y = p = 0$ on $\partial D \times \Omega$ and applying integration by parts on the first term only, gets us

$$\begin{aligned}
&\frac{\partial}{\partial y}\Big(\int_{D\times\Omega} y\nabla(a\nabla p) \, dx d\mathbb{P} - \int_{\partial D\times\Omega} y\, a\nabla p \, ds d\mathbb{P} + \int_{D\times\Omega} \mathbf{b}\cdot\nabla p \, y \, dx d\mathbb{P} \\
&\qquad + \int_{D\times\Omega} (f+u)p \, dx d\mathbb{P}, h\Big)_{\mathcal{X}}.
\end{aligned}$$

Finally, taking the derivative with respect to y we arrive at

$$(\nabla_y (p, c(y,u))_{\mathcal{X}}, h)_{\mathcal{X}} = (\nabla(a\nabla p) + \mathbf{b}\cdot\nabla p, h)_{\mathcal{X}}. \tag{3.1.18}$$

Thus, adding (3.1.17) and (3.1.18) together and equating 0 we have the so-called *adjoint equation*,

$$-\nabla(a(x,\omega)\nabla p(x,\omega)) - \mathbf{b}(x,\omega)\cdot\nabla p(x,\omega) \;=\; y - y_d + \gamma(y - \mathbb{E}[y]). \tag{3.1.19}$$

Next, the $\nabla_u \mathcal{L}$ reads

$$\begin{aligned}
\Big(\nabla_u \mathcal{J} + \nabla_u\Big(p, c(y,u)\Big)_{\mathcal{X}}, h\Big)_{\mathcal{X}} &= \frac{\partial}{\partial u}\Big(\frac{1}{2}\|y - y^d\|^2_{\mathcal{X}} + \frac{\gamma}{2}\|y - \mathbb{E}[y]\|^2_{\mathcal{X}} + \frac{\mu}{2}\|u\|^2_{\mathcal{U}} \\
&\qquad + (p, c(y,u))_{\mathcal{X}}, h\Big)_{\mathcal{X}} \\
&= (\mu u + p, h)_{\mathcal{X}}.
\end{aligned} \tag{3.1.20}$$

Notice, since $u \in \mathcal{U} = L^2(D)$, $(p,h)_{\mathcal{X}}$ must hold for all $h \in \mathcal{U}$ (as opposed to $h \in \mathcal{X}$), then we have $(p,h)_{\mathcal{X}} = (\mathbb{E}[p], h)_{\mathcal{U}}$. Also, using the trivial equality $(\mu u, h)_{\mathcal{X}} = (\mu u, h)_{\mathcal{U}}$, and setting (3.1.20) equal to 0 yield

$$\mu u + \mathbb{E}[p] = 0. \tag{3.1.21}$$

The equation (3.1.21) is also known the gradient of the reduced cost functional (3.1.13), $\nabla \mathcal{J}(u)$.

Finally, the derivative with respect to p reads

$$(\nabla_p \mathcal{L}, h)_{\mathcal{X}} = (c(y,u), h)_{\mathcal{X}}\,. \tag{3.1.22}$$

Equating to 0, we get

$$-\nabla(a(x,\omega)\nabla y(x,\omega)) + \mathbf{b}(x,\omega)\cdot\nabla y(x,\omega) \;=\; f+u, \tag{3.1.23}$$

Putting together (3.1.19), (3.1.21), (3.1.23) we obtain the optimality system. □

Using the (bi)-linear forms defined in (3.1.7a)-(3.1.7b), we obtain the weak form of optimality conditions.

$$a[y,v] - [u,v] = [f,v], \qquad v \in \mathcal{Y}, \tag{3.1.24a}$$

$$a[q,p] = [y-y^d, q] + \gamma\big[y - \mathbb{E}[y], q\big], \qquad q \in \mathcal{Y}, \tag{3.1.24b}$$

$$\mathbb{E}\Big[\int_D (p+\mu u)\,dx\Big] = 0, \qquad u \in \mathcal{U}, \tag{3.1.24c}$$

Having defined our problem and gave optimality conditions, now we are enabled to begin solution procedure in the next chapter.

CHAPTER 4

SOLUTION METHODS

In this chapter we detail the solution procedure to our robust deterministic optimal control problem (3.1.2)–(3.1.3). First, we discretize our problem in probability space using the standard MC method. We state the corresponding optimality conditions of the semi-discretized optimal control problem. Then, some error estimates will be derived in the semi-discrete setting. Next, we provide the fully-discrete optimal control problem, obtained by the standard continuous finite element method and derive the estimates for the fully-discrete scheme. Finally, the stochastic gradient descent method with Polyak's and Nesterov's momentums, are proposed as alternative optimization methods to solve the fully-discrete optimal control problem. We also provide the convergence estimates and computational cost of both variations.

To make notation easier we will omit the subscript in $\|\cdot\|_{L^2(\mathcal{D})}$ and write only $\|\cdot\|$. In addition, C denotes a generic positive constant independent of the mesh size h and differs in various estimates.

4.1 Approximation in Probability Space

In this section, we approximate the optimization problem (3.1.2)-(3.1.3) in probability space by applying the Monte Carlo (MC) method. The advantages are its easy implementation, interpretability, and flexible nature enabling parallel computing or stochastic methods effortlessly.

Denoting $y_i = y(x, \omega_i) \in \mathcal{Y}_\omega = H_0^1(D)$ with $\overrightarrow{\omega} = \{\omega_i\}_{i=1}^N$, we define the bilinear

forms for each ω_i as follows

$$\begin{aligned} a_\omega(y_i, v) &= \int_D \big(a(x,\omega_i)\nabla y_i \cdot \nabla v + \mathbf{b}(x,\omega_i)\cdot \nabla y_i\, v\big)\, dx,\ \forall y_i, v \in \mathcal{Y}_\omega, \\ (u, v) &= \int_D uv\, dx, \qquad \forall u \in \mathcal{U}, v \in \mathcal{Y}_\omega. \end{aligned}$$

Then, the corresponding semi–discrete optimization problem reads:

$$\begin{aligned} \min_{\widehat{u}\in\mathcal{U}} \widehat{\mathcal{J}}(\widehat{u}) = \frac{1}{2}E^{\vec{\omega}}_{MC}\left[\int_D \big(y_\omega(\widehat{u}) - y^d\big)^2\, dx\right] \\ + \frac{\gamma}{2}E^{\vec{\omega}}_{MC}\left[\int_D \big(y_\omega(\widehat{u}) - E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})]\big)^2\, dx\right] + \frac{\mu}{2}E^{\vec{\omega}}_{MC}\left[\int_D \widehat{u}^2\, dx\right], \end{aligned} \tag{4.1.2}$$

subject to

$$a_\omega(y_i, v) - (\widehat{u}, v) = (f, v), \qquad v \in \mathcal{Y}_\omega,\ i = 1, \ldots, N, \tag{4.1.3}$$

where $E^{\vec{\omega}}_{MC}[y_\omega] = \frac{1}{N}\sum_{i=1}^{N} y(x,\omega_i)$ is a Monte Carlo approximation for the expectation operator $\mathbb{E}$, and the subscript ω is used in expectation operator to underline the dependence on the random variable ω. Here, ω_i are $i.i.d.$ in Ω.

It follows from the strict convexity of $\widehat{\mathcal{J}}(\widehat{u})$ and the convexity of the admissible set $\mathcal{U}$ that the semi-discrete formulation (4.1.2)–(4.1.3) has a unique solution $(y_i(\widehat{u}), \widehat{u})$ if and only if there exists an adjoint function $p_i(\widehat{u}) \in \mathcal{Y}_\omega$ such that the triplet $(y_i(\widehat{u}), \widehat{u}, p_i(\widehat{u}))$ satisfies the following system of optimality conditions for $i = 1, \ldots, N$:

$$a_\omega(y_i(\widehat{u}), v) - (\widehat{u}, v) = (f, v), \qquad v \in \mathcal{Y}_\omega, \tag{4.1.4a}$$

$$a(q, p_i(\widehat{u})) = (y_i(\widehat{u}) - y^d, q) + \gamma\big(y_i(\widehat{u}) - E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})], q\big), \qquad q \in \mathcal{Y}_\omega, \tag{4.1.4b}$$

$$E^{\vec{\omega}}_{MC}\Big[\int_D (p_\omega(\widehat{u}) + \mu\widehat{u})\, dx\Big] = 0, \qquad \widehat{u} \in \mathcal{U}. \tag{4.1.4c}$$

Theorem 4.1.1. *Let $\big(y(u), u, p(u)\big)$ and $\big(y(\widehat{u}), \widehat{u}, p(\widehat{u})\big)$ be the solutions of problem (3.1.24) and (4.1.4), respectively. Then, we have*

$$\frac{\mu}{4}\mathbb{E}[\|u - \widehat{u}\|^2] + \gamma'\mathbb{E}[\|y(u) - y(\widehat{u})\|^2] \le \frac{1}{2N\mu}\mathbb{E}[\|p(u)\|^2], \tag{4.1.5}$$

provided that $\gamma' = \left(1 + \gamma - \frac{\gamma^2}{\mu c_{cr}^2 N^2}\right) > 0$, where c_{cr} is the coercivity constant.

Proof. By the optimality conditions in (3.1.24c) and (4.1.4c), we have

$$\begin{aligned}\mu\|u-\widehat{u}\|^2 &= \big(E^{\vec{\omega}}_{MC}[p_\omega(\widehat{u})]-\mathbb{E}[p(u)],u-\widehat{u}\big)\\ &\leq \big(E^{\vec{\omega}}_{MC}[p_\omega(u)]-\mathbb{E}[p(u)],u-\widehat{u}\big)\\ &\quad+\big(E^{\vec{\omega}}_{MC}[p_\omega(\widehat{u})]-E^{\vec{\omega}}_{MC}[p_\omega(u)],u-\widehat{u}\big).\end{aligned}\tag{4.1.6}$$

An application of Young's inequality for the first term in (4.1.6) gives us

$$\big(E^{\vec{\omega}}_{MC}[p_\omega(u)]-\mathbb{E}[p(u)],u-\widehat{u}\big)\leq\frac{1}{2\mu}\|\mathbb{E}[p(u)]-E^{\vec{\omega}}_{MC}[p_\omega(u)]\|^2+\frac{\mu}{2}\|u-\widehat{u}\|^2.\tag{4.1.7}$$

With the help of the following expression obtained from the bilinear forms for $i=1,\ldots,N$

$$\begin{aligned}(u-\widehat{u},p_i(\widehat{u})-p_i(u)) &= a_\omega(y_i(u)-y_i(\widehat{u}),p_i(\widehat{u})-p_i(u))\\ &= (1+\gamma)\big(y_i(\widehat{u})-y_i(u),y_i(u)-y_i(\widehat{u})\big)\\ &\quad+\gamma\big(E^{\vec{\omega}}_{MC}[y_\omega(u)]-E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})],y_i(u)-y_i(\widehat{u})\big),\end{aligned}$$

we write a bound for the second term in (4.1.6)

$$\begin{aligned}&\big(u-\widehat{u},E^{\vec{\omega}}_{MC}[p_\omega(\widehat{u})]-E^{\vec{\omega}}_{MC}[p_\omega(u)]\big)\\ &\quad\leq-(1+\gamma)E^{\vec{\omega}}_{MC}[\|y_\omega(u)-y_\omega(\widehat{u})\|^2]\\ &\qquad+\gamma\Big(E^{\vec{\omega}}_{MC}[y_\omega(u)]-E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})],E^{\vec{\omega}}_{MC}[y_\omega(u)]-E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})]\Big).\end{aligned}\tag{4.1.8}$$

An application of the coercivity of the bilinear form $a_\omega(\cdot,\cdot)$ and Young's inequality gives us

$$\begin{aligned}&\gamma\Big(E^{\vec{\omega}}_{MC}[y_\omega(u)]-E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})],E^{\vec{\omega}}_{MC}[y_\omega(u)]-E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})]\Big)\\ &\quad=\frac{\gamma}{N^2}\|\sum_{i=1}^N y_i(u)-y_i(\widehat{u})\|^2\\ &\quad\leq\frac{\gamma}{N^2}\sum_{i=1}^N\|y_i(u)-y_i(\widehat{u})\|^2\\ &\quad\leq\frac{\gamma}{c_{cr}N^2}\sum_{i=1}^N a_\omega(y_i(u)-y_i(\widehat{u}),y_i(u)-y_i(\widehat{u}))\\ &\quad=\frac{\gamma}{c_{cr}N}\big(u-\widehat{u},E^{\vec{\omega}}_{MC}[y_\omega(u)]-E^{\vec{\omega}}_{MC}[y_\omega(\widehat{u})]\big)\\ &\quad\leq\frac{\mu}{4}\|u-\widehat{u}\|^2+\frac{\gamma^2}{\mu c_{cr}^2N^2}E^{\vec{\omega}}_{MC}[\|y_\omega(u)-y_\omega(\widehat{u})\|^2].\end{aligned}\tag{4.1.9}$$

Inserting (4.1.7)–(4.1.9) into (4.1.6) yields

$$\frac{\mu}{4}\|u-\widehat{u}\|^2+\underbrace{\left(1+\gamma-\frac{\gamma^2}{\mu c_{cr}^2 N^2}\right)}_{\gamma'}E_{MC}^{\overrightarrow{\omega}}[\|y_\omega(u)-y_\omega(\widehat{u})\|^2]$$
$$\leq \frac{1}{2\mu}\|\mathbb{E}[p(u)]-E_{MC}^{\overrightarrow{\omega}}[p_\omega(u)]\|^2. \tag{4.1.10}$$

By taking expectation of (4.1.10) with respect to $\overrightarrow{\omega}$ with the fact that the MC estimator is unbiased, i.e., $\mathbb{E}[E_{MC}^{\overrightarrow{\omega}}[X(\omega)]]=\mathbb{E}[X]$ for a random variable $X:\Omega\to\mathbb{R}$, we obtain

$$\begin{aligned}
\frac{\mu}{4}\|u-\widehat{u}\|^2+\gamma'\mathbb{E}[\|y(u)-y(\widehat{u})\|^2] &\leq \frac{1}{2\mu}\mathbb{E}[\|\mathbb{E}[p(u)]-E_{MC}^{\overrightarrow{\omega}}[p_\omega(u)]\|^2]\\
&\leq \frac{1}{2\mu}\mathbb{E}[\frac{1}{N^2}\sum_{i=1}^{N}\|p_i(u)-\mathbb{E}[p(u)]\|^2]\\
&\leq \frac{1}{2\mu}\frac{1}{N}\mathbb{E}[\|p(u)-\mathbb{E}[p(u)]\|^2]\\
&\leq \frac{1}{2\mu}\frac{1}{N}\mathbb{E}[\|p(u)\|^2],
\end{aligned}$$

which is the desired result. □

To finalize the discretization process, all that remains is to discretize our problem (4.1.2)–(4.1.3) in the spatial domain.

4.2 Approximation in Physical Space

Now, together with the MC approximation in probability space, we consider the fully discretized optimal control problem, discretized by using the continuous finite element method in the spatial domain.

Let $\{\mathcal{T}_h\}_h$ be a set of shape-regular simplicial triangulations of D. Each mesh $\mathcal{T}_h$ consists of closed triangles such that $\overline{D}=\bigcup_{K\in\mathcal{T}_h}\overline{K}$ holds. The space of piecewise-continuous linear polynomials on $K\in\mathcal{T}_h$ is denoted by $\mathbb{P}^1(K),\ \forall K\in\mathcal{T}_h$. The diameter of an element K and the maximum value of the element diameter are denoted by h_K and $h=\max_{K\in\mathcal{T}_h}h_K$, respectively.

We introduce the spaces of the discrete state and control, respectively, by

$$\mathcal{Y}_\omega^h = \left\{ y_\omega^h \in C^0(D) \,:\, y_\omega^h \,|_K \in \mathbb{P}^1(K) \quad \forall K \in \mathcal{T}_h, \quad y_\omega^h|_{\partial\mathcal{D}} = 0 \right\}, \tag{4.2.1a}$$

$$\mathcal{U}^h = \left\{ u^h \in C^0(D) \,:\, u^h \,|_K \in \mathbb{P}^1(K) \quad \forall K \in \mathcal{T}_h \right\}. \tag{4.2.1b}$$

We note that $\mathcal{Y}_\omega^h \subset \mathcal{Y}_\omega$. Then, the fully discretized optimal control problem is as follows

$$\begin{aligned} \min_{\widehat{u}^h \in \mathcal{U}^h} \widehat{\mathcal{J}}^h(\widehat{u}^h) =& \frac{1}{2} E_{MC}^{\vec{\omega}}\big[\|y_\omega^h(\widehat{u}^h) - y^d\|^2\big] \\ &+ \frac{\gamma}{2} E_{MC}^{\vec{\omega}}\big[\|y_\omega^h(\widehat{u}^h) - E_{MC}^{\vec{\omega}}[y_\omega^h(\widehat{u}^h)]\|^2\big] + \frac{\mu}{2} E_{MC}^{\vec{\omega}}\big[\|\widehat{u}^h\|^2\big] \end{aligned} \tag{4.2.2a}$$

such that $y_i^h \in \mathcal{Y}_\omega^h$ solving

$$a_\omega(y_i^h, v^h) = (f + \widehat{u}^h, v^h), \quad \forall v^h \in \mathcal{Y}_\omega^h,\ i = 1, \ldots, N. \tag{4.2.2b}$$

Analogously, the fully-discrete formulation (4.2.2) has a unique solution $(y_i^h, \widehat{u}^h)$ if and only if there exists an adjoint function $p_i^h \in \mathcal{Y}_\omega^h$ such that the triplet $(y_i^h, \widehat{u}^h, p_i^h)$ satisfies the following system of optimality conditions:

$$a_\omega(y_i^h, v^h) - (\widehat{u}^h, v^h) = (f, v^h), \qquad v^h \in \mathcal{Y}_\omega^h, \tag{4.2.3a}$$

$$a_\omega(q^h, p_i^h) = (y_i^h - y^d, q^h) + \gamma\big(y_i^h - E_{MC}^{\vec{\omega}}[y_\omega^h], q^h\big), \qquad q^h \in \mathcal{Y}_\omega^h, \tag{4.2.3b}$$

$$E_{MC}^{\vec{\omega}}\Big[\int_{\mathcal{D}} (p_\omega^h + \mu\widehat{u}^h)\, dx\Big] = 0, \qquad u^h \in \mathcal{U}^h. \tag{4.2.3c}$$

Lemma 4.2.1. *Assume that* $\big(y(\widehat{u}), \widehat{u}, p(\widehat{u})\big)$ *and* $\big(y^h(\widehat{u}^h), \widehat{u}^h, p^h(\widehat{u}^h)\big)$ *be the solutions of problem (4.1.4) and (4.2.3), respectively. Then, there holds*

$$\begin{aligned} \frac{\mu}{2}\|\widehat{u} - \widehat{u}^h\|^2 \;\le\; & \frac{1}{2\mu} E_{MC}^{\vec{\omega}}[\|\widetilde{p}_\omega^h(\widehat{u}) - p_\omega(\widehat{u})\|^2] + \frac{1+2\gamma}{2} E_{MC}^{\vec{\omega}}[\|y_\omega(\widehat{u}) - y_\omega^h(\widehat{u})\|^2] \\ & + \left(\gamma - \frac{1}{2}\right) E_{MC}^{\vec{\omega}}[\|y_\omega(\widehat{u}) - y_\omega^h(\widehat{u}^h)\|^2], \end{aligned} \tag{4.2.4}$$

where $\widetilde{p}_i^h(\widehat{u})$ *solves*

$$a_\omega^*(\widetilde{p}_i^h(\widehat{u}), v^h) = \big(y_i(\widehat{u}) - y^d + \gamma(y_i(\widehat{u}) - \mathbb{E}[y_\omega(\widehat{u})]), v^h\big), \ \ \forall v^h \in \mathcal{Y}_\omega^h, \ \ i = 1, \ldots, N. \tag{4.2.5}$$

Proof. Optimality conditions (4.1.4c) and (4.2.3c) give us

$$\begin{aligned} \mu\|\widehat{u} - \widehat{u}^h\|^2 &= \big(\mu(\widehat{u} - \widehat{u}^h), \widehat{u} - \widehat{u}^h\big) \\ &= \Big(E_{MC}^{\vec{\omega}}[p_\omega^h(\widehat{u}^h)] - E_{MC}^{\vec{\omega}}[p_\omega(\widehat{u})], \widehat{u} - \widehat{u}^h\Big) \\ &= \Big(E_{MC}^{\vec{\omega}}[p_\omega^h(\widehat{u}^h) - \widetilde{p}_\omega^h(\widehat{u})], \widehat{u} - \widehat{u}^h\Big) + \Big(E_{MC}^{\vec{\omega}}[\widetilde{p}_\omega^h(\widehat{u}) - p_\omega(\widehat{u})], \widehat{u} - \widehat{u}^h\Big). \end{aligned} \tag{4.2.6}$$

We start by estimating the first term in (4.2.6). With the help of the bilinear systems in (4.2.3) and (4.2.5), we obtain

$$\begin{aligned}\left(p_i^h(\widehat{u}^h)-\widetilde{p}_i^h(\widehat{u}),\widehat{u}-\widehat{u}^h\right) &= a_\omega(y_i^h(\widehat{u})-y_i^h(\widehat{u}^h),p_i^h(\widehat{u}^h)-\widetilde{p}_i^h(\widehat{u}))\\ &= (1+\gamma)\underbrace{\left(y_i^h(\widehat{u}^h)-y_i(\widehat{u}),y_i^h(\widehat{u})-y_i^h(\widehat{u}^h)\right)}_{M_1}\\ &\quad+\gamma\underbrace{\left(E_{MC}^{\vec{\omega}}[y_\omega(\widehat{u})]-E_{MC}^{\vec{\omega}}[y_\omega^h(\widehat{u}^h)],y_i^h(\widehat{u})-y_i^h(\widehat{u}^h)\right)}_{M_2}.\end{aligned} \tag{4.2.7}$$

Then, for $i=1,\dots,N$, Young's inequality yields

$$\begin{aligned}M_1 &= \left(y_i^h(\widehat{u}^h)-y_i(\widehat{u}),y_i(\widehat{u})-y_i^h(\widehat{u}^h)\right)+\left(y_i^h(\widehat{u}^h)-y_i(\widehat{u}),y_i^h(\widehat{u})-y_i(\widehat{u})\right)\\ &\le -\|y_i(\widehat{u})-y_i^h(\widehat{u}^h)\|^2+\frac{1}{2}\|y_i^h(\widehat{u}^h)-y_i(\widehat{u})\|^2+\frac{1}{2}\|y_i(\widehat{u})-y_i^h(\widehat{u})\|^2\\ &= -\frac{1}{2}\|y_i(\widehat{u})-y_i^h(\widehat{u}^h)\|^2+\frac{1}{2}\|y_i(\widehat{u})-y_i^h(\widehat{u})\|^2,\end{aligned} \tag{4.2.8}$$

and

$$\begin{aligned}M_2 &= \left(E_{MC}^{\vec{\omega}}[y_\omega(\widehat{u})-y_\omega^h(\widehat{u}^h)],y_i^h(\widehat{u})-y_i(\widehat{u})\right)\\ &\quad+\left(E_{MC}^{\vec{\omega}}[y_\omega(\widehat{u})-y_\omega^h(\widehat{u}^h)],y_i(\widehat{u})-y_i^h(\widehat{u}^h)\right)\\ &\le \|E_{MC}^{\vec{\omega}}[y_\omega(\widehat{u})-y_\omega^h(\widehat{u}^h)]\|^2+\frac{1}{2}\|y_i^h(\widehat{u})-y_i(\widehat{u})\|^2+\frac{1}{2}\|y_i(\widehat{u})-y_i^h(\widehat{u}^h)\|^2.\end{aligned} \tag{4.2.9}$$

Inserting (4.2.8) and (4.2.9) into (4.2.7), we get

$$\begin{aligned}\left(p_i^h(\widehat{u}^h)-\widetilde{p}_i^h(\widehat{u}),\widehat{u}-\widehat{u}^h\right) &\le -\frac{1}{2}\|y_i(\widehat{u})-y_i^h(\widehat{u}^h)\|^2+\frac{1+2\gamma}{2}\|y_i(\widehat{u})-y_i^h(\widehat{u})\|^2\\ &\quad+\gamma\|E_{MC}^{\vec{\omega}}[y_\omega(\widehat{u})-y_\omega^h(\widehat{u}^h)]\|^2.\end{aligned} \tag{4.2.10}$$

By using the operator $E_{MC}^{\vec{\omega}}[\cdot]$ on (4.2.10), inserting it to (4.2.6), and then applying Young's inequality and Fubini's theorem, we obtain

$$\begin{aligned}\mu\|\widehat{u}-\widehat{u}^h\|^2 &= \left(E_{MC}^{\vec{\omega}}[\widetilde{p}_\omega^h(\widehat{u})-p_\omega(\widehat{u})],\widehat{u}-\widehat{u}^h\right)+\frac{1+2\gamma}{2}E_{MC}^{\vec{\omega}}[\|y_\omega(\widehat{u})-y_\omega^h(\widehat{u})\|^2]\\ &\quad+\left(\gamma-\frac{1}{2}\right)E_{MC}^{\vec{\omega}}[\|y_\omega(\widehat{u})-y_\omega^h(\widehat{u}^h)\|^2]\\ &\le \frac{1}{2\mu}E_{MC}^{\vec{\omega}}[\|\widetilde{p}_\omega^h(\widehat{u})-p_\omega(\widehat{u})\|^2]+\frac{1+2\gamma}{2}E_{MC}^{\vec{\omega}}[\|y_\omega(\widehat{u})-y_\omega^h(\widehat{u})\|^2]\\ &\quad+\frac{\mu}{2}\|\widehat{u}-\widehat{u}^h\|^2+\left(\gamma-\frac{1}{2}\right)E_{MC}^{\vec{\omega}}[\|y_\omega(\widehat{u})-y_\omega^h(\widehat{u}^h)\|^2],\end{aligned} \tag{4.2.11}$$

which completes the proof. □

Next, we find a bound for $E^{\overrightarrow{\omega}}_{MC}[\|y_\omega(\widehat{u})-y^h_\omega(\widehat{u}^h)\|^2]$ in terms of a priori error estimates of the state solution.

Lemma 4.2.2. *Assume that* $\big(y(\widehat{u}),\widehat{u},p(\widehat{u})\big)$ *and* $\big(y^h(\widehat{u}^h),\widehat{u}^h,p^h(\widehat{u}^h)\big)$ *be the solutions of problem (4.1.4) and (4.2.3), respectively. Then, there exists a constant* C *such that*

$$E^{\overrightarrow{\omega}}_{MC}[\|y_\omega(\widehat{u})-y^h_\omega(\widehat{u}^h)\|^2]\leq C\Big(E^{\overrightarrow{\omega}}_{MC}[\|y_\omega(\widehat{u})-\widetilde{y}^h_\omega(\widehat{u})\|^2]+\|\widehat{u}-\widehat{u}^h\|^2\Big),$$

where $\widetilde{y}^h_i(\widehat{u})$ *solves*

$$a_\omega(\widetilde{y}^h_i(\widehat{u}),v^h)=\big(\widehat{u}+f,v^h\big)\quad\forall v^h\in\mathcal{Y}^h_\omega,\quad i=1,\dots,N. \tag{4.2.12}$$

Proof. By the coercivity of $a_\omega(\cdot,\cdot)$, for $i=1,\dots,N$ we have

$$\begin{aligned}&\|y_i(\widehat{u})-y^h_i(\widehat{u}^h)\|^2\\&\quad\leq\frac{1}{c_{cr}}\Big(a_\omega\big(y_i(\widehat{u})-y^h_i(\widehat{u}^h),y_i(\widehat{u})-\widetilde{y}^h_i(\widehat{u})\big)+a_\omega\big(y_i(\widehat{u})-y^h_i(\widehat{u}^h),\widetilde{y}^h_i(\widehat{u})-y^h_i(\widehat{u}^h)\big)\Big).\end{aligned} \tag{4.2.13}$$

Now, we obtain a bound for the first term in (4.2.13) by the continuity of $a_\omega(\cdot,\cdot)$ and Young's inequality

$$\begin{aligned}a_\omega(y_i(\widehat{u})-y^h_i(\widehat{u}^h),y_i(\widehat{u})-\widetilde{y}^h_i(\widehat{u}))&\leq c_{cnt}\|y_i(\widehat{u})-y^h_i(\widehat{u}^h)\|\|y_i(\widehat{u})-\widetilde{y}^h_i(\widehat{u})\|\\&\leq\frac{c_{cr}}{4}\|y_i(\widehat{u})-y^h_i(\widehat{u}^h)\|^2+\frac{c^2_{cnt}}{c_{cr}}\|y_i(\widehat{u})-\widetilde{y}^h_i(\widehat{u})\|^2.\end{aligned} \tag{4.2.14}$$

Next, an application of the bilinear forms and Young's inequality yields an estimate for the second term in (4.2.13)

$$\begin{aligned}&a_\omega(y_i(\widehat{u})-y^h_i(\widehat{u}^h),\widetilde{y}^h_i(\widehat{u})-y^h_i(\widehat{u}^h))\\&\quad=\big(\widehat{u}-\widehat{u}^h,\widetilde{y}^h_i(\widehat{u})-y^h_i(\widehat{u}^h)\big)\\&\quad\leq\big(\widehat{u}-\widehat{u}^h,\widetilde{y}^h_i(\widehat{u})-y_i(\widehat{u})\big)+\big(\widehat{u}-\widehat{u}^h,y_i(\widehat{u})-y^h_i(\widehat{u}^h)\big)\\&\quad\leq\frac{1}{2}\|\widehat{u}-\widehat{u}^h\|^2+\frac{1}{2}\|\widetilde{y}^h_i(\widehat{u})-y_i(\widehat{u})\|^2+\frac{1}{c_{cr}}\|\widehat{u}-\widehat{u}^h\|^2+\frac{c_{cr}}{4}\|y_i(\widehat{u})-y^h_i(\widehat{u}^h)\|^2.\end{aligned} \tag{4.2.15}$$

Inserting (4.2.14) and (4.2.15) into (4.2.13), we obtain

$$\frac{1}{2}\|y_i(\widehat{u})-y^h_i(\widehat{u}^h)\|^2\leq\frac{2c^2_{cnt}+c_{cr}}{2c^2_{cr}}\|y_i(\widehat{u})-\widetilde{y}^h_i(\widehat{u})\|^2+\frac{c_{cr}+2}{2c^2_{cr}}\|\widehat{u}-\widehat{u}^h\|^2. \tag{4.2.16}$$

Last, the usage of the operator $E^{\overrightarrow{\omega}}_{MC}[\cdot]$ on (4.2.16) produces the desired result. □

Before concluding the error bound for the FE solution, we state a standard finite element error estimate which will be needed in the rest of the thesis.

Theorem 4.2.3. ([15],Theorem 5.4.8) *Suppose that there exists a global interpolator $\mathcal{I}^h$ for all the members of the family $\{\mathcal{T}_h\}_h$ constituting an approximation of order m, also satisfying*

$$\|u - \mathcal{I}^h u\|_{H^1(D)} \leq Ch^{m-1}|u|_{H^m(D)},$$

where C is a non-negative constant. Further, suppose we have $\mathcal{I}^h(\mathcal{Y} \cap C^k(D)) \subset \mathcal{Y}^h_\omega$. Then, the following holds

$$\begin{aligned} \|u - u_h\|_{L^2(D)} &\leq Ch\|u - u_h\|_{H^1(D)} \\ &\leq Ch^m|u|_{H^m(D)}. \end{aligned} \tag{4.2.17}$$

Theorem 4.2.4. *Assume that $\big(y(\widehat{u}), \widehat{u}, p(\widehat{u})\big)$ and $\big(y^h(\widehat{u}^h), \widehat{u}^h, p^h(\widehat{u}^h)\big)$ be the solutions of problem (4.1.4) and (4.2.3), respectively. Then, there exists a constant $C > 0$ independent of h such that*

$$\|\widehat{u} - \widehat{u}^h\|^2 \leq Ch^4\Big(\mathbb{E}[\|y_\omega(\widehat{u})\|^2_{H^2(D)}] + \mathbb{E}[\|p_\omega(\widehat{u})\|^2_{H^2(D)}]\Big). \tag{4.2.18}$$

Proof. Adding the results of Lemma 4.2.1, and Lemma 4.2.2, and applying Theorem 4.2.3 we obtain

$$\begin{aligned} \|\widehat{u} - \widehat{u}^h\|^2 &\leq C\Big(E^{\overrightarrow{\omega}}_{MC}[\|\widetilde{p}^h_\omega(\widehat{u}) - p_\omega(\widehat{u})\|^2] + E^{\overrightarrow{\omega}}_{MC}[\|y_\omega(\widehat{u}) - y^h_\omega(\widehat{u})\|^2] \\ &\quad + E^{\overrightarrow{\omega}}_{MC}[\|y_\omega(\widehat{u}) - \widetilde{y}^h_\omega(\widehat{u})\|^2]\Big) \\ &\leq Ch^4\Big(E^{\overrightarrow{\omega}}_{MC}[|y_\omega(\widehat{u})\|^2_{H^2(D)}] + E^{\overrightarrow{\omega}}_{MC}[\|p_\omega(\widehat{u})\|^2_{H^2(D)}]\Big). \end{aligned} \tag{4.2.19}$$

By taking the expectation of (4.2.19) with respect to $\overrightarrow{\omega}$ with the fact that the MC estimator is unbiased, the desired result is obtained. □

4.3 Stochastic Gradient Descent with Momentum for Fully Discrete Problem

The fully discretized optimal control problem (4.2.2) obtained by the Monte Carlo in probability space and by the finite element method in the spatial domain, now is approximated by using the stochastic momentum methods, i.e., Polyak's and Nesterov's momentum methods.

Before stating the optimization techniques, we need to show the essential properties called the smoothness and the strong convexity of the functional $\widehat{\mathcal{J}}^h$.

Lemma 4.3.1. *For the cost functional $\widehat{\mathcal{J}}^h$ as defined in (4.2.2) the following smoothness condition holds*

$$\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_2^h) - \nabla\widehat{\mathcal{J}}^h(\widehat{u}_1^h)\| \leq L\|\widehat{u}_2^h - \widehat{u}_1^h\| \qquad \forall \widehat{u}_1^h,\, \widehat{u}_2^h \in \mathcal{U}^h, \tag{4.3.1}$$

with $L = \mu + \frac{1+\gamma}{c_{cr}^2} + \frac{\gamma}{c_{cr}}$, where c_{cr} is the coercivity constant.

Proof. Without any loss of generality, we shall prove the claim for a single sample ω and then apply the MC estimator to generalize what is proven to all N number of samples. For all $\widehat{u}_1^h, \widehat{u}_2^h \in \mathcal{U}^h$, the optimality condition (4.2.3c) is written in terms of the gradient of the cost functional for a single sample $\widehat{\mathcal{J}}_i^h$

$$\nabla\widehat{\mathcal{J}}_i^h(\widehat{u}_2^h) - \nabla\widehat{\mathcal{J}}_i^h(\widehat{u}_1^h) = \mu(\widehat{u}_2^h - \widehat{u}_1^h) + p_i^h(\widehat{u}_2^h) - p_i(\widehat{u}_1^h), \quad \text{for } i = 1, \ldots, N. \tag{4.3.2}$$

Using the coercivity of the bilinear form $a_\omega(\cdot,\cdot)$ and the Cauchy-Schwarz inequality yields

$$\begin{aligned}
\|p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h)\|^2 &\leq \frac{1}{c_{cr}} a_\omega^*\big(p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h), p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h)\big) \\
&= \frac{1}{c_{cr}}\Big((1+\gamma)(y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h)), p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h)\Big) \\
&\quad + \gamma\Big(E_{MC}^{\overrightarrow{\omega}}\big[y_\omega^h(\widehat{u}_1^h) - y_\omega^h(\widehat{u}_2^h)\big], p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h)\Big) \\
&\leq \frac{1+\gamma}{c_{cr}}\|p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h)\|\|y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h)\| \\
&\quad + \frac{\gamma}{N}\sum_{i=1}^{N}\|y_i^h(\widehat{u}_1^h) - y_i^h(\widehat{u}_2^h)\|\|p_i^h(\widehat{u}_2^h) - p_i^h(\widehat{u}_1^h)\|.
\end{aligned} \tag{4.3.3}$$

Similarly, we obtain

$$\begin{aligned}
\|y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h)\|^2 &\leq \frac{1}{c_{cr}} a_\omega\big(y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h), y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h)\big) \\
&= \frac{1}{c_{cr}}\big(\widehat{u}_2^h - \widehat{u}_1^h, y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h)\big) \\
&\leq \frac{1}{c_{cr}}\|y_i^h(\widehat{u}_2^h) - y_i^h(\widehat{u}_1^h)\|\|\widehat{u}_2^h - \widehat{u}_1^h\|.
\end{aligned} \tag{4.3.4}$$

Combining (4.3.3) and (4.3.4) in (4.3.2) and using the linearity of the operator $E_{MC}^{\overrightarrow{\omega}}[\cdot]$, we obtain the desired result. □

Lemma 4.3.2. *The cost functional $\widehat{\mathcal{J}}^h$ given in (4.2.2) is μ–strongly convex such that*

$$\mu\|\widehat{u}_2^h - \widehat{u}_1^h\|^2 \leq \big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_2^h) - \nabla\widehat{\mathcal{J}}^h(\widehat{u}_1^h), \widehat{u}_2^h - \widehat{u}_1^h\big) \quad \forall \widehat{u}_1^h,\, \widehat{u}_2^h \in \mathcal{U}^h. \tag{4.3.5}$$

Proof. For every $\widehat{u}_1^h,\, \widehat{u}_2^h \in \mathcal{U}^h$, the bilinear form $a_\omega(\cdot,\cdot)$ yields that

$$\begin{aligned}
&\big(\widehat{u}_2^h - \widehat{u}_1^h, \nabla\widehat{\mathcal{J}}^h(\widehat{u}_2^h) - \nabla\widehat{\mathcal{J}}^h(\widehat{u}_1^h)\big) \\
&\quad= \big(\widehat{u}_2^h - \widehat{u}_1^h, E_{MC}^{\overrightarrow{\omega}}[\mu(\widehat{u}_2^h - \widehat{u}_1^h) + p_\omega^h(\widehat{u}_2^h) - p_\omega^h(\widehat{u}_1^h)]\big) \\
&\quad= \mu E_{MC}^{\overrightarrow{\omega}}\big[\|\widehat{u}_2^h - \widehat{u}_1^h\|^2\big] + \big(\widehat{u}_2^h - \widehat{u}_1^h, E_{MC}^{\overrightarrow{\omega}}[p_\omega^h(\widehat{u}_2^h) - p_\omega^h(\widehat{u}_1^h)]\big) \\
&\quad= \mu E_{MC}^{\overrightarrow{\omega}}\big[\|\widehat{u}_2^h - \widehat{u}_1^h\|^2\big] + E_{MC}^{\overrightarrow{\omega}}\big[a_\omega(y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h), p_\omega^h(\widehat{u}_2^h) - p_\omega^h(\widehat{u}_1^h))\big] \\
&\quad= \mu E_{MC}^{\overrightarrow{\omega}}[\|\widehat{u}_2^h - \widehat{u}_1^h\|^2] + E_{MC}^{\overrightarrow{\omega}}\Big[(1+\gamma)\big(y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h), y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h)\big)\Big] \\
&\qquad- \gamma E_{MC}^{\overrightarrow{\omega}}\Big[\Big(E_{MC}^{\overrightarrow{\omega}}[y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h)], y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h)\Big)\Big] \\
&\quad\geq \mu\|\widehat{u}_2^h - \widehat{u}_1^h\|^2 + (1+\gamma)E_{MC}^{\overrightarrow{\omega}}\big[\|y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h)\|^2\big] \\
&\qquad- \gamma E_{MC}^{\overrightarrow{\omega}}\big[\|y_\omega^h(\widehat{u}_2^h) - y_\omega^h(\widehat{u}_1^h)\|^2\big] \\
&\quad\geq \mu\|\widehat{u}_2^h - \widehat{u}_1^h\|^2,
\end{aligned} \tag{4.3.6}$$

which completes the proof. □

We note that L–smoothness (4.3.1) and μ–strong convexity (4.3.5) can also be expressed as the following, see, e.g., [16],

$$\widehat{\mathcal{J}}^h(\widehat{u}^h) \leq \widehat{\mathcal{J}}^h(\widehat{u}_j^h) + \big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h), \widehat{u}^h - \widehat{u}_j^h\big) + \frac{L}{2}\|\widehat{u}^h - \widehat{u}_j^h\|^2, \tag{4.3.7}$$

and

$$\widehat{\mathcal{J}}^h(\widehat{u}^h) \geq \widehat{\mathcal{J}}^h(\widehat{u}_j^h) + \big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h), \widehat{u}^h - \widehat{u}_j^h\big) + \frac{\mu}{2}\|\widehat{u}^h - \widehat{u}_j^h\|^2, \tag{4.3.8}$$

respectively. Then, from (4.3.7) and (4.3.8), one can easily derive

$$\frac{\mu}{2}\|\widehat{u}_j^h - \widehat{u}^h\|^2 \leq \widehat{\mathcal{J}}^h(\widehat{u}_j^h) - \widehat{\mathcal{J}}^h(\widehat{u}^h) \leq \frac{L}{2}\|\widehat{u}_j^h - \widehat{u}^h\|^2, \tag{4.3.9}$$

by switching $\widehat{u}^h$ and $\widehat{u}_j^h$ and using the fact $\nabla\widehat{\mathcal{J}}^h(\widehat{u}^h) = 0$.

4.3.1 Stochastic Polyak's Momentum

Now, let us investigate Polyak's momentum on stochastic gradient descent, a method also known as stochastic heavy ball, for the approximation of the fully discrete optimal control problem (4.2.2). Let $\overrightarrow{\omega}_j = (\omega^{(1)_j}, \ldots, \omega^{(N_j)_j})$ be the N_j $i.i.d.$ realizations

which are drawn independent of the previous iterations, then the update formula of the stochastic Polyak's momentum is as demonstrated

$$\widehat{u}_{j+1}^h = \widehat{u}_j^h - \alpha E_{MC}^{\overrightarrow{\omega}_j}[\nabla \widehat{\mathcal{J}}^h(\widehat{u}_j^h)] + \beta_P(\widehat{u}_j^h - \widehat{u}_{j-1}^h), \tag{4.3.10}$$

where α is the step–size of the algorithm and β_P corresponds to positive Polyak's momentum parameter. It essentially computes the gradient making an advancement in that direction, and follow that by a correction in a direction related to the previous step. By adding the correction term, it allows larger step size α and accelerate convergence.

To proceed with the analysis of the Polyak's momentum, we first define the history of iteration as $\mathcal{F}_j = \{u_0, \ldots, u_j\}$, and introduce the following lemma, which states the convergence of nonnegative real numbers under the specified conditions.

Lemma 4.3.3. *[50, Lemma 9] For a given $F_0 = F_1 \geq 0$ let $\{F_j\}_{j\geq 0}$ be a nonnegative sequence of real numbers satisfying the following condition*

$$F_{j+1} \leq a_1 F_j + a_2 F_{j-1}, \qquad \forall j \geq 1, \tag{4.3.11}$$

where $a_2 \geq 0,\ a_1 + a_2 < 1$ and at least one of a_1, a_2 is positive. Then, the sequence $\{F_j\}_{j\geq 0}$ satisfies

$$F_{j+1} \leq q^j(1+\delta)F_0, \qquad \forall j \geq 1, \tag{4.3.12}$$

where $q = \frac{a_1+\sqrt{a_1^2+4a_2}}{2} < 1$ and $\delta = q - a_1$. Furthermore, the following holds, and holds with the equality if and only if $a_2 = 0$,

$$q \geq a_1 + a_2.$$

Proof. First, let $\delta = \frac{-a_1+\sqrt{a_1^2+4a_2}}{2}$ and observe that

$$(a_1 + \delta)\delta - a_2 = 0 \tag{4.3.13}$$

holds. Now, see that when $a_2 > 0$, we have $q > 0$. When $a_2 = 0$, then by assumption $a_1 > 0$ and $q > 0$. Also, it can be seen directly from $a_1 + a_2 < 1$ that $q < 1$. From the fact that $a_2 \geq 0$ we can say $\delta \geq 0$, and from (4.3.13) we have $a_2 \leq (a_1 + \delta)\delta$. With

those at hand, and since $a_1 = q - \delta$ and $a_2 = q\delta$, we have $a_1 + a_2 \leq q$. Adding δF_j to both sides of (4.3.11), we get

$$F_{j+1} + \delta F_j \leq q(F_j + \delta F_{j-1}). \tag{4.3.14}$$

By unraveling the recurrence (4.3.14) we obtain (4.3.12). □

Theorem 4.3.4. *Let $\widehat{u}_0^h = \widehat{u}_1^h \in U^h$ and $\{\widehat{u}_j^h\}_{j\geq 0}$ be the sequence of random iterates produced by stochastic Polyak's momentum (4.3.10). Under the assumptions of $1 < \alpha L < 2$ and $0 \leq \beta_P$, and the expressions*

$$a_1 = \frac{6\beta_P - 4\alpha\mu + 4\beta_P^2 - 3L\alpha\beta_P - \alpha\beta_P\mu + 2L\alpha^2\mu + 2}{\alpha\mu - L\alpha + 2}, \tag{4.3.15a}$$

$$a_2 = \frac{2\beta_P + 4\beta_P^2}{\alpha\mu - L\alpha + 2} \tag{4.3.15b}$$

satisfying $a_1 + a_2 < 1$, the stochastic Polyak's momentum (4.3.10) yields

$$\mathbb{E}\big[\|\widehat{u}_j^h - \widehat{u}^h\|^2\big] \leq q^j(1+\delta)\|\widehat{u}_0^h - \widehat{u}^h\|^2, \tag{4.3.16}$$

where $q = \frac{a_1 + \sqrt{a_1^2 + 4a_2}}{2} < 1$ and $\delta = q - a_1$.

Proof. Using the update formula (4.3.10), decompose the expression for $i = 1, \ldots, N$ as follows:

$$\begin{aligned}
\|\widehat{u}_{j+1}^h - \widehat{u}^h\|^2 &= \|\widehat{u}_j^h - \alpha\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i + \beta_P(\widehat{u}_j^h - \widehat{u}_{j-1}^h) - \widehat{u}^h\|^2 \\
&= \underbrace{\|\widehat{u}_j^h - \alpha\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i - \widehat{u}^h\|^2}_{T_1} \\
&\quad + \underbrace{2\big(\widehat{u}_j^h - \alpha\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i - \widehat{u}^h, \beta_P(\widehat{u}_j^h - \widehat{u}_{j-1}^h)\big)}_{T_2} \\
&\quad + \underbrace{\beta_P^2\|\widehat{u}_j^h - \widehat{u}_{j-1}^h\|^2}_{T_3}.
\end{aligned} \tag{4.3.17}$$

By (4.3.8) and (4.3.9), the first expression T_1 on the right–hand side of (4.3.17) can be rewritten as

$$\begin{aligned}
T_1 &= \|\widehat{u}_j^h - \widehat{u}^h\|^2 - 2\alpha\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i, \widehat{u}_j^h - \widehat{u}^h\big) + \alpha^2\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\|^2 \\
&\leq \|\widehat{u}_j^h - \widehat{u}^h\|^2 - 2\alpha\big(\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i - \widehat{\mathcal{J}}^h(\widehat{u}^h)_i\big) \\
&\quad - \alpha\mu\|\widehat{u}^h - \widehat{u}_j^h\|^2 + \alpha^2\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\|^2 \\
&\leq \big(1 - 2\alpha\mu\big)\|\widehat{u}_j^h - \widehat{u}^h\|^2 + \alpha^2\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\|^2.
\end{aligned} \tag{4.3.18}$$

Then, we use the smoothness property (4.3.7) and the update formula in (4.3.10) to find a bound for $\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\|^2$ in (4.3.18)

$$\begin{aligned}\widehat{\mathcal{J}}^h(\widehat{u}_{j+1}^h)_i &\leq \widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i + \big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\big)^T\big(\widehat{u}_{j+1}^h - \widehat{u}_j^h\big) + \frac{L}{2}\|\widehat{u}_{j+1}^h - \widehat{u}_j^h\|^2 \\ &= \widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i + \left(\frac{L\alpha^2}{2} - \alpha\right)\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\|^2 + \frac{L\beta_P^2}{2}\|\widehat{u}_j^h - \widehat{u}_{j-1}^h\|^2 \\ &\quad + \big(\beta_P - L\alpha\beta_P\big)\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\big)^T\big(\widehat{u}_j^h - \widehat{u}_{j-1}^h\big). \end{aligned} \tag{4.3.19}$$

Rearranging (4.3.19) with the expression (4.3.9), Young's inequality, the convexity expression of the $\widehat{\mathcal{J}}^h$

$$\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h), \widehat{u}_{j-1}^h - \widehat{u}_j^h\big) \leq \widehat{\mathcal{J}}^h(\widehat{u}_{j-1}^h) - \widehat{\mathcal{J}}^h(\widehat{u}_j^h), \tag{4.3.20}$$

and the assumption $1 < \alpha L < 2$, we obtain

$$\begin{aligned}&\left(\frac{2\alpha - L\alpha^2}{2}\right)\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\|^2 \\ &\leq \widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i - \widehat{\mathcal{J}}^h(\widehat{u}_{j+1}^h)_i + \frac{L\beta_P^2}{2}\|\widehat{u}_j^h - \widehat{u}_{j-1}^h\|^2 \\ &\quad + \big(\beta_P - L\alpha\beta_P\big)\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\big)^T\big(\widehat{u}_j^h - \widehat{u}_{j-1}^h\big) \\ &\leq \frac{L}{2}\|\widehat{u}_j^h - \widehat{u}^h\|^2 - \frac{\mu}{2}\|\widehat{u}_{j+1}^h - \widehat{u}^h\|^2 + L\beta_P^2\|\widehat{u}_j^h - \widehat{u}^h\|^2 + L\beta_P^2\|\widehat{u}^h - \widehat{u}_{j-1}^h\|^2 \\ &\quad + \big(L\alpha\beta_P - \beta_P\big)\left(\widehat{\mathcal{J}}^h(\widehat{u}_{j-1}^h)_i - \widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\right) \\ &\leq \left(\frac{L}{2} + L\beta_P^2 - \big(L\alpha\beta_P - \beta_P\big)\frac{\mu}{2}\right)\|\widehat{u}_j^h - \widehat{u}^h\|^2 - \frac{\mu}{2}\|\widehat{u}_{j+1}^h - \widehat{u}^h\|^2 \\ &\quad + \left(L\beta_P^2 + \big(L\alpha\beta_P - \beta_P\big)\frac{L}{2}\right)\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2. \end{aligned} \tag{4.3.21}$$

Inserting (4.3.21) into (4.3.18), we get

$$\begin{aligned}T_1 \leq &\left(1 - 2\alpha\mu + \frac{2\alpha}{2 - \alpha L}\left(\frac{L}{2} + L\beta_P^2 - \big(L\alpha\beta_P - \beta_P\big)\frac{\mu}{2}\right)\right)\|\widehat{u}_j^h - \widehat{u}^h\|^2 \\ &- \frac{\alpha\mu}{2 - \alpha L}\|\widehat{u}_{j+1}^h - \widehat{u}^h\|^2 + \frac{2\alpha}{2 - \alpha L}\left(L\beta_P^2 + \big(L\alpha\beta_P - \beta_P\big)\frac{L}{2}\right)\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2. \end{aligned} \tag{4.3.22}$$

Next, using the identity $2(a-b, b-c) = \|a-c\|^2 - \|c-b\|^2 - \|a-b\|^2$, we rearrange

the second expression T_2 as follows

$$\begin{aligned}
T_2 &= 2\beta_P\big(\widehat{u}_j^h - \widehat{u}^h, \widehat{u}_j^h - \widehat{u}_{j-1}^h\big) + 2\alpha\beta_P\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i, \widehat{u}_{j-1}^h - \widehat{u}_j^h\big) \\
&= 2\beta_P\big(\widehat{u}_j^h - \widehat{u}^h, \widehat{u}_j^h - \widehat{u}^h\big) + 2\beta_P\big(\widehat{u}_j^h - \widehat{u}^h, \widehat{u}^h - \widehat{u}_{j-1}^h\big) \\
&\quad +2\alpha\beta_P\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i, \widehat{u}_{j-1}^h - \widehat{u}_j^h\big) \\
&= 2\beta_P\|\widehat{u}_j^h - \widehat{u}^h\|^2 + 2\beta_P\big(\widehat{u}_j^h - \widehat{u}^h, \widehat{u}^h - \widehat{u}_{j-1}^h\big) + 2\alpha\beta_P\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i, \widehat{u}_{j-1}^h - \widehat{u}_j^h\big) \\
&= \beta_P\|\widehat{u}_j^h - \widehat{u}^h\|^2 + \beta_P\|\widehat{u}_j^h - \widehat{u}_{j-1}^h\|^2 - \beta_P\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2 \\
&\quad +2\alpha\beta_P\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i, \widehat{u}_{j-1}^h - \widehat{u}_j^h\big). \qquad (4.3.23)
\end{aligned}$$

Then, by the Young's inequality a bound for the second term in (4.3.23) becomes

$$\begin{aligned}
\beta_P\|\widehat{u}_j^h - \widehat{u}_{j-1}^h\|^2 &= \beta_P\|\widehat{u}_j^h - \widehat{u}^h\|^2 + \beta_P\|\widehat{u}^h - \widehat{u}_{j-1}^h\|^2 + 2\beta_P(\widehat{u}_j^h - \widehat{u}^h, \widehat{u}^h - \widehat{u}_{j-1}^h) \\
&\le \beta_P\|\widehat{u}_j^h - \widehat{u}^h\|^2 + \beta_P\|\widehat{u}^h - \widehat{u}_{j-1}^h\|^2 \\
&\quad + 2\beta_P\left(\frac{\|\widehat{u}_j^h - \widehat{u}^h\|^2}{2} + \frac{\|\widehat{u}^h - \widehat{u}_{j-1}^h\|^2}{2}\right) \\
&= 2\beta_P\|\widehat{u}_j^h - \widehat{u}^h\|^2 + 2\beta_P\|\widehat{u}^h - \widehat{u}_{j-1}^h\|^2. \qquad (4.3.24)
\end{aligned}$$

Moreover, by (4.3.20) and (4.3.9), we have

$$\begin{aligned}
2\alpha\beta_P\big(\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i, \widehat{u}_{j-1}^h - \widehat{u}_j^h\big) &\le 2\alpha\beta_P\big(\widehat{\mathcal{J}}^h(\widehat{u}_{j-1}^h)_i - \widehat{\mathcal{J}}^h(\widehat{u}_j^h)_i\big) \\
&\le \alpha\beta_P L\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2 - \alpha\beta_P\mu\|\widehat{u}_j^h - \widehat{u}^h\|^2. \qquad (4.3.25)
\end{aligned}$$

Inserting (4.3.24) and (4.3.25) into (4.3.23), we obtain

$$T_2 \le \big(\beta_P + \alpha\beta_P L\big)\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2 + \big(3\beta_P - \alpha\beta_P\mu\big)\|\widehat{u}_j^h - \widehat{u}^h\|^2. \qquad (4.3.26)$$

Lastly, the term T_3 in (4.3.17) is bounded by using the identity $2(a,b) \le a^2 + b^2$

$$\begin{aligned}
T_3 &= \beta_P^2\|(\widehat{u}_j^h - \widehat{u}^h) + (\widehat{u}^h - \widehat{u}_{j-1}^h)\|^2 \\
&\le 2\beta_P^2\|\widehat{u}_j^h - \widehat{u}^h\|^2 + 2\beta_P^2\|\widehat{u}^h - \widehat{u}_{j-1}^h\|^2. \qquad (4.3.27)
\end{aligned}$$

Combining the bounds in (4.3.22), (4.3.26), and (4.3.27) in (4.3.17), we have

$$\begin{aligned}
&\left(1 + \frac{\alpha\mu}{2 - \alpha L}\right)\|\widehat{u}_{j+1}^h - \widehat{u}^h\|^2 \\
&\le \left(\frac{2\alpha}{2 - \alpha L}\left(L\beta_P^2 + \big(L\alpha\beta_P - \beta_P\big)\frac{L}{2}\right) + \beta_P + \alpha\beta_P L + 2\beta_P^2\right)\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2 \\
&\quad + \left(1 - 2\alpha\mu + \frac{2\alpha}{2 - \alpha L}\left(\frac{L}{2} + L\beta_P^2 - \big(L\alpha\beta_P - \beta_P\big)\frac{\mu}{2}\right) + 3\beta_P - \alpha\beta_P\mu + 2\beta_P^2\right) \\
&\qquad \times \|\widehat{u}_j^h - \widehat{u}^h\|^2. \qquad (4.3.28)
\end{aligned}$$

Now, taking expectation of (4.3.28) with respect to the sampling history $\mathcal{F}_j$ yields

$$\mathbb{E}\big[\|\widehat{u}_{j+1}^h - \widehat{u}^h\|^2|\mathcal{F}_j\big] \leq a_1\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2 + a_2\|\widehat{u}_{j-1}^h - \widehat{u}^h\|^2, \tag{4.3.29}$$

where

$$a_1 = \frac{6\beta_P - 4\alpha\mu + 4\beta_P^2 - 3L\alpha\beta_P - \alpha\beta_P\mu + 2L\alpha^2\mu + 2}{\alpha\mu - L\alpha + 2},$$
$$a_2 = \frac{2\beta_P + 4\beta_P^2}{\alpha\mu - L\alpha + 2}.$$

Finally, taking one more expectation of (4.3.29) and applying Lemma (4.3.3), we obtain the desired result. □

4.3.2 Stochastic Nesterov's Momentum

We introduce the Nesterov's accelerated gradient method in the stochastic approximation setting as an alternative to the stochastic Polyak's momentum given in Section (4.3.1). Given $\widehat{u}_0^h$ and with $\widehat{u}_{-1}^h = \widehat{u}_0^h$, the Nesterov's momentum reads as the following

$$y_{j+1} = \widehat{u}_j^h + \beta_N\big(\widehat{u}_j^h - \widehat{u}_{j-1}^h\big), \tag{4.3.30a}$$
$$\widehat{u}_{j+1}^h = y_{j+1} - \alpha E_{MC}^{\vec{\omega}}\big[\nabla\widehat{\mathcal{J}}^h(y_{j+1})\big], \tag{4.3.30b}$$

where α is the fixed step-size and β_N is the Nesterov momentum parameter. As opposed to the Polyak's momentum, Nesterov's momentum first makes an advancement in the direction of previous iterations, and then compute the gradient and make a correction.

Setting

$$r_j := y_j - \widehat{u}^h \qquad \text{and} \qquad v_j := \widehat{u}_j^h - \widehat{u}_{j-1}^h,$$

we can transform the update formula (4.3.30) to

$$v_{j+1} = \beta_N v_j - \alpha E_{MC}^{\vec{\omega}}\big[\nabla\widehat{\mathcal{J}}^h(y_{j+1})\big], \tag{4.3.31a}$$
$$r_{j+1} = r_j + \beta_N^2 v_{j-1} - \alpha(1+\beta_N)E_{MC}^{\vec{\omega}}\big[\nabla\widehat{\mathcal{J}}^h(y_j)\big], \tag{4.3.31b}$$

or equivalently

$$\begin{bmatrix} r_{j+1} \\ v_j \end{bmatrix} = \begin{bmatrix} I & \beta_N^2 I \\ 0 & \beta_N I \end{bmatrix}\begin{bmatrix} r_j \\ v_{j-1} \end{bmatrix} - \alpha\begin{bmatrix} (1+\beta_N)I \\ I \end{bmatrix} E_{MC}^{\vec{\omega}}\big[\nabla\widehat{\mathcal{J}}^h(y_j)\big]. \tag{4.3.32}$$

We note that $r_1 = \widehat{u}_0^h - \widehat{u}^h$ and $v_0 = 0$ provided that $\widehat{u}_{-1}^h = \widehat{u}_0^h$. Moreover, by the following expression for the twice continuously differentiable functions $\widehat{\mathcal{J}}^h$

$$\nabla \widehat{\mathcal{J}}^h(t) = \nabla \widehat{\mathcal{J}}^h(x) + \left(\int_0^1 \nabla^2 \widehat{\mathcal{J}}^h(x + s(t-x)) ds \right) (t-x),$$

the approximated gradient can be written as follows

$$\nabla \widehat{\mathcal{J}}^h(y_j) = \widetilde{H}_j r_j + z_j, \tag{4.3.33}$$

where

$$\widetilde{H}_j = E_{MC}^{\overrightarrow{\omega}} \left[\int_0^1 \nabla^2 \widehat{\mathcal{J}}^h(\widehat{u}^h + r_j\, s)\, ds \right], \quad \text{and} \quad z_j = E_{MC}^{\overrightarrow{\omega}} \big[\nabla \widehat{\mathcal{J}}^h(\widehat{u}^h) \big].$$

Inserting (4.3.33) into (4.3.32) and unravelling the iteration we obtain

$$\begin{bmatrix} r_{j+1} \\ v_j \end{bmatrix} = A_j \cdots A_1 \begin{bmatrix} r_1 \\ v_0 \end{bmatrix} - \alpha \begin{bmatrix} (1+\beta_N) I \\ I \end{bmatrix} z_j - \alpha \sum_{k=1}^{j-1} (A_j \cdots A_{k+1}) \begin{bmatrix} (1+\beta_N) I \\ I \end{bmatrix} z_k, \tag{4.3.34}$$

where

$$A_j = \begin{bmatrix} I - \alpha(1+\beta_N)\widetilde{H}_j & \beta_N^2 I \\ -\alpha \widetilde{H}_j & \beta_N I \end{bmatrix}. \tag{4.3.35}$$

Before stating the convergence result of Nesterov's momentum in the stochastic approximation setting, we give a bound for the spectral norm of A_j.

Lemma 4.3.5. *[5, Lemma 3] The spectral norm of A_j defined in (4.3.35) is bounded by*

$$\|A_j\| \leq \max_{\lambda \in [\mu, L]} R_\lambda(\alpha, \beta_N) = R(\alpha, \beta_N), \tag{4.3.36}$$

where

$$\begin{aligned} C_\lambda(\alpha, \beta_N) &= \big(1 - \alpha(1+\beta_N)\lambda\big)^2 + \alpha^2\lambda^2 + \beta_N^2(1+\beta_N^2), \\ \widetilde{\Delta}_\lambda(\alpha, \beta_N) &= C_\lambda(\alpha, \beta_N)^2 - 4\beta_N^2(1-\alpha\lambda)^2, \\ R_\lambda(\alpha, \beta_N) &= \frac{1}{2}\sqrt{C_\lambda(\alpha, \beta_N) + \sqrt{\widetilde{\Delta}_\lambda(\alpha, \beta_N)}}. \end{aligned}$$

Proof. Since our cost functional $\widehat{\mathcal{J}}$ is $L-$smooth and $\mu-$strongly convex, eigenvalues λ of $\tilde{H}_j$ lie in $[\mu, L]$. By the original work of Polyak [61], there exists an eigenvalue λ such that $\|A_j\|$ is equal to the spectral norm of

$$B(\lambda) = \begin{bmatrix} 1-\alpha(1+\beta)\lambda & \beta^2 \\ -\alpha\lambda & \beta \end{bmatrix}. \tag{4.3.37}$$

Calculating $\|B(\lambda)\|^2 = B(\lambda)^\top B(\lambda)$ we find the characteristic polynomial

$$z^2 - C_\lambda(\alpha,\beta)z + \beta^2(1-\alpha\lambda)^2 = 0.$$

The largest root corresponds to $R_\lambda(\alpha,\beta)^2 = \|B(\lambda)\|^2$. Taking the squared-root gives us the desired result. □

Now, we provide the convergence result of the Nesterov's momentum (4.3.30) in the discrete MC setting by following the analysis done in [5].

Theorem 4.3.6. *Assume that for a given the step length α and Nesterov's momentum parameter β_N, we have*

$$R(\alpha,\beta_N) = \max_{\lambda\in[\mu,L]} R_\lambda(\alpha,\beta_N) < 1.$$

If we run the Nesterov's momentum (4.3.30) in the discrete MC setting (that is, finite–sum setting), then for all $j \geq 0$ it holds

$$\mathbb{E}\Big[\|y_{j+1} - \widehat{u}^h\|\Big] \leq R(\alpha,\beta_N)^j\|y_1 - \widehat{u}^h\| + \frac{\alpha\sqrt{(1+\beta_N)^2+1}}{1-R(\alpha,\beta_N)}\mathbb{E}\big[\|\nabla\widehat{\mathcal{J}}^h(\widehat{u}_j^h)\|\big]. \tag{4.3.38}$$

Proof. By the submultiplicativity of matrix norms

$$\|A_j\cdots A_{k+1}\| \leq \prod_{l=k+1}^{j}\|A_l\|$$

and Lemma (4.3.5), we have

$$\|A_j\cdots A_{k+1}\| \leq \prod_{l=k+1}^{j}\|A_l\| \leq R(\alpha,\beta)^{j-k}. \tag{4.3.39}$$

An application of the triangle inequality on (4.3.34) with the bound (4.3.39) yields

$$\left\|\begin{bmatrix} r_{j+1} \\ v_j \end{bmatrix}\right\| \leq R(\alpha,\beta_N)^j\left\|\begin{bmatrix} r_1 \\ v_0 \end{bmatrix}\right\| + \alpha\sqrt{(1+\beta_N)^2+1}\sum_{k=1}^{j}R(\alpha,\beta_N)^{j-k}\|z_j\|. \tag{4.3.40}$$

Taking expectation of both sides in (4.3.40) with respect to the sampling history $\mathcal{F}_j$ and using the unbiasedness property of the MC estimator $E_{MC}^{\overrightarrow{\omega}}[\cdot]$, we obtain

$$\begin{aligned}\mathbb{E}\Big[\|y_{j+1} - \widehat{u}^h\|^2\Big] &\leq \mathbb{E}\Big[\| \begin{bmatrix} r_{j+1} \\ v_j \end{bmatrix} \|\Big] \\ &\leq R(\alpha, \beta_N)^j \|y_1 - \widehat{u}^h\|^2 + \frac{\alpha\sqrt{(1+\beta_N)^2+1}}{1 - R(\alpha, \beta_N)} \mathbb{E}\big[\|\nabla \widehat{\mathcal{J}}^h(\widehat{u}_j^h)\|\big],\end{aligned}$$

which is the desired result. □

We note that $\mathbb{E}\big[\|\nabla \widehat{\mathcal{J}}^h(\widehat{u}_j^h)\|\big] = 0$ if the minimizer $\widehat{u}^h$ of $\widehat{\mathcal{J}}^h(\widehat{u}^h)$ defined in (4.2.2) also minimizes each $\widehat{\mathcal{J}}^h(\widehat{u}^h)_i$ for all $i = 1, \ldots, N$, called as interpolation condition.

Next, combining Theorems (4.1.1), (4.2.4), (4.3.4), and (4.3.6), an error bound for the approximate solution $\widehat{u}_j^h$ defined in (4.3.10) or in (4.3.30) in terms of the iteration number j, the mesh size h, and the sample size N is obtained.

Theorem 4.3.7. *Let u, $\widehat{u}_{j,P}^h$, and $\widehat{u}_{j,N}^h$ be the solutions of (3.1.13), (4.3.10), and (4.3.30), respectively. Assume that the interpolation condition holds and the conditions of Theorems (4.1.1), (4.2.4), (4.3.4), and (4.3.6) are satisfied. Then, the following global error estimates hold*

$$\mathbb{E}[\|\widehat{u}_{j,P}^h - u\|^2] \leq C\left(q^j + \frac{1}{N} + h^4\right), \tag{4.3.41a}$$

and

$$\mathbb{E}[\|\widehat{u}_{j,N}^h - u\|^2] \leq C\left(R(\alpha, \beta_N)^{2j} + \frac{1}{N} + h^4\right). \tag{4.3.41b}$$

Lastly, we discuss the complexity of the fully discretized problem based on the stochastic momentum methods in order to obtain $\mathbb{E}[\|\widehat{u}_{j,\cdot}^h - u\|^2] \leq C\epsilon^2$, i.e., $\mathcal{O}(\epsilon)$.

Corollary 4.3.8. *To achieve $\mathcal{O}(\epsilon)$ by applying stochastic Polyak's momentum (4.3.10) and stochastic Nesterov's momentum (4.3.30), the computational works are bounded by*

$$W_P \approx \epsilon^{-2-\frac{1}{2}} |\log(\epsilon^2)|, \qquad \textit{and} \qquad W_N \approx \epsilon^{-2-\frac{1}{2}} |\log(\epsilon)|, \tag{4.3.42}$$

respectively.

Proof. Assuming that the error is equidistributed over the terms in Theorem (4.3.7), we directly obtain

$$N \approx \epsilon^{-2}, \qquad \text{and} \qquad h \approx \epsilon^{1/2}. \tag{4.3.43}$$

Next, by Theorem (4.3.4) with $R_P = \|\widehat{u}^h_{0,P} - \widehat{u}^h\|$, we have

$$q^j(1+\delta)R_P^2 \leq \epsilon^2 \quad \longrightarrow \quad \frac{(1+\delta)R_P^2}{\epsilon^2} \leq \left(\frac{1}{q}\right)^j \quad \longrightarrow \quad \frac{\log\left(\frac{(1+\delta)R_P^2}{\epsilon^2}\right)}{\log\left(\frac{1}{q}\right)} \leq j. \tag{4.3.44}$$

In the similar way, by Theorem (4.3.6) with $C_N = \frac{\alpha\sqrt{(1+\beta_N)^2+1}}{1-R(\alpha,\beta_N)} E^{\vec{\omega}}_{MC}\left[\|\nabla \widehat{J}^h(\widehat{u}^h_j)\|\right]$ and $R_N = \|y_1 - \widehat{u}^h\|$, we compute

$$\frac{\log\left(\frac{(1-R(\alpha,\beta_N))R_N}{\epsilon(1-R(\alpha,\beta_N))-C_N}\right)}{2\log\left(\frac{1}{R(\alpha,\beta_N)}\right)} \leq j. \tag{4.3.45}$$

Combining (4.3.44) and (4.3.45) with (4.3.43), the desired results are obtained. □

CHAPTER 5

NUMERICAL EXPERIMENTS

In this chapter, we numerically study different examples of robust deterministic optimal control problem subject to convection-diffusion equation with uncertain data. We solve two different cases where only diffusion coefficient is random, only convection coefficient is random. We use standard continuous FEM paired with the MC method to solve the PDEs with uncertainty. In optimization, Polyak's and Nesterov's momentum methods are used. In the numerical experiments, we investigate the behavior of solution for various data sets such as the momentum parameter β_P or β_N, the step length α, and the mini-batch size N.

In each case, we consider a random field having a covariance function of the form

$$\mathcal{C}_\eta(\boldsymbol{x},\boldsymbol{y}) = \kappa^2 \prod_{n=1}^{2} e^{-|x_n-y_n|/\ell_n} \quad \forall(\boldsymbol{x},\boldsymbol{y}) \in D, \tag{5.0.1}$$

where we denote the correlation length by ℓ_n, and the standard deviation by κ. In the KL expansion, the eigenpairs of the covariance function (5.0.1) are computed explicitly as done in [Chapter 7 p. 300, [51]]. Further, we present our algorithm of stochastic gradient with momentum for the solution of optimal control problems constrained by PDEs with uncertain data in Algorithm 1. We note that the state equation (3.1.3) has nonhomogeneous Dirichlet boundary conditions in the benchmark examples.

5.1 Randomness in Diffusion Parameter

Our first problem is a robust deterministic optimal control problem containing random diffusion coefficient on the domain $D = [-1,1]^2$. We take the source function

Algorithm 1 Stochastic Gradient with Momentum Terms

Given a step-size α, tolerance ϵ_{tol}, maximum iteration $\max_{iter}$, and mini-batch size $\overline{N}$.
Set $u \leftarrow 0$ and $k \leftarrow 0$.
for $k \leq \max_{iter}$ **do**
  $\widetilde{p} \leftarrow 0$.
  for $i = 1, \ldots, \overline{N}$ **do**
    Generate KL expansion of the random field η_i.
    Solve the state equation $\rightarrow y(a_i, u)$.
    Solve the adjoint equation $\rightarrow p(a_i, u)$.
    Update $\widetilde{p} = \widetilde{p} + p(a_i, u)/\overline{N}$.
  end for
  Compute the gradient $\nabla \mathcal{J} = \mu u + \widetilde{p}$.
  if $\|\nabla \mathcal{J}\| \leq \epsilon_{tol}$ **then**
    STOP
  end if
  Apply Polyak's momentum (4.3.10) or Nesterov momentum (4.3.30).
  $k \leftarrow k + 1$.
end for

$f(\boldsymbol{x}) = 0$. The random diffusion coefficient is taken as $a(\boldsymbol{x}, \omega) = \eta(\boldsymbol{x}, \omega)$ where $\eta(\boldsymbol{x}, \omega)$ is a uniform random field with a unity mean and has the covariance function (5.0.1). The convection term is taken as $\boldsymbol{b} = (0, 1)^\top$. The boundary condition is the following nonhomogeneous Dirichlet boundary condition

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

As the desired state y^d, we take the solution of the same problem as the constraint equation, only with $u(\boldsymbol{x}) = 0$. We represent the random coefficient by the KL expansion with the correlation length $\ell_1 = \ell_2 = 1$, and the standard deviation $\kappa = 0.05$. We truncate the KL expansion at $N_{KL} = 41$ so that we cover the 99% of the sum of the eigenvalues of the covariance function. The $i.i.d.$ random variables ξ are chosen from the uniform distribution in $[-1, 1]$.

There is no known analytic solution to our problem; hence, we solve the problem by the gradient descent method to generate a reference solution. We take the MC approximation size $N = 5000$, the step size $\alpha = 10$, and the spatial mesh size $h = 2^{-7}$ for $\mu = 10^{-2}$ and $\gamma = 0$ (resp, $\gamma = 1$). The stopping criteria is chosen as $\nabla \widehat{\mathcal{J}}^h(\widehat{u}^h) \leq 10^{-8}$, and it is met after 43 (resp. 47) iterations for $\gamma = 0$ (resp, $\gamma = 1$). Mean of the solution $\mathbb{E}[y^*]$ and the corresponding control u^* are taken as reference solutions and depicted in Figure 5.1. We find that the value of cost function is $\mathcal{J}^* = 2.3214e-5$ (resp, $\mathcal{J}^* = 4.6433e-5$).

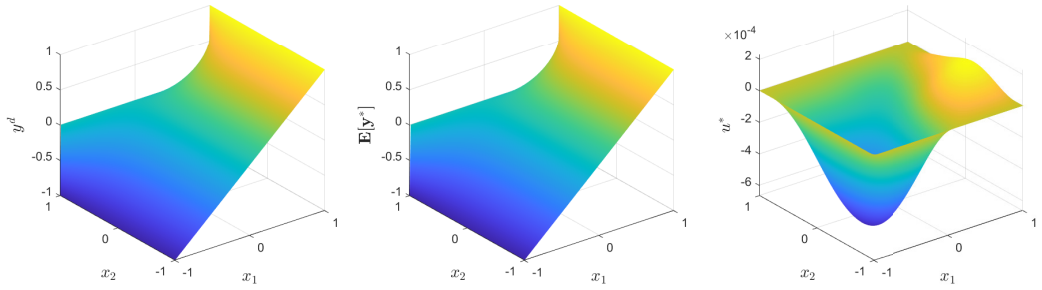

Figure 5.1: Example 5.1:The desired state y^d (left), the optimal state $\mathbb{E}[y^*]$ (middle), and the optimal control u^* computed by the gradient descent method with $N = 5000$, $h = 2^{-7}$, $\mu = 10^{-2}$, and $\gamma = 0$.

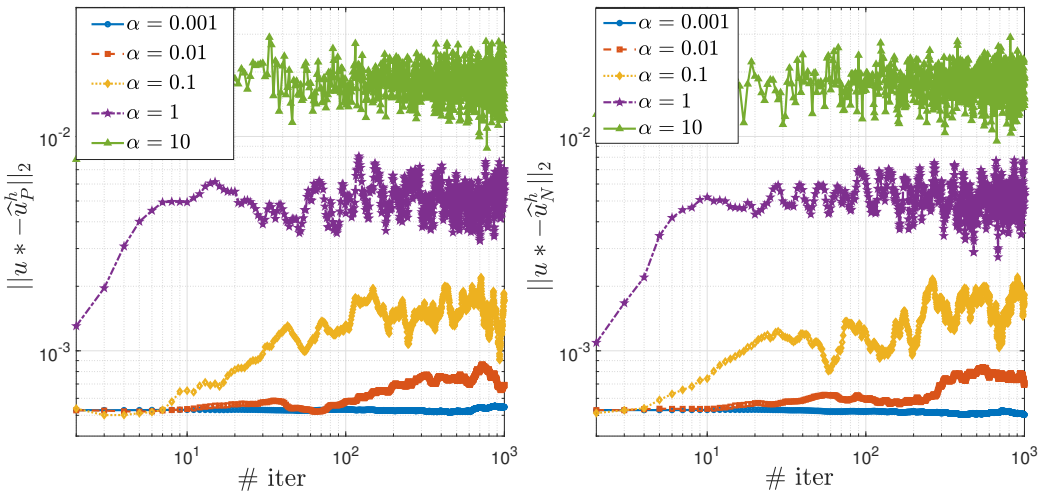

Figure 5.2: Example 5.1: Convergence results of the stochastic Polyak's momentum (left) and stochastic Nesterov momentum (right) with $\overline{N} = 1$, $h = 2^{-7}$, $\mu = 10^{-2}$, $\gamma = 0$, and $\beta_P = \beta_N = 0.6$ for several values of the step size $\alpha = \alpha_0$. All results are averaged over 10 independent runs.

We first run the Algorithm 1 for 10 independent realizations up to 1000 iterations,

using only one sample $\overline{N} = 1$, a fixed step size $\alpha = \alpha_0$ for the momentum parameters $\beta_P = \beta_N = 0.6$, and we take the average of those 10 independent realizations. It is observed in the Figure 5.2 that the methods do not converge when the step size is taken fixed; therefore, we use a diminishing step size $\alpha_j = \alpha_0/j$ with $\alpha_0 = 10$ as discussed in [29, 54]. In Figure 5.3, we give the convergence of stochastic momentum methods with different momentum parameters for both $\gamma = 0$, and $\gamma = 1$. We note that, addition of momentum term improves the convergence substantially after a few iterations. Further, we observe that the larger values of momentum parameter tend to perform better in terms of convergence of the control u. The computed values of the objective function $\widehat{\mathcal{J}}^h$ and the relative error $\frac{|\mathcal{J}^* - \widehat{\mathcal{J}}^h|}{\mathcal{J}^*}$ for both momentum variants are given in Table 5.1, and Table 5.2, respectively.

Table 5.1: Example 5.1: Values of the objective function $\widehat{\mathcal{J}}_P^h$ and the relative error of the objective function $\frac{|\mathcal{J}^* - \widehat{\mathcal{J}}_P^h|}{\mathcal{J}^*}$ obtained by the stochastic Polyak's momentum with $\gamma = 0, 1$. All results are averaged over 10 independent runs.

	$\gamma = 0$				$\gamma = 1$			
β_P	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
$\widehat{\mathcal{J}}_P^h$	3.15e-5	2.75e-5	2.41e-5	1.59e-5	2.28e-5	2.46e-5	2.73e-5	3.36e-5
$\frac{\lvert\mathcal{J}^* - \widehat{\mathcal{J}}_P^h\rvert}{\mathcal{J}^*}$	3.56e-1	1.84e-1	3.88e-2	3.16e-1	5.08e-1	4.70e-1	4.13e-1	2.77e-1

Table 5.2: Example 5.1: Values of the objective function $\widehat{\mathcal{J}}_N^h$ and the relative error of the objective function $\frac{|\mathcal{J}^* - \widehat{\mathcal{J}}_N^h|}{\mathcal{J}^*}$ obtained by the stochastic Nesterov momentum with $\gamma = 0, 1$. All results are averaged over 10 independent runs.

	$\gamma = 0$				$\gamma = 1$			
β_N	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
$\widehat{\mathcal{J}}_N^h$	2.85e-5	3.39e-5	1.75e-5	2.99e-5	2.45e-5	2.68e-5	3.19e-5	2.27e-5
$\frac{\lvert\mathcal{J}^* - \widehat{\mathcal{J}}_N^h\rvert}{\mathcal{J}^*}$	2.26e-1	4.60e-1	2.47e-1	2.90e-1	4.73e-1	4.22e-1	3.14e-1	5.12e-1

Now, in Figure 5.4 we study the convergence behavior of the control u for different α_0 values, where we take the momentum values $\beta_P = \beta_N = 0.6$ and the risk-aversion parameter $\gamma = 0$. We observe that the larger values of α_0 produce better performance in terms of acceleration. Finally, we investigate the effect of different mini-batch sizes $\overline{N}$ in Figure 5.5 and conclude that the larger mini-batch values do not necessarily yield better results.

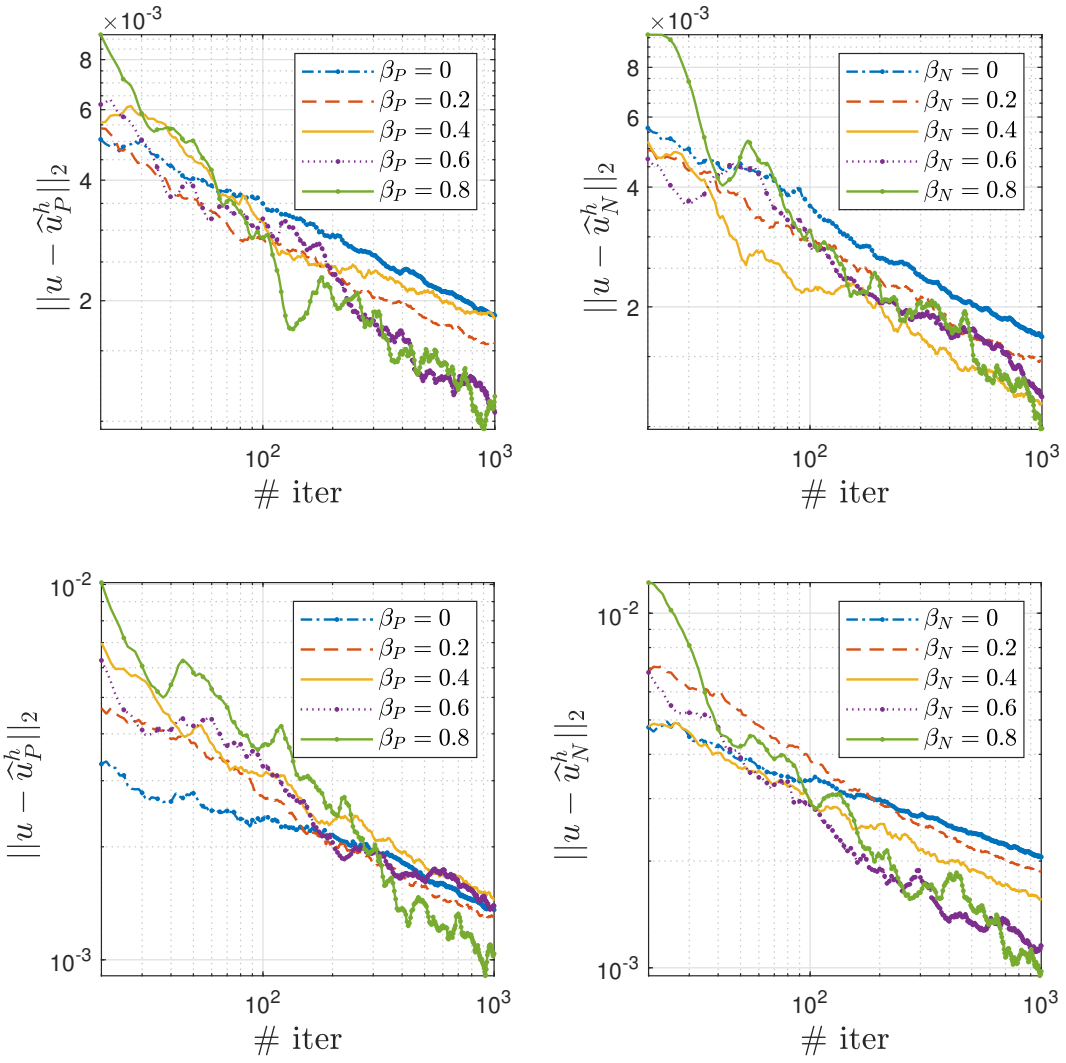

Figure 5.3: Example 5.1: Convergence results of the stochastic Polyak's momentum (left) and stochastic Nesterov momentum (right) with various values of momentum parameters and risk–aversion parameters $\gamma = 0$ (top) and $\gamma = 1$ (bottom). All results are averaged over 10 independent runs.

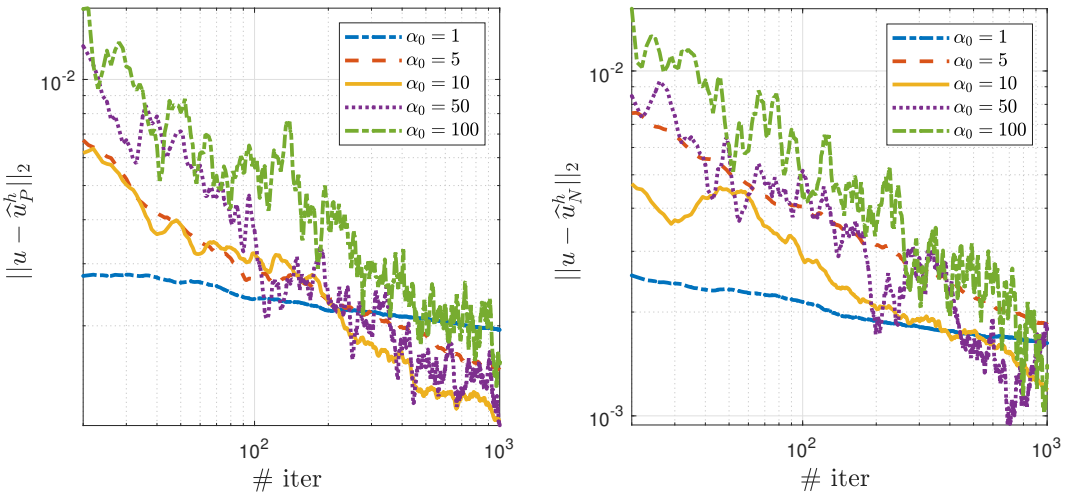

Figure 5.4: Example 5.1: Convergence results of the stochastic Polyak's momentum (left) and stochastic Nesterov momentum (right) with $\beta_P = \beta_N = 0.6$, $\gamma = 0$, and $\overline{N} = 1$ for several values of α_0. All results are averaged over 10 independent runs.

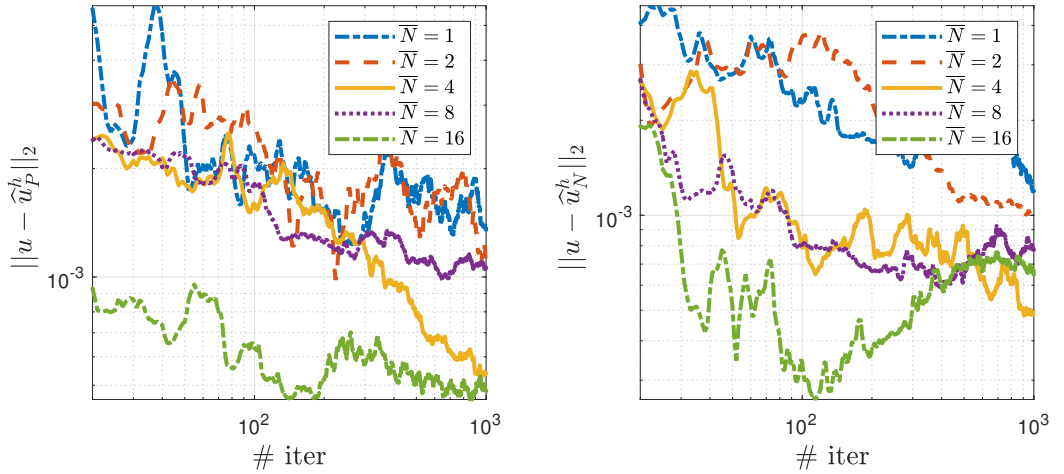

Figure 5.5: Example 5.1: Convergence results of stochastic Polyak's momentum (left) and stochastic Nesterov momentum (right) with $\beta_P = \beta_N = 0.6$, $\gamma = 0$, and $\alpha_0 = 10$ for several values of the mini–batch size $\overline{N}$. All results are averaged over 1 run.

5.2 Randomness in Convection Parameter

As a second example we choose our problem to be on the spatial domain $D = [0, 1]^2$ with source term $f(\boldsymbol{x}) = 0$. The diffusion coefficient is deterministic and taken as $a(\boldsymbol{x}, \omega) = 0.1$, while the convection coefficient is random in the constraint equation

such that

$$\mathbf{b}(\boldsymbol{x},\omega) := \left(\cos\left(\frac{1}{5}\eta(\boldsymbol{x},\omega)\right), \sin\left(\frac{1}{5}\eta(\boldsymbol{x},\omega)\right)\right)^T, \tag{5.2.1}$$

where $\eta(\boldsymbol{x},\omega)$ is a uniform field with zero mean and covariance function (5.0.1). To represent $\mathbf{b}(\boldsymbol{x},\omega)$ by the KL expansion with the standard deviation $\kappa = 0.05$ and correlation length as $\ell_1 = \ell_2 = 0.5$, we use $N_{KL} = 85$ to account for the 99% of the sum of eigenvalues of (5.0.1). The $i.i.d.$ random variables ξ are chosen from the uniform distribution in $[-1, 1]$ as the previous example. The boundary conditions are nonhomogenous Dirichlet such that

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

where $S \subset \partial D$ is defined as

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

Similar to the previous example, we choose the desired state y^d as the solution of the same problem as the constraint equation, only with $u(\boldsymbol{x}) = 0$. Again, there is no analytic solution, so we use the gradient descent method to obtain a reference solution for the case $\mu = 10^{-2}$, and $\gamma = 0$ (resp, $\gamma = 1$). In the gradient descent simulations, we take a fixed step size $\alpha = 10$ and the MC sample size $N = 5000$ with the spatial mesh size $h = 2^{-7}$. The stopping criteria $\nabla \widehat{\mathcal{J}}^h(\widehat{u}^h) \leq 10^{-8}$ is satisfied at 3rd iteration for both $\gamma = 0, 1$. We find that the cost functional value at the optimal control is $\mathcal{J}^* = 1.0505e-6$ (resp. $\mathcal{J}^* = 2.1880e-6$). In Figure 5.7, we display the desired state y^d, the optimal state reference solution $\mathbb{E}[y^*]$, and the optimal control reference solution u^*.

We run the Algorithm 1 for 10 independent realizations all up to 2500 iterations, with $\overline{N} = 1$ and step size as $\alpha = \alpha_0/j$ with $\alpha_0 = 10$. In Figure 5.6 we show the convergence for different momentum parameters β_P, β_N and risk-aversion parameters $\gamma = 0, 1$. We observe the similar tendency to the previous Example 5.1 that the larger momentum parameters yield better convergence behavior. In Table 5.3 and Table 5.4 we present the optimal cost functional values $\widehat{\mathcal{J}}^h$ and the relative error of the objective function $\frac{|\mathcal{J}^* - \widehat{\mathcal{J}}^h|}{\mathcal{J}^*}$ for the stochastic Polyak's momentum and stochastic Nesterov's momentum, respectively.

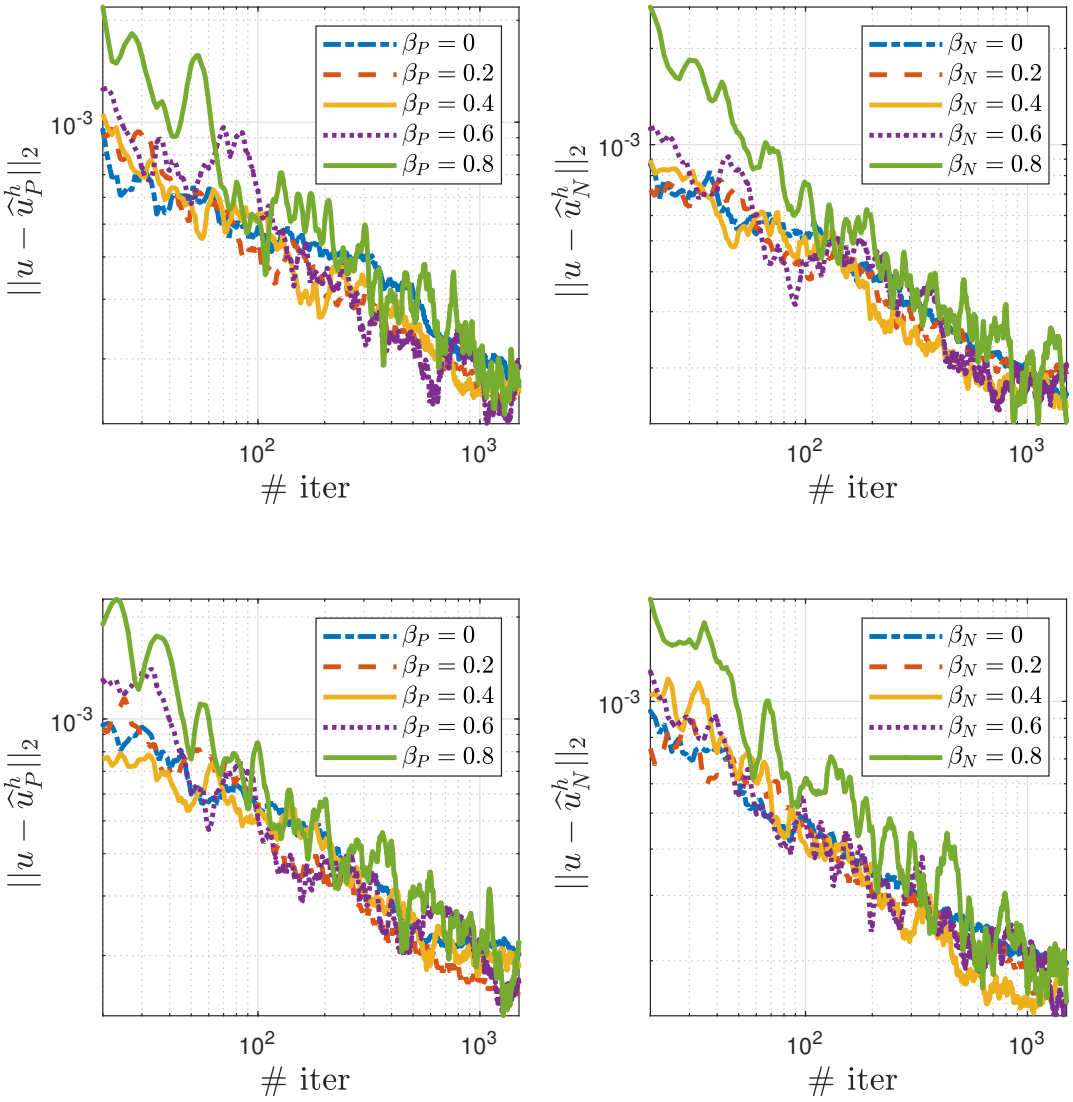

Figure 5.6: Example 5.2: Convergence behavior of the stochastic Polyak's momentum (left) and stochastic Nesterov momentum (right) with various values of momentum parameters and risk–aversion parameters $\gamma = 0$ (top) and $\gamma = 1$ (bottom). All simulations are averaged over 10 trials.

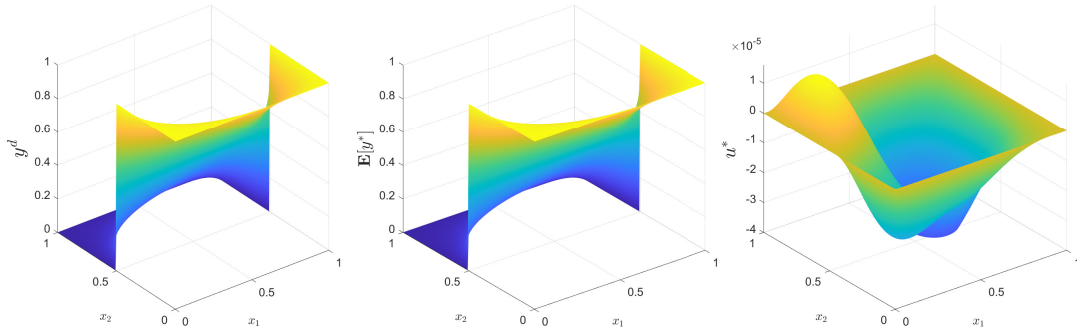

Figure 5.7: Example 5.2: The desired state y^d(left), the optimal state $\mathbb{E}[y^*]$ (middle), and the optimal control u^* computed by the gradient descent with $N = 1000$, $h = 2^{-7}$, $\mu = 10^{-2}$, and $\gamma = 0$.

Table 5.3: Example 5.2: Computed values of the objective function $\widehat{\mathcal{J}}_P^h$ and the relative error of the objective function $\frac{|\mathcal{J}^* - \widehat{\mathcal{J}}_P^h|}{\mathcal{J}^*}$ obtained by the stochastic Polyak's momentum with $\gamma = 0, 1$. All results are averaged over 10 independent runs.

	$\gamma = 0$				$\gamma = 1$			
β_P	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
$\widehat{\mathcal{J}}_P^h$	2.02e-6	1.29e-6	1.10e-6	1.03e-6	1.07e-6	1.09e-6	1.13e-6	1.08e-6
$\frac{\lvert\mathcal{J}^* - \widehat{\mathcal{J}}_P^h\rvert}{\mathcal{J}^*}$	9.19e-1	2.31e-1	5.02e-2	2.12e-2	5.13e-1	5.00e-1	4.83e-1	5.05e-1

Table 5.4: Example 5.2: Values of the objective function $\widehat{\mathcal{J}}_N^h$ and the relative error of the objective function $\frac{|\mathcal{J}^* - \widehat{\mathcal{J}}_N^h|}{\mathcal{J}^*}$ obtained by the stochastic Nesterov momentum with $\gamma = 0, 1$. All results are averaged over 10 independent runs.

	$\gamma = 0$				$\gamma = 1$			
β_N	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
$\widehat{\mathcal{J}}_N^h$	1.08e-6	4.97e-7	2.16e-6	1.04e-6	1.21e-6	8.49e-6	1.39e-6	7.10e-7
$\frac{\lvert\mathcal{J}^* - \widehat{\mathcal{J}}_N^h\rvert}{\mathcal{J}^*}$	3.14e-2	5.27e-1	1.06e-0	1.00e-2	4.45e-1	6.12e-1	3.66e-1	6.76e-1

As a final investigation for this example, we study the convergence behavior for different α_0 and $\overline{N}$ values in Figure 5.8 and Figure 5.9, respectively. In Figure 5.9 we see that the larger mini-batches do not necessarily improve convergence as the previous example.

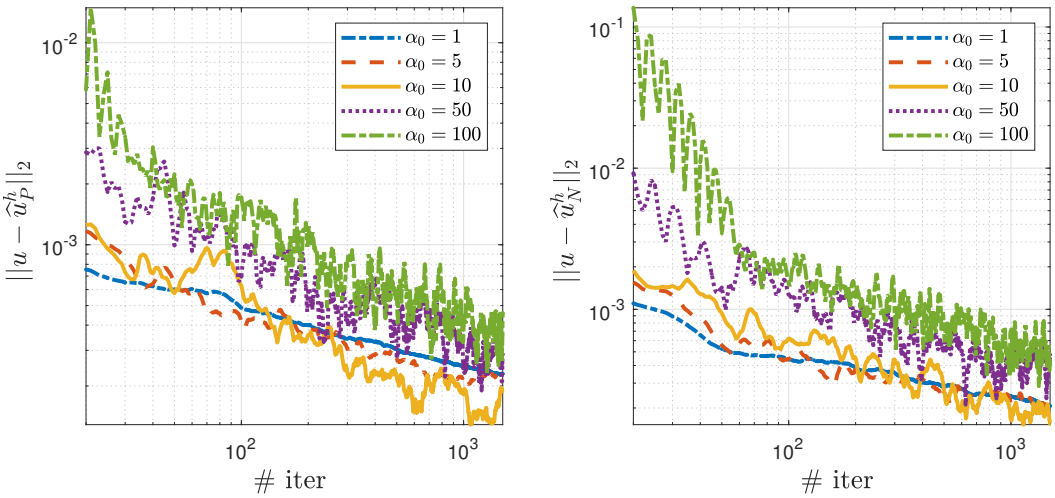

Figure 5.8: Example 5.2: Convergence behavior of the stochastic Polyak's momentum with $\beta_P = 0.6$, $\gamma = 0$, and $\overline{N} = 1$ (left) and stochastic Nesterov momentum (right) with $\beta_N = 0.8$, $\gamma = 1$, and $\overline{N} = 1$ for several values of α_0. All results are averaged over 10 independent runs.

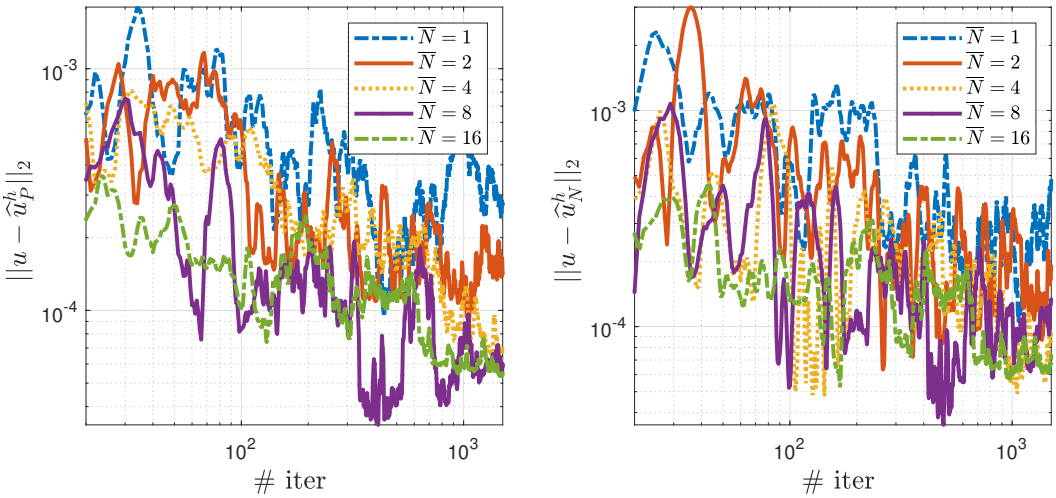

Figure 5.9: Example 5.2: Convergence behavior of stochastic Polyak's momentum with $\beta_P = 0.6$, $\gamma = 0$, and $\alpha_0 = 10$ (left) and stochastic Nesterov momentum (right) with $\beta_N = 0.8$, $\gamma = 1$, and $\alpha_0 = 10$ for several values of the mini–batch size $\overline{N}$. All results are averaged over 1 run.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, we have studied the numerical methods to solve robust deterministic optimal control problems constrained by convection-diffusion equations with uncertain coefficients. In solving the problem, we used the Monte Carlo method in probability space, the standard continuous Finite Element Method in the spatial domain; and in optimization, we used the stochastic gradient method with Polyak's and Nesterov's momentums added, separately.

It is seen in the numerical results that the addition of momentum terms is beneficial in all examples. Also, we observe that the fixed step size α does not meet expectations, so we use a diminishing step size. For our choice of initial step size, it is clear that larger momentum parameters result in better convergence results. Further, we see that increasing the mini-batch size does not necessarily result in accelerated convergence.

In future work, since the computational cost posed by the MC method is significant, employing the Multi-Level Monte Carlo method [2] can make a good improvement. In the spatial domain, it is known that for small values of the diffusion coefficient, the convection-diffusion equation becomes unstable, and so-called "boundary layers" are formed. We haven't gone to that domain in this thesis yet, for future work, we may employ a stabilization method such as "Streamline Upwind Petrov-Galerkin method" (SUPG)[19], or we may resort to Discontinuous Galerkin method [27, 47]. Also, recent optimization theory shows that there are alternative stochastic gradient with momentum methods such as "Katyusha momentum" [3] and "accelerated proximal stochastic variance reduced gradient (ASVRG)" [65] which we can employ.

REFERENCES

[1] G. S. A. Alexanderian, N. Petra and O. Ghattas, Mean-variance risk-averse optimal control of systems governed by PDEs with random parameter fields using quadratic approximations, SIAM/ASA Journal of Uncertainty Quantification, 5(1), pp. 1166–1192, 2017.

[2] 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.

[3] Z. Allen-Zhu, Katyusha: The first direct acceleration of stochastic gradient methods, The Journal of Machine Learning Research, 18(1), p. 8194–8244, 2017.

[4] W. Alt and U. Mackenroth, Convergence of finite element approximations to state constrained convex parabolic boundary control problems, SIAM Journal on Control and Optimization, 27(4), pp. 718–736, 1989.

[5] M. Assran and M. Rabbat, On the convergence of Nesterov's accelerated gradient method in stochastic settings, in H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 410–420, PMLR, 2020.

[6] I. Babuška, F. Nobile, and R. Tempone, A stochastic collocation method for elliptic partial differential equations with random input data, SIAM Journal on Numerical Analysis, 45(3), pp. 1005–1034, 2007.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] A. Borzì, Multigrid and sparse-grid schemes for elliptic control problems with random coefficients, Computing and Visualization in Science, 13, pp. 153–160, 2010.

[12] A. Borzì, V. Schulz, C. Schillings, and G. von Winckel, On the treatment of distributed uncertainties in PDE constrained optimization, Gesellschaft für Angewandte Mathematik und Mechanik Mitteilungen, 33(2), pp. 230–246, 2010.

[13] 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.

[14] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning, SIAM Review, 60(2), pp. 223–311, 2018.

[15] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin, second edition, 2002.

[16] S. Bubeck, Convex optimization: Algorithms and complexity, Foundations and Trends in Machine Learning, 8(3-4), p. 231–357, 2015.

[17] P. Chen and A. Quarteroni, Weighted reduced basis method for stochastic optimal control problems with elliptic PDE constraint, SIAM/ASA Journal on Uncertainty Quantification, 2(1), pp. 364–396, 2014.

[18] 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.

[19] 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.

[20] A. d'Aspremont, Smooth optimization with approximate gradient, SIAM Journal on Optimization, 19(3), pp. 1171–1183, 2008.

[21] M. Deb, I. Babuška, and J. Oden, Solution of stochastic partial differential equations using Galerkin finite element techniques, Computer Methods in Applied Mechanics and Engineering, 190, pp. 6359–6372, 2001.

[22] O. Devolder, F. Glineur, and Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, Mathematical Programming, 146, p. 37–75, 2013.

[23] C. R. Dietrich and G. N. Newsam, Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix, SIAM Journal on Scientific Computing, 18, pp. 1088–1107, 1997.

[24] N. Flammarion and F. Bach, From averaging to acceleration there is only a step-size, Conference on Learning Theory, pp. 658–695, 2015.

[25] P. Frauenfelder, C. Schwab, and R. A. Todor, Finite elements for elliptic problems with stochastic coefficients, Computer Methods in Applied Mechanics and Engineering, 194(2-5), pp. 205–228, 2005.

[26] 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.

[27] 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(7), pp. 763–797, 2019.

[28] C. Geiersbach and G. Pflug, Projected stochastic gradients for convex constrained problems in Hilbert spaces, SIAM Journal of Optimization, 29, pp. 2079–2099, 2019.

[29] C. Geiersbach and W. Wollner, A stochastic gradient method with mesh refinement for PDE–constrained optimization under uncertainity, SIAM Journal on Scientific Computing, 42(5), pp. A2750–A2772, 2020.

[30] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, Global convergence of the heavy-ball method for convex optimization, IEEE European Control Conference (ECC), pp. 310–315, 2015.

[31] M. B. Giles, Multilevel Monte Carlo methods, Acta Numerica, 24, p. 259–328, 2015.

[32] I. G. Graham, F. Y. Kuo, D. Nuyens, R. Scheichl, and I. H. Sloan, Quasi–Monte Carlo methods for elliptic PDEs with random coefficients and applications, Journal of Computational Physics, 230, pp. 3668–3694, 2011.

[33] 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.

[34] E. Haber, M. Chung, and F. Herrmann, An effective method for parameter estimation with PDE constraints with multiple right–hand side, SIAM Journal of Optimization, 22, pp. 739–757, 2012.

[35] 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.

[36] J. Jacod and P. Protter, *Probability essentials*, Universitext, Springer-Verlag, Berlin, second edition, 2003.

[37] H. Jarchow, *Locally convex spaces*, B.G. Teubner, Stuttgart, 1981.

[38] A. Keese, A review of recent developments in the numerical solution of stochastic partial differential equations (stochastic finite elements), Technical Report Informatikbericht Nr.: 2003-06, Department of Mathematics and Computer Science, Technical University Braunschweig, 2003.

[39] J. Kiefer and J. Wolfowitz, Stochastic estimation of the maximum of a regression function, Annals of Mathematical Statistics, 30, pp. 462–466, 1952.

[40] 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.

[41] D. P. Kouri and T. M. Surowiec, Risk–averse PDE-constrained optimization using the conditional value-at-risk, SIAM Journal of Optimization, 26(1), pp. 365–396, 2016.

[42] A. Kunoth and C. Schwab, Sparse adaptive tensor Galerkin approximations of stochastic PDE-constrained control problems, SIAM/ASA Journal on Uncertainty Quantification, 4, pp. 1034–1059, 2016.

[43] I. Lasiecka, Ritz–Galerkin approximation of the time optimal boundary control problem for parabolic systems with Dirichlet boundary conditions, SIAM Journal on Control and Optimization, 22(3), pp. 477–500, 1984.

[44] 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.

[45] H.-C. Lee and M. D. Gunzburger, Comparison of approaches for random PDE optimization problems based on different matching functionals, Computers and Mathematics with Applications, 73(8), pp. 1657 – 1672, 2017.

[46] H.-C. Lee and J. Lee, A stochastic Galerkin method for stochastic control problems, Communications in Computational Physics, 14, pp. 77–106, 2013.

[47] 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, pp. 2012–2038, 2012.

[48] N. Li, J. Fiordilino, and X. Feng, Ensemble time–stepping algorithm for the convection-diffusion equation with random diffusivity, Journal of Scientific Computing, 79(2), pp. 1271–1293, 2019.

[49] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[50] N. Loizou and P. Richtárik, Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods, Computational Optimization and Applications, 77(3), p. 653–710, 2020.

[51] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge University Press, New York, 2014.

[52] F. J. Marín, J. Martínez-Frutos, and F. Periago, Control of random PDEs: An overview, in *Recent Advances in PDEs: Analysis, Numerics and Control: In Honor of Prof. Fernández-Cara's 60th Birthday*, pp. 193–210, Springer International Publishing, Cham, 2018.

[53] M. Martin, S. Krumscheid, and F. Nobile, Complexity analysis of stochastic gradient methods for PDE-constrained optimal control problems with uncertain parameters, ESAIM: Mathematical Modelling and Numerical Analysis, 55(4), pp. 1599–1633, 2021.

[54] M. Martin, F. Nobile, and P. Tsilifis, Multilevel stochastic gradient method for PDE–constrained optimal control problems with uncertain parameters, Technical report, 2019, arXiv:1912.11900.

[55] W. J. Morokoff and R. E. Caflisch, Quasi-Monte Carlo integration, Journal of Computational Physics, 122(2), pp. 218–230, 1995.

[56] K. W. Morton, *Numerical Solution of Convection–Diffusion Problems*, Chapman & Hall, London, Glasgow, New York, 1996.

[57] F. Negri, A. Manzoni, and G. Rozza, Reduced basis approximation of parametrized optimal flow control problems for the Stokes equations, Computers and Mathematics with Applications, 69(4), pp. 319–336, 2015.

[58] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal of Optimization, 19(4), p. 1574–1609, 2009.

[59] Y. Nesterov, A method for solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$, Soviet Mathematics - Doklady, 27, p. 372–367, 1983.

[60] J. Peiró and S. Sherwin, *Finite Difference, Finite Element and Finite Volume Methods for Partial Differential Equations*, pp. 2415–2446, Springer Netherlands, Dordrecht, 2005.

[61] B. T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics, 4, pp. 1–17, 1964.

[62] H. Robbins and S. Monro, A stochastic approximation method, Annals of Mathematical Statistics, 22, pp. 400–407, 1951.

[63] 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.

[64] M. Schmidt, N. Roux, and F. Bach, Convergence rates of inexact proximal-gradient methods for convex optimization, in J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, Curran Associates, Inc., 2011.

[65] F. Shang, L. Jiao, K. Zhou, J. Cheng, Y. Ren, and Y. Jin, ASVRG: Accelerated proximal SVRG, in J. Zhu and I. Takeuchi, editors, *Proceedings of The 10th Asian Conference on Machine Learning*, volume 95 of *Proceedings of Machine Learning Research*, pp. 815–830, PMLR, 14–16 Nov 2018.

[66] G. Stefanou and P. Manolis, Assessment of spectral representation and Karhunen–Loève expansion methods for the simulation of Gaussian stochastic fields, Computer Methods in Applied Mechanics and Engineering, 196, pp. 2465–2477, 2007.

[67] T. J. Sullivan, *Stochastic Galerkin Methods*, Springer International Publishing, Cham, 2015, ISBN 978-3-319-23395-6.

[68] Q. Sun and J. Ming, Multilevel Monte Carlo finite element method for a stochastic optimal control problem, 2016, arXiv preprint arXiv:1512.08403.

[69] 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.

[70] 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.

[71] X. Wan, D. Xiu, and G. E. Karniadakis, Stochastic solutions for the two-dimensional advection-diffusion equation, SIAM Journal on Scientific Computing, 26(2), pp. 578–590, 2004.

[72] T. Wanga and J. Knap, Stochastic gradient descent for semilinear elliptic equations with uncertainties, Journal of Computational Physics, 426, p. 109945, 2021.

[73] W. Zhao, J. Huang, and W.-A. Yong, Lattice Boltzmann method for stochastic convection-diffusion equations, SIAM/ASA Journal on Uncertainty Quantification, 9, pp. 536–563, 2021.

[74] E. Zuazua, Averaged control, Automatica, 50(12), pp. 3077–3087, 2014.

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☐

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : .....
Adı / Name : .....
Bölümü / Department : .....

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : .....
.....
.....
.....
.....

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ☐ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☐
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih** / Date .....