Monte Carlo and Multilevel Monte Carlo Methods with Applications in Financial Engineering

by

DEVANG SINHA

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI-781039, INDIA

June, 2025

Monte Carlo and Multilevel Monte Carlo Methods with Applications in Financial Engineering

A Thesis submitted
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

by

Devang Sinha
(Roll Number: 196123105)

to the

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

June, 2025

Declaration

I hereby declare that the work contained in this thesis entitled **"Monte Carlo and Multilevel Monte Carlo Methods with Applications in Financial Engineering"** was done by me, under the supervision of **Prof. Siddhartha Pratim Chakrabarty**, Professor, Department of Mathematics, Indian Institute of Technology Guwahati for the award of the degree of Doctor of Philosophy and this work has not been submitted elsewhere for a degree.

June, 2025

Devang Sinha

Roll No. 196123105

Department of Mathematics

Indian Institute of Technology Guwahati

Certificate

It is certified that the work contained in this thesis entitled **"Monte Carlo and Multilevel Monte Carlo Methods with Applications in Financial Engineering"** by **Devang Sinha,** a student in the Department of Mathematics, Indian Institute of Technology Guwahati, for the award of the degree of Doctor of Philosophy has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

June, 2025

Prof. Siddhartha Pratim Chakrabarty

Professor

Department of Mathematics

Indian Institute of Technology Guwahati

Dedicated

to

My Family

Acknowledgements

I would like to express my heartfelt gratitude to everyone who supported and guided me throughout my research journey.

First and foremost, I am deeply thankful to my supervisor, Prof. Siddhartha Pratim Chakrabarty, for his invaluable guidance, patience, and unwavering support over the years. His insightful advice, constant encouragement, and high academic standards have been instrumental in shaping this thesis. His active involvement at every stage and persistent push for excellence helped me navigate numerous challenges, making this work possible.

I extend my sincere thanks to the members of my doctoral committee, Prof. N. Selvaraju, Dr. Ayon Ganguly, and Dr. Ashok Singh Sairam, for their regular reviews, constructive feedback, and valuable suggestions, which significantly improved the quality of this work. I am also grateful to all the Department of Mathematics, IIT Guwahati faculty members, for their direct and indirect support throughout my research.

To my friends and colleagues in the department, thank you for your constant motivation, camaraderie, and assistance. This journey would not have been the same without the shared moments of joy, collaboration, and late-night discussions. I would especially like to thank Deb Narayan Barik and Ashish Poonia for the extensive discussions and insights they shared during the various stages of writing this thesis. I am also sincerely thankful to Kuriakose, Deepa, Nabanita, Rik, Adri, Gurinder, Gaurav, Kannan, Mandeep, Sunil, Saurabh, Sirsendu, Anjali, Shilpi, Ramendra and Prakash for their support, which kept me motivated throughout. My labmates, Sachin and Ayushman, deserve special mention for their friendly conversations and, not least, for providing sweets every now and then.

I gratefully acknowledge the *Indian Institute of Technology Guwahati* for providing an excellent academic and research environment. I am also highly thankful to the *Ministry of Human Resource Development, Government of India* for the financial assistance that supported the completion of my research.

I sincerely appreciate the technical staff, Mr Santanu Majumdar, Mr Pranpratim Borgohain, Mr Pranab Jyoti Boro, and the administrative staff for their consistent help and support throughout my time at the institute.

I sincerely thank both the examiners of the thesis for their valuable comments and suggestions.

Lastly, and most importantly, I express my deepest gratitude to my family. Words cannot fully capture my appreciation for my parents, Mrs Meena Sinha and Mr Mukesh Kumar Sinha, and my sister, Ms Divyasha Sinha, for their unconditional love, sacrifices, and unwavering belief in me. Their support has been the cornerstone of all my efforts.

IIT Guwahati Devang Sinha

Abstract

We undertake an in-depth investigation of the Monte Carlo simulation approach and its various extensions to tackle computationally demanding problems that frequently arise in the field of Quantitative Finance. Specifically, we concentrate on three major domains: derivative pricing, risk management, and portfolio optimization, each of which involves significant computational complexity. By leveraging advanced Monte Carlo techniques, we aim to improve numerical efficiency, accuracy, and convergence rates in these financial applications. In the context of derivative pricing, we explore an innovative hybrid algorithm that combines Multilevel Richardson-Romberg extrapolation (ML2R) with adaptive importance sampling to enhance the overall computational efficiency of Monte Carlo estimators. The ML2R technique provides a systematic means of reducing bias while improving the accuracy of numerical approximations, whereas adaptive importance sampling seeks to minimize estimator variance through an optimally chosen change of measure. We establish theoretical guarantees for the convergence of this hybrid methodology, ensuring that it remains robust even when applied to the pricing of financial derivatives. To validate our approach, we conduct numerical experiments within the quantitative finance framework, demonstrating the superior performance of our hybrid method in comparison to the standard ML2R method. Regarding risk management, we delve into the role of stochastic optimization within a biased sampling framework, with a particular emphasis on the Sample Average Approximation (SAA) method. The SAA framework is widely used in stochastic programming to approximate the optimal value of a decision problem by replacing the expected value with a sample-based empirical mean. We investigate the uniform convergence properties of SAA and analyze the computational cost associated with achieving an accurate estimation of the optimal value. Additionally, we incorporate the Multilevel Monte Carlo (MLMC) method within the SAA framework to improve computational efficiency in solving stochastic optimization problems. We demonstrate that by leveraging MLMC one can reduce computational costs to achieve desired accuracy. As part of our analysis, we conduct a root-mean-squared error (RMSE) study, assessing the trade-off between computational effort and estimation accuracy. To substantiate our theoretical insights, we perform numerical simulations in which we estimate Conditional Value-at-Risk (CVaR)—a widely used risk measure—in the context of a Geometric Brownian Motion (GBM) model and a nested expectation

setting. Our empirical results illustrate the benefits of integrating MLMC with SAA, particularly in terms of reducing variance and improving the precision of CVaR estimation under stochastic dynamics. Finally, in the domain of portfolio optimization, we focus on the efficient computation of the minimum-CVaR portfolio, an essential problem in financial risk management that involves constructing a portfolio that minimizes the risk measure CVaR. To this end, we study a variance-reduced variant of Stochastic Gradient Langevin Dynamics (SGLD) to solve the minimum-CVaR portfolio optimization problem efficiently. The SGLD algorithm is particularly well-suited for high-dimensional optimization problems with noisy gradient information, as it incorporates stochastic noise to improve convergence properties. By introducing a variance-reduction technique, we aim to enhance the stability and accuracy of the SGLD algorithm while ensuring faster convergence to the optimal portfolio allocation. We provide rigorous non-asymptotic error bounds for the Expected Excess Risk, quantifying the precision of our variance-reduced SGLD method in solving the optimization problem. Furthermore, we conduct extensive numerical experiments to evaluate the practical effectiveness of our proposed methodology, demonstrating its ability to achieve improved portfolio allocations with lower computational costs. Our results underscore the potential of variance-reduced SGLD as a powerful tool for risk-averse portfolio optimization in high-dimensional settings.

Contents

List of Figures

List of Tables

Chapter 1

An Introduction to Monte Carlo Methods and Variants

1.1 Introduction

In the broader area of Computational Finance, the mere establishment of the existence of the solution to a problem is not sufficient to achieve tangible financial solutions (from a finance perspective) for the problem that has been posed. Accordingly, as is the case for many applications, we seek a solution that (in practice) approximates the solution being sought. In Quantitative Finance, with the exception of a handful of cases, where one can arrive either at an analytical or a semi-analytical solution, for the most part, one needs to devise efficient methods to arrive at the desired approximate solution to the posed problem. These particular situations necessitate resorting to robust computational techniques.

At the heart of this thesis lies a specific computational technique widely used in the finance industry, namely, the Monte Carlo simulation approach. In our study, we intend to explore various variants of the standard Monte Carlo procedure to solve the problems that arise in the computational finance paradigm. Specifically, we intend to look into three broad categories of problems: option pricing, risk estimation, and portfolio management. We intend to employ problem-specific Monte Carlo procedures to solve them numerically. Accordingly, this chapter provides a concise overview of the Monte Carlo method, including a brief exploration of computational complexity. Further, we give a brief overview of two Monte Carlo variants, namely Multilevel Monte Carlo and

Langevin Monte Carlo, as they form the central part of our study.

1.2 Monte Carlo Simulation

Monte Carlo simulation has become an indispensable tool in finance, addressing challenges such as derivative securities pricing, portfolio management, and risk mitigation. Its widespread adoption is driven by the need to simulate high-dimensional stochastic models arising from the increasing complexity and dimensionality of financial problems. In most financial applications, the primary objective is to estimate the expected value of a random variable or function of a random variable. A standard Monte Carlo simulation achieves this by generating samples of the random variable from its distribution and approximating the expected value using the sample average. Mathematically speaking, if we let,

$$\hat{Y} \coloneqq \frac{1}{N}\sum_{k=1}^{N} Y_k, \tag{1.2.1}$$

then $\hat{Y}$ represent the Monte Carlo estimator for $\mathbb{E}[Y]$, where $\{Y_1, Y_2, \ldots, Y_N\}$ are independent and identically distributed (i.i.d.) samples of Y. Since $Y_k \sim \mathcal{L}(Y)$, where $\mathcal{L}(Y)$ represented the distribution of the random variable Y, we have $\mathbb{E}[\hat{Y}] = \mathbb{E}[Y]$, meaning $\hat{Y}$ is an *unbiased* estimator of $\mathbb{E}[Y]$. An important question in this setup is determining the number of samples required to achieve a sufficiently accurate approximation. To address this, we rely on root-mean-squared error (RMSE) analysis, which quantifies the accuracy of the Monte Carlo estimator. Mathematically, the RMSE is defined as:

$$\mathrm{RMSE}(\hat{Y}) \coloneqq \sqrt{\mathbb{E}[(\hat{Y} - \mathbb{E}[Y])^2]} \tag{1.2.2}$$

This measure provides a clear framework for assessing the estimator's accuracy and helps establish the sample size required for a reliable approximation. A simple calculation suggests that,

$$\mathrm{RMSE}(\hat{Y}) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\mathrm{Var}(Y_k)} = \sqrt{\frac{\mathrm{Var}(Y)}{N}} \tag{1.2.3}$$

where $\mathrm{Var}(Y)$ denotes the variance of the random variable Y. To this end, if we require $\mathrm{RMSE}(\hat{Y}) \leq \epsilon$ for some $0 < \epsilon < 1$, then under the assumptions that $\mathrm{Var}(Y) < \infty$, (1.2.3) suggests that the number of samples increases as the square of the inverse of the desired accuracy, *i.e.* $N = \mathcal{O}(\epsilon^{-2})$. In principle, this represents the optimal complexity

achievable, assuming direct access to samples from the Law of Y. However, as one might expect, this ideal scenario is not always realistic. In the following subsection, we analyze the number of samples required in the context where direct access to samples from $\mathcal{L}(Y)$ is unavailable. Instead, we rely on an approximation of the random variable Y.

1.2.1 Sample Complexity in Biased Monte Carlo Paradigm

In many practical scenarios, practitioners often lack access to samples from the true distribution of random variables. For instance, in financial contexts, if the underlying assets are governed by stochastic differential equations (SDEs) that do not admit an analytical solution, then numerical schemes must be employed to approximate the asset prices over a finite time horizon. However, this introduces an additional layer of approximation [40]. Such approximations also appear in the nested simulation framework, which is an internal part of risk estimation [14, 41]. This additional layer of approximation increases the computational complexity required to render RMSE of $\mathcal{O}(\epsilon)$. In order to estimate the computational cost with an extra layer of approximation, we assume the existence of a family of random variables $(Y_h)_{h\in\mathfrak{B}}$, such that,

$$\mathbb{E}[Y_h] \to \mathbb{E}[Y] \text{ as } h \to 0, \tag{1.2.4}$$

where $\mathfrak{B}$ is the set of *bias* parameters such that,

$$\frac{\mathfrak{B}}{n} \subset \mathfrak{B}, \qquad \forall n \in \mathbb{N}$$

and $\mathfrak{B} \cup \{0\}$ is a compact set. To this end, suppose we have access to the N i.i.d copies of Y_h for some bias parameter $h \in \mathfrak{B}$. Let,

$$\hat{Y}_h = \frac{1}{N}\sum_{k=1}^{N} Y_h^k \tag{1.2.5}$$

be a estimator of $\mathbb{E}[Y]$. Since $\mathbb{E}[\hat{Y}_h] \neq \mathbb{E}[Y]$, we say $\hat{Y}_h$ is a *biased* estimator. Now the following result provides an estimate of the computational cost required for $\text{RMSE}(\hat{Y}_h) \leq \epsilon$.

Theorem 1.2.1. *Suppose there exists an $\alpha > 0$ and $c_1 \neq 0$ such that,*

$$\mathbb{E}[Y_h] = \mathbb{E}[Y] + c_1 h^\alpha + o(h^\alpha) \tag{1.2.6}$$

and, $\sigma^2 = \sup_{h \in \mathfrak{B}} \text{Var}(Y_h) < \infty$*, then the computational cost required for* $\sqrt{\mathbb{E}[(\hat{Y}_h - \mathbb{E}[Y])^2]} \leq \epsilon$ *is* $\mathcal{O}(\epsilon^{-(2+\frac{1}{\alpha})})$.

Proof. Refer to [68, Chapter 9] . □

In the above result, the equation (1.2.6) is referred to as first-order *weak expansion*, with α denoting the weak rate of convergence. Further, it is evident from the above result that higher the rate of convergence, the lower is the computational requirement to achieve the desired accuracy. The estimation of a weak rate of convergence is often dependent on the problem under consideration. For example, the asset price is often modelled on geometric Brownian motion or gBm in a financial setup. In this scenario, the *Euler-Maruyama* is a widely used discretization scheme used to estimate the price of the asset at a finite time horizon [32]. Then, under some regularity conditions, it can be shown that $\alpha = 1$ [55, 68]. Consequently, the computational cost required to achieve $\text{RMSE}(\hat{Y}_h) \leq \epsilon$ is $\mathcal{O}(\epsilon^{-3})$. A natural question in these scenarios is whether we can do anything better. And the answer is yes. But before we dwell on the methods to improve the computational cost in a biased sampling setup, we briefly review some standard variance reduction techniques as they play a central part in our thesis.

1.2.2 Overview of Variance Reduction

A conventional way to improve the efficiency of the standard Monte Carlo framework is to employ the so-called variance reduction techniques. These methods often exploit the features of the problem at our disposal to develop strategies leading to variance reduction and, thereby, improved computational efficiency. Here, we provide an overview of two of these variance-reduction techniques that would be prevalent in our work. The interested reader may refer to [40, 68] for a complete and rigorous exposition.

The first variance reduction technique that we review is popularly known as the *control variate*. As before, let $Y_1, \ldots, Y_N$ be N i.i.d samples simulated from $\mathcal{L}(Y)$ and we intend to estimate estimate $\mathbb{E}[Y]$. Suppose, now, for each Y_i, we are able to simulate X_i where we have the knowledge of $\mathbb{E}[X]$. Then this knowledge could be used to simulate $\bar{Y}_i$'s such that $\mathbb{E}[\bar{Y}_i] = \mathbb{E}[Y]$ and $\text{Var}(\bar{Y}_i) < \text{Var}(Y_i)$. To do so, we define,

$$\bar{Y}_i = Y_i - \lambda(X_i - \mathbb{E}[X]) \tag{1.2.7}$$

for some parameter $\lambda \in \mathbb{R}$. Since $\mathbb{E}[X_i] = \mathbb{E}[X]$, we have,

$$\begin{aligned}\mathbb{E}[\,\bar{Y}\,] &= \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\bar{Y}_i\right] \\ &= \mathbb{E}[\,Y\,], \end{aligned} \tag{1.2.8}$$

therefore, $\bar{Y}$ is an unbiased estimator of $\mathbb{E}[Y]$. A simple calculation in the above setup, (refer [40]), suggests that, if,

$$\lambda = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}.$$

then,

$$\frac{\mathrm{Var}(\bar{Y})}{\mathrm{Var}(Y)} = 1 - \rho_{XY}^2$$

where, ρ_{XY} denoted the correlation between X and Y. A natural takeaway from this discussion is that the higher the correlation between the random variable X and Y, the better the variance reduction. However, in practice, we do not know about $\mathrm{Var}(Y)$. Therefore, the practicality of this setup is far from obvious. The interested reader may refer to [40] to study the implementation of the above procedure. However, we would like to emphasize that control variate is among the most efficient and widely used variance reduction techniques.

Another variance reduction technique we recall is known as *importance sampling*. The underlying idea behind the importance sampling framework is to change the probability measure in order to reduce the variance. To formalize this idea, let Y be a $\mathbb{R}^d$ valued random variable and f be the underlying probability density function. Further let $h : \mathbb{R}^d \to \mathbb{R}$ be a function. We know that,

$$\mathbb{E}[h(Y)] = \int_{\mathbb{R}^d} h(y)f(y)dy \approx \frac{1}{N}\sum_{k=1}^{N} h(Y_k). \tag{1.2.9}$$

where $Y_1, \ldots, Y_N$ are sampled from f. Suppose now there exists another probability density function h such that $f(y) > 0 \implies g(y) > 0$. Then

$$\begin{aligned}\mathbb{E}[h(y)] &= \int_{\mathbb{R}^d} h(y)f(y)\,dy \\ &= \int_{\mathbb{R}^d} h(y)\frac{f(y)}{g(y)}g(y)\,dy = \mathbb{E}\left[h(Y)\frac{f(Y)}{g(Y)}\right]. \end{aligned} \tag{1.2.10}$$

Consequently,

$$\hat{Y}_g \coloneqq \frac{1}{N}\sum_{k=1}^{N} h(Y_k)\frac{f(Y_k)}{g(Y_k)} \tag{1.2.11}$$

is an unbiased Monte Carlo estimator of $\mathbb{E}[Y]$ where $Y_1, \ldots, Y_N$ are sampled from g. The effectiveness of the importance sampling framework is steered by selecting the probability density function g. A simple variance analysis undertaken in [40, 68] suggests that if $h(y)$ is non-negative, and

$$g \propto \frac{f(y)h(y)}{\int f(y)h(y)}, \tag{1.2.12}$$

then $\hat{Y}_g$ is a zero-variance estimator. However, such a choice of g is impractical as it assumes the knowledge of $\mathbb{E}[h(Y)]$. Nevertheless, the analysis suggests that choosing a probability density function proportional to $h(y)f(y)$ would lead to an estimator with a lower variance. The interested reader may refer to the above citations to study various examples in the option pricing paradigm. We want to conclude that importance sampling could lead to substantial variance reduction. However, implementing this procedure could be tricky and lead to infinite variance if g is not chosen carefully.

1.3 Multilevel Monte Carlo

In Section 1.2, we saw how introducing bias in the Monte Carlo framework could increase computational cost while estimating the expected value with high accuracy. Following this, we also observe that if we employ variance reduction techniques, we could, in theory, improve the computational cost associated with achieving the same degree of accuracy. The Multilevel Monte Carlo procedure or MLMC exploits the control variate technique to ensure computational savings and achieve the desired accuracy.

The MLMC technique is rooted in the method of parametric integration introduced by Heinrich in [44] for estimating expectations of the form $\mathbb{E}[f(x, \lambda)]$, where x is a finite-dimensional random variable and λ is a parameter. Suppose $\lambda \in [0, 1]$, then one can estimate $\mathbb{E}[f(x, 0)]$ and $\mathbb{E}[f(x, 1)]$ and then use the control variate $\frac{f(x,0)+f(x,1)}{2}$ to approximate $\mathbb{E}\left[f\left(x, \frac{1}{2}\right)\right]$. One can extend this process to recursively approximate $\mathbb{E}[f(x, \lambda)]$ for other intermediate values of λ within $[0, 1]$. Giles extended the approach in [32] to develop an MLMC path simulation method. Unlike Heinrich's work, where the random variable is finite-dimensional, the approach in [32] deals with an infinite-dimensional random variable defined over a Brownian path without relying on parametric

integration. However, the underlying control variate perspective remains closely related to that in [44].

The underlying dynamics of MLMC deal with running independent Monte Carlo simulations with varying degrees of bias and accumulating everything together to get the final result. To be more specific, the procedure uses various levels of resolutions, namely, $l = 1, \ldots, L$, with $l = 1$ being the coarsest and $l = L$ being the finest level. From the perspective of the control variate, the simulation executed in the coarse path is then used as the control variate to carry out the estimation on a more refined path. We present the formalization of this idea as elaborated upon in the following presentation.

Let P be a functional of an underlying random variable X defined on some probability space, and let P_l be the approximation of P on level l. Therefore, in order to estimate the value of the expected value of P on the finest level *i.e.,* $\mathbb{E}[P_L]$, we can make use of the following formula,

$$\mathbb{E}[P_L] = \mathbb{E}[P_1] + \sum_{l=2}^{L} \mathbb{E}[P_l - P_{l-1}].$$

Now, for a given computational cost, the idea is to make an independent estimate, using the standard Monte Carlo, on the terms on the right-hand side so as to minimize the variance. Suppose that, Y_0 is the estimator for $\mathbb{E}[P_1]$ making use of N_1 samples, and let Y_l be the estimator for $\mathbb{E}[P_l - P_{l-1}]$ making use of N_l samples. Then, in the simplest case, we have,

$$Y_l = \frac{1}{N_l} \sum_{k=1}^{N_l} \left(P_l^k - P_{l-1}^k \right).$$

Thus, the combined MLMC estimator is given by, $\widehat{Y} = \sum_{l=1}^{L} Y_l$. Also we observe that, $\mathbb{E}[\widehat{Y}] = \mathbb{E}[P_L]$. With all the preludes being presented in the preceding discussion, we are now in a position to look at the seminal results due to Giles [32].

Theorem 1.3.1. *Let X be $\mathbb{R}^d$-valued random variable defined on a probability space Ω and let $P : \mathbb{R}^d \to R$ be the functional of the random variable. Let $h_l = \mathbf{h}/m^{l-1}$, where $\mathbf{h}$ is the bias parameter of the coarsest level, and $m \in \mathbb{N}$ with $m > 1$. Let X_l denote the corresponding level l numerical approximation of X and let $P_l := P(X_l)$. If there exist independent estimators Y_l, based on N_l Monte Carlo samples, and positive constants $\alpha, \beta, \gamma, c_1, c_2, c_3, c_4$ such that $\alpha \geq \frac{1}{2} \min(\gamma, \beta)$ and*

(i) $|\mathbb{E}[P_l - P]| \leq c_1 h_l^{\alpha}$.

(ii) $\mathbb{E}[Y_l] = \begin{cases} \mathbb{E}[P_1], & l = 1, \\ \mathbb{E}[P_l - P_{l-1}], & l > 1. \end{cases}$.

(iii) $\mathrm{Var}[Y_l] \leq c_2 N_l^{-1} h_l^{\beta}$.

(iv) $C_l \leq c_3 N_l h_l^{-\gamma}$, *where* C_l *is the computational complexity of* Y_l

then there exists a positive constant c_4 *such that for any* $\epsilon < e^{-1}$, *there are values* L *and* N_l *for which the multilevel estimator* $\widehat{Y} = \sum_{l=1}^{L} Y_l$, *has a RMSE with bound,*

$$\sqrt{\mathbb{E}\left[\left(\widehat{Y} - \mathbb{E}[P]\right)^2\right]} < \epsilon,$$

with a computational complexity C, *having the bound,*

$$C \leq \begin{cases} c_4 \epsilon^{-2}, & \beta > \gamma, \\ c_4 \epsilon^{-2} (\log \epsilon)^2, & \beta = \gamma, \\ c_4 \epsilon^{-2 - \frac{(\gamma - \beta)}{\alpha}}, & 0 < \beta < \gamma. \end{cases}$$

Proof. Refer to [32, 68] □

From the above theorem, it is evident that the rate of variance convergence plays a vital role in determining the computational complexity of the MLMC method. In the original study, Giles introduced the MLMC in the context of the option pricing problem where the asset is driven by a stochastic differential equation of the form,

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t \tag{1.3.1}$$

In order to simulate the above SDE, the Euler-Maruyama scheme was used, which is given as,

$$X_{n+1} = X_n + a(X_n, t_n)h + b(X_n, t_n)\Delta W_n, \quad \Delta W_n \sim N(0, \sqrt{h}), \tag{1.3.2}$$

where h is the time-discretization. On level l, the time discretization would be $h_l = \dfrac{\mathbf{h}}{m^{l-1}}$ for some $m \geq 2$, where $\mathbf{h}$ is the discretization of the coarsest level. In order to estimate

the variance of the MLMC, one needed to examine the strong convergence properties of the underlying numerical scheme [55],

$$\left(\mathbb{E}\left[\|X(T) - X_T^l\|^2\right]\right)^{1/2} = O(h_l^{\beta/2}),$$

where h_l is the discretization parameter. On the other hand, while dealing with path-dependent functionals, we often measure the error in sup norm, *i.e.,*

$$\left(\mathbb{E}\left[\sup_{0\leq n\leq 2^l} \|X(nh_l) - X_n^l\|^2\right]\right)^{1/2} = O(h_l^{\beta/2})$$

Even in the best case of a global Lipschitz continuous payoff P, the Euler-Maruyama does not achieve $\beta > 1$, the requirement for the optimal multilevel simulation.

Before proceeding further, we would like to highlight the importance of strong and weak convergence in the context of Monte Carlo and MLMC methods, particularly with respect to algorithm design. As observed in Theorem 1.2.1, the RMSE in the Monte Carlo setting is influenced by the weak convergence rate, denoted by α, and the sample variance of the Monte Carlo estimator. Therefore, algorithm design in this case focuses on constructing approximation schemes that yield a faster rate of weak convergence. In contrast, MLMC relies on both weak and strong convergence rates. The weak convergence rate governs how quickly the bias decays with the level l (corresponding to the step size h_l), and thus influences the choice of the finest level required to meet a given accuracy target. Meanwhile, the strong convergence rate dictates how quickly the variance of level differences $\mathrm{Var}(P_l - P_{l-1})$ decays. This is a key factor in achieving computational efficiency, as the cost savings in MLMC rely on these variances being small at finer levels. Therefore, designing approximation schemes that exhibit not only improves weak convergence but also a high rate of strong convergence is a critical aspect of implementing effective multilevel algorithms.

1.3.1 MLMC with Milstein Discretization

In another article, Giles [31], presented the Milstein scheme for discretization of an SDE, an approach that led to improvement in variance estimation, thereby resulting in the overall computational cost being $O(\epsilon^{-2})$ in the best case scenario. Mathematically, the

Milstein scheme for a one-dimensional stochastic differential equation is given as,

$$X_{n+1} = X_n + a(X_n, t_n)h + b(X_n, t_n)\Delta W_n + \frac{1}{2}b(X_n, t_n)\frac{\partial b}{\partial X}(X_n, t_n)\left((\Delta W_n)^2 - h\right) \quad (1.3.3)$$

Further, he demonstrated that it could be more prudent to make use of different estimators for the coarser and the finer level for the two levels being considered, *i.e.,* P_l^f with l being the finer level, and P_l^c with l being the coarser level. In this case, we require,

$$\mathbb{E}\left[P_l^f\right] = \mathbb{E}\left[P_l^c\right],$$

so that,

$$\mathbb{E}\left[P_L^f\right] = \mathbb{E}\left[P_1^f\right] + \sum_{l=2}^{L}\mathbb{E}\left[P_l^f\right] - \mathbb{E}\left[P_{l-1}^c\right].$$

This method offers flexibility in the process of construction of an approximation, for which $P_l^f - P_{l-1}^c$ is much smaller than the standard $P_l - P_{l-1}$, resulting in a larger value for β, which is the rate of variance convergence in the theorem stated above. The numerical results presented, as well as the thorough numerical analysis carried out in [34], demonstrates the efficacy of the method in case of option pricing problems.

1.3.2 Antithetic MLMC

With the improvement in the variance convergence resulting from the usage of the Milstein scheme, as well as the flexibility offered by the estimator, the extension of the scheme to the higher dimension SDEs is obvious. However, the application of the Milstein scheme for dimensions higher than two is computationally very intensive due to the simulation of the Lévy area that appears in the Milstein discretization. For more discussion on the Lévy area, please refer to [37]. In order to tackle this issue, the authors in [36] introduced the antithetic MLMC estimator based on the classical antithetic variance reduction technique [40]. The idea of antithetic MLMC exploits the flexibility of the general MLMC estimator defined above. Here, based on the coarse path simulation, we have $P_{l-1}^c = P(X_{l-1}^c)$, whereas,

$$P_l^f = \frac{P\left(X_l^f\right) + P\left(X_l^a\right)}{2}.$$

Here, X^a is the antithetic pair of the level l simulation. Note that X^a is defined so that it has the same distribution as X^f, conditional on X^c. Therefore,

$$\mathbb{E}\left[P\left(X^f\right)\right] = \mathbb{E}\left[P\left(X^a\right)\right].$$

Also,

$$P\left(X^f\right) - P\left(X^c\right) \approx -\left(P\left(X^a\right) - P\left(X^c\right)\right),$$

which implies that,

$$\frac{P\left(X^f\right) + P\left(X^a\right)}{2} \approx P\left(X^c\right),$$

thereby resulting in $P_l^f - P_{l-1}^c$ having lower variance than $P_l - P_{l-1}$. Now, in the case of multidimensional SDE, setting the Lévy area equal to zero and using the Milstein scheme in combination with the antithetic estimator, we can achieve,

$$\mathrm{Var}\left[\frac{1}{2}\left(P\left(X^f\right) + P\left(X^a\right)\right) - P\left(X^c\right)\right] = O\left(h^2\right),$$

which is the same as the order obtained by the Milstein scheme for the scalar case.

1.3.3 Multilevel Richardson-Romberg Extrapolation

Another major development was brought with the introduction of Richardson-Romberg extrapolation in the multilevel paradigm. In [32], Giles explored the Richardson extrapolation in the context of both the MLMC and the standard Monte Carlo. The MLMC on its own was significantly better than the Richardson extrapolation. However, together, they worked even better. Lemaire and Pages [60] took this approach and undertook further comprehensive error analysis. They combined the method developed in [32] and Multistep Richardson extrapolation in order to minimize the cost of simulation. The extension relied on the fact that $\mathbb{E}[P_h] - \mathbb{E}[P]$ can be expanded as a polynomial function of h, where P_h is a strong approximation of P. In general, suppose P_h is an approximation of P, based on a discretization parameter h. Then,

$$P_h - P = ah^\alpha + O(h^{2\alpha}).$$

Replacing h by $2h$, accompanied by some basic calculations, we obtain,

$$\widetilde{P} = \frac{2^{\alpha}}{2^{\alpha}-1}P_h - \frac{1}{2^{\alpha}-1}P_{2h},$$

such that,

$$\widetilde{P} - P = O(h^{2\alpha}).$$

With this as the motivation, we consider the bias error $\mathbb{E}[P_h] - \mathbb{E}[P]$. In many applications, the bias error is expanded as,

$$\mathbb{E}[P_h] - \mathbb{E}[P] = \sum_{n=1}^{L} a_n h^{n\alpha} + O(h^{L\alpha}),$$

where L is the finest level of the discretization. In case of the MLMC, we have, $h = \left(\frac{1}{2}\right)^l$ on level l. Therefore,

$$\mathbb{E}[P_l] - \mathbb{E}[P] = \sum_{n=1}^{L} a_n 2^{-nl\alpha} + O(2^{-Ll\alpha}).$$

All that is needed then is to determine a set of weights w_l such that,

$$\sum_{l=1}^{L} w_l = 1, \text{ and } \sum_{l=1}^{L} w_l 2^{-n\alpha l} = 0, \ n = 1, 2, \ldots, L,$$

so that,

$$\left(\sum_{l=1}^{L} w_l \mathbb{E}[P_l]\right) - \mathbb{E}[P] = \mathcal{O}(2^{-\alpha L^2}).$$

Defining, $\widetilde{W_l} := \sum_{l'=l}^{L} w_{l'}$, we can derive the Multilevel Richardson-Romberg extrapolation estimator, given by,

$$\widehat{Y} = \sum_{l=1}^{L} Y_l, \text{ where } Y_l = \frac{1}{N_l}\widetilde{W_l}\sum_{k=1}^{N_l} P_l^k - P_{l-1}^k.$$

With this as the estimator, it was proved that, for $\beta = \gamma$, the overall cost reduces to $\mathcal{O}(\epsilon^{-2}|\log(\epsilon)|)$, while for $\beta < \gamma$ the cost reduces to $\mathcal{O}(\epsilon^{-2}2^{(\gamma-\beta)\sqrt{\log_2 \epsilon/\alpha}})$. The analysis

is supported by the numerical experiments, which demonstrated considerable savings. Therefore, this method is a useful extension to standard MLMC when $\beta \leq \gamma$. In Chapter 3, we review the underlying assumptions and methodology for estimating the weights $\widetilde{W}_l$.

Besides the above improvements, there were various developments pertaining to the multilevel algorithm. For instance, authors in [38] studied a novel combination of MLMC in the quasi-Monte Carlo paradigm. The numerical results presented showed the effectiveness of QMC for SDE applications. However, the results presented in the paper were not supported by any theoretical development in this context. In [72], Rhee and Glynn introduced a new approach to constructing an unbiased estimator, given a family of biased estimators. The idea presented by them is closely related to that of MLMC, wherein the finest level of estimation is chosen, contingent on the level of accuracy. However, in [72], they attempt to produce an unbiased estimator using a total of N samples and performing each simulation on level l with probability p_l. The results presented in [72] demonstrate a significant improvement in the computational cost over the standard MLMC. Also, they prove the square root convergence of the estimator, given that the strong order of convergence is greater than $\frac{1}{2}$ for the path functionals. However, the estimator constructed in [72] has an infinite expected cost whenever $\beta \leq \gamma$.

Apart from the aforementioned developments, there were various improvements in the multilevel setup pertaining to the problem under consideration. We revisit these improvements in the context of option proving problem and risk estimation in Chapter 2, where we thoroughly review the amalgamation of variance reduction techniques and MLMC.

1.4 Langevin Monte Carlo

Another variant of the Monte Carlo methods that dwells in the landscape of optimization belongs to a class of algorithms that derive its foundation from the Langevin dynamics. Mathematically, the dynamic is represented by an SDE given as,

$$dZ_t = -\nabla F(Z_t)dt + \sqrt{2\delta^{-1}}dW_t, \;\; Z_0 = z_0 \in \mathbb{R}^d. \tag{1.4.1}$$

where, F is the potential function, δ is the temperature parameter, and $(W_t)_{t\geq 0}$ is a $d-$dimensional Brownian motion. The above SDE is widely used in physics [78] and also for sampling from Gibbs distribution via Markov Chain Monte Carlo [26]. An extensive study undertaken in [12, 73] suggests that the SDE admits a unique solution under

appropriate conditions on F. The primary feature of the above dynamics is that it leaves the target distribution invariant. The density of the target distribution is given as, [26],

$$\pi_\delta(x) \coloneqq \frac{e^{-\delta F(x)}}{\int_{\mathbb{R}^d} e^{-\delta F(s)} ds}. \tag{1.4.2}$$

However, as it often is, simulating the exact solution of the above SDE is not always possible, and we have to resort to numerical schemes to approximate the target solution. In practice, the simulation is undertaken via Euler-Maruayma discretization given as,

$$Z_{n+1} = Z_n - \nabla F(Z_n)h + \sqrt{2\delta^{-1}}\Delta W_n \tag{1.4.3}$$

where h is the discretization parameter. This approximate sampling from π_δ via Euler-Maruyama discretization is popularly known as *Langevin Monte Carlo* (LMC) or *Unadjust Langevein Monte Carlo*. It should be pointed out that the numerical scheme defines the Markov kernel for which π_δ is not invariant [26]. Therefore, in order to justify the sampling via the numerical scheme, it is necessary to quantify the accumulation of error throughout the algorithm. In this regard, the interested reader may refer to the extensive studies undertaken in [23, 26, 27, 28, 80], that aim at quantifying the error between the $\mathcal{L}(Z_n)$ and π_δ via weak approximation, total variation and Wasserstein distance.

Moving on, in the context of optimization, the Langevin Monte Carlo is used to solve the optimization problem defined as,

$$\mathfrak{p}^* = \min_{\theta \in \mathbb{R}^d} \{F(\theta) \coloneqq \mathbb{E}[f(\theta, X)]\}, \tag{1.4.4}$$

where X is an $\mathbb{R}^q$ valued random variable. The relevance of the above dynamics stems from the fact that, as $\delta \to \infty$, the target measure concentrates around the minimizer of F. In order to solve the above problem, we perform the LMC simulation starting with θ_0, which is given as,

$$\theta_{n+1} = \theta_n - h\nabla F(\theta_n) + \sqrt{2\delta^{-1}}\Delta W_n, n \in \mathbb{N} \tag{1.4.5}$$

and approximate $\mathfrak{p}^*$ by $\mathbb{E}[F(\theta_n)]$ for some $n \in \mathbb{N}$. The underlying idea is to generate θ_N that minimizes the Excess Risk (ER), which is defined as,

$$\text{ER} := \mathbb{E}[F(\theta_n)] - \mathfrak{p}^*. \tag{1.4.6}$$

An important requirement to run the scheme described in (1.4.5) is the calculation of the gradient of the $F(\theta)$, *i.e.*, $\nabla F(\theta)$. In many practical applications, it is often difficult to estimate $F(\theta)$, let alone its gradient. Moreover, we may not have any knowledge about the distribution of X but have access to the sample from the distribution. A natural approach to resolve the first issue is to construct an approximation of $F(\theta)$ by generating $(X_k)_{1\leq j\leq n}$ i.i.d sample from Law(X) and taking the sample average, *i.e.*,

$$F_N(\theta) = \frac{1}{N}\sum_{k=1}^{N} f(\theta, X_k). \tag{1.4.7}$$

A similar approximation is considered if we have access to the samples of X without any knowledge about its distribution. Consequently, the gradient is approximated as,

$$\nabla F_N(\theta) = \frac{1}{N}\sum_{k=1}^{N} \nabla f(\theta, X_k). \tag{1.4.8}$$

With the above approximation, we solve the following Empirical Minimization problem, given as,

$$\mathfrak{p}_N^* = \min_{\theta\in\mathbb{R}^d} F_N(\theta). \tag{1.4.9}$$

The sample average approximation described above poses another challenge in our simulation. Suppose N is very large, which it often is, then the numerical scheme requires us to estimate $\nabla F_N(\theta)$ at each iteration step, which could be extremely time-consuming and slow down the convergence. One could look into the so-called stochastic gradient method to cater to this issue, where one gives up the accuracy for faster convergence. Accordingly, our scheme becomes,

$$\theta_{n+1} = \theta_n - h\nabla v_n + \sqrt{2\delta^{-1}}\Delta W_n \tag{1.4.10}$$

where v_n is a conditionally unbiased estimator of $\nabla F_N(\theta_n)$, *i.e.*, $\mathbb{E}[v_n] = \nabla F_N(\theta_n)$. A natural candidate for v_n is $\nabla f(\theta_n, X_{i_{n+1}})$, where $i_{n+1} \sim \text{Unif}\{1, 2, \ldots, N\}$. However, for a better approximation one could sample $B_n \subset \{1, \ldots, N\}$ such that $|B_n| = B < N$ for all n, and set $v_n = \frac{1}{N}\sum_{j\in B_n} f(\theta_n, X_j)$. This approach is popularly known as *Stochastic Gradient Langevin Dynamics* or SGLD in the literature.

Stochastic Gradient Langevin Dynamics have been extensively used to solve the optimization problem that arises in the machine learning paradigm. The interested reader

may refer to [1, 77, 82, 84, 87] for the application of SGLD in machine learning. The SGLD methods have traditionally been used to solve the convex optimization problem. However, [71] laid down the theoretical foundation for solving the optimization problem in the non-convex setting. Building upon the theoretical foundation in [71], various variants and improvements have been undertaken to solve the optimization problem. The interested reader may refer to [16, 86, 90] and references therein for the extensive study on the application of SGLD on non-convex optimization. Although the majority of the study with SGLD has been directed towards solving optimization problems in the machine learning domain, we intend to revisit this method to solve optimization problems that appear in the financial engineering paradigm.

1.5 Organization of the Thesis

The remainder of the thesis concentrates on three major problems that appear in the area of quantitative finance, namely, Derivative pricing, Risk management and Portfolio optimization. We intend to study the Monte Carlo method and its variants in order to solve this problem efficiently. Accordingly, in Chapter 2, we undertake a comprehensive review of the recent developments in the variance reduction techniques in the paradigm of Multilevel Monte Carlo simulation, where we dwell on the Importance sampling framework in the context of Multilevel Monte Carlo and review its effectiveness while pricing financial derivatives. Further, we look into the adaptive sampling framework in the context of nested simulation, which appears in Value-at-Risk and Conditional Value-at-Risk estimation and its extension to MLMC with the aim of achieving better computational cost via a Multilevel framework.

Taking cues from the discussion undertaken in Chapter 2, we investigate the Multilevel Richardson-Romberg extrapolation with the importance of the sampling procedure in Chapter 3 to achieve greater computational efficiency. We extend this approach by leveraging the Central Limit Theorem in the context of the ML2R setup. We perform a rigorous convergence study of the hybrid algorithm and demonstrate the efficacy of our hybrid setup via numerical examples in the paradigm of quantitative finance.

In Chapter 4, we consider a stochastic optimization problem where we assume the sampling is biased. In the biased sampling framework, we investigate the Non-Asymptotic properties of Sample Average Approximation (SAA), in order to solve the stochastic optimization problem. In this regard, we perform a complete computational cost analysis and determine the sample complexity required to achieve an accurate solution. Further,

we extend the approach to the multilevel setup and undertake a rigorous analysis to establish uniform convergence and determine the computational cost. We further consider an RMSE error analysis and perform numerical experiments in the risk estimation paradigm to demonstrate the efficacy of our study.

In Chapter 5, we again consider a stochastic optimization problem where we employ variance reduction techniques in the Langevin Monte Carlo paradigm to improve the results obtained in the context of a minimum CVaR-portfolio optimization problem. We undertake a comprehensive analysis of our method where we determine the bound for Excess Risk (1.4.6). Finally, in Chapter 6, we discuss the future prospects of our study.

Chapter 2

A Review of Efficient Multilevel Monte Carlo Algorithms for Derivative Pricing and Risk Management

In this chapter, we discuss four methodological contributions to the paradigm of MLMC simulation. The first two algorithms provide substantial improvement in the computational cost in the landscape of derivative pricing, whereas the latter two dwell in the landscape of risk estimation.

2.1 Importance Sampling Multilevel Algorithm

Since the advent of MLMC in literature, one of the directions of its progression has been through various attempts to combine this algorithm with the already existing variance reduction techniques. For instance, as discussed in Chapter 1, authors in [36, 37] studied and analyzed the combination of antithetic variates and MLMC in order to bypass the Levy area simulation (encountered while using the Milstein discretization scheme) in order to simulate higher-dimensional SDEs. However, in this section, we primarily focus on the combination of an importance sampling algorithm and a multilevel estimator.

The idea of incorporating importance sampling with multilevel estimators is derived from the seminal paper by Arouna [7]. Arouna's idea relied upon the parametric change of measure and the deployment of a search algorithm to approximate the optimal change of the measure parameter in order to minimize the variance of the standard Monte Carlo estimator. Before we discuss the research undertaken in the area of MLMC pertaining to

importance sampling algorithm, we give a brief overview of the parametric importance sampling approach.

Consider a general problem of estimating $\mathbb{E}[G(X)]$, where X is a d-dimensional random variable. If $f(x)$ is the multivariate density function of X, then,

$$\mathbb{E}\left[G(X)\right] = \int G(x)f(x)dx = \int G(x+\theta)f(x+\theta)dx = \int h(\theta,x)f(x)dx,$$

where, $h(\theta,x) = \dfrac{G(x+\theta)f(x+\theta)}{f(x)}$. This implies that, $\mathbb{E}[G(X)] = \mathbb{E}[h(\theta,X)]$. Therefore, we need to determine the optimal value of θ such that $\mathrm{Var}[h(\theta,X)]$ is minimum. Mathematically, this is represented as,

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} \mathrm{Var}[h(\theta,X)]. \tag{2.1.1}$$

In order to solve the above problem, one can resort to the usage of the Robbins-Monro algorithm that deals with a sequence of random variable $(\theta_i)_{i\in\mathbb{N}}$, which approximates θ^* accurately. However, the convergence of this algorithm requires certain restrictive conditions, which are known as the non-explosion condition (NEC) (see, e.g. [2]),

$$\mathbb{E}\left[h^2\left(\theta,X\right)\right] \leq C\left(1+|\theta|^2\right) \text{ for all } \theta\in\mathbb{R}^d.$$

In order to deal with this restrictive condition, the authors in [19, 45] introduced a truncation-based procedure which was furthered in [6, 58]. An unconstrained procedure to approximate θ^*, by using the regularity of the density function in an extensive manner, was introduced in [59] along with the proof of convergence of the algorithm. Besides, the stochastic approximation algorithm, one can also use a deterministic algorithm such as sample average approximation, which, while being computationally expensive, provides for a better approximation to θ^*.

In problems dealing with the pricing of the options, for the most part the underlying stochastic process $(X_t)_{0\leq t\leq T}$ (where $T>0$ is a finite time horizon), is governed by some SDEs. The general form of these SDEs is given as follows:

$$dX_t = b(X_t)dt + \sum_{j=1}^{q}\sigma_j(X_t)dW_t^j,\ X_0 = x\in\mathbb{R}^d, \tag{2.1.2}$$

where $W := \begin{pmatrix} W_1 & W_2 & \dots & W_q \end{pmatrix}$ is a q-dimensional Brownian motion on a filtered probability space $(\Omega, (\mathcal{F}_t)_{0\le t\le T}, \mathbb{P})$ with $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma_j : \mathbb{R}^d \to \mathbb{R}^d$ being the functions satisfying the following,

Assumption 2.1.1. *There exists a constant $K_{b,\sigma} > 0$ such that for all $x, y \in \mathbb{R}^d$,*

$$|b(x) - b(y)| + \sum_{j=1}^{q} |\sigma_j(x) - \sigma_j(y)| < K_{b,\sigma}|x - y|. \tag{2.1.3}$$

Assumption (2.1.1) ensures the existence and the uniqueness of the solution to (2.1.2) [55]. For the most part, constructing an analytical or semi-analytical solution to (2.1.2) is not possible, and therefore we need to rely on discretization schemes such as Euler or Milstein in order to simulate the SDEs. For a detailed discussion on these discretization schemes, interested readers may refer to [55]. Further, following the idea of [7], we consider a family of stochastic process $(X_t(\theta))_{0\le t\le T}$, with $\theta \in \mathbb{R}^d$, being governed by the following SDE:

$$dX_t(\theta) = (b(X_t(\theta)) + \sigma(X_t(\theta))\theta)dt + \sum_{j=1}^{q} \sigma_j(X_t(\theta))dW_t^j,\ \sigma(x) = \begin{pmatrix} \sigma_1(x) & \dots & \sigma_q(x) \end{pmatrix}. \tag{2.1.4}$$

As a consequence of Girsanov's Theorem, we know that there exists a risk-neutral probability measure $\mathbb{P}_\theta$, which is equivalent to $\mathbb{P}$ such that,

$$\left.\frac{d\mathbb{P}_\theta}{d\mathbb{P}}\right|_{\mathcal{F}_t} = \exp\left(-\langle\theta, W_t\rangle - \frac{1}{2}|\theta|^2 t\right), \tag{2.1.5}$$

under which the process $(\theta t + W_t)_{0\le t\le T}$ is a Brownian motion. Therefore,

$$\mathbb{E}_{\mathbb{P}}\left[G(X_T)\right] = \mathbb{E}_{\mathbb{P}_\theta}\left[G(X_T(\theta))\right] = \mathbb{E}_{\mathbb{P}}\left[G(X_T(\theta))e^{-\langle\theta, W_T\rangle - \frac{1}{2}|\theta|^2 T}\right]. \tag{2.1.6}$$

Therefore, following the discussion above, we have,

$$\mathbb{E}\left[G(X_T)\right] = \mathbb{E}\left[h(\theta, X_T)\right].$$

Here, $h(\theta, X_T) = G(X_T(\theta))e^{-\langle\theta, W_T\rangle - \frac{1}{2}|\theta|^2 T}$. Now the idea of importance sampling Monte

Carlo method is to estimate $\mathbb{E}\left(G(X_T)\right)$, where θ is given by,

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} \text{Var}\ \left(G(X_T(\theta))e^{-\langle\theta,W_T\rangle-\frac{1}{2}|\theta|^2T}\right). \tag{2.1.7}$$

In the context of the multilevel estimator, we present two approaches studied in [3, 10, 52], adapting the ideas studied by authors in [2, 7] and extending it to multilevel scenarios. Under the parametric change of measure, the general multilevel estimator is defined as,

$$\mathbb{E}\left[Y_L\right] = \mathbb{E}\left[Y_1^{\theta_1}\right] + \sum_{l=2}^{L}\mathbb{E}\left[Y_l^{\theta_l} - Y_{l-1}^{\theta_l}\right], \text{ where } Y_l^{\theta} = G(X_l^{\theta})e^{-\langle\theta_l,W_T^l\rangle-\frac{1}{2}|\theta_l|^2T}. \tag{2.1.8}$$

Under the framework of the multilevel estimator, the parametric importance sampling estimator looks like as follows:

$$\widehat{Y}_L^{\theta} = \frac{1}{N_1}\sum_{k=1}^{N_1} Y_1^{k,\theta_1} + \sum_{l=2}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l}\left(Y_l^{k,\theta_l} - Y_{l-1}^{k,\theta_l}\right). \tag{2.1.9}$$

Considering the variance of the above estimator, we have [52],

$$\text{Var}[\widehat{Y}_L^{\theta}] = \frac{1}{N_1}\text{Var}[Y_1^{\theta_1}] + \sum_{l=2}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l}\frac{(m-1)T}{m^{l-1}}\text{Var}[Y_l^{\theta_l} - Y_{l-1}^{\theta_l}]. \tag{2.1.10}$$

Therefore, as discussed, in order to solve the problem of minimizing the overall variance of the estimator described above, we intend to minimize the variance at each level of resolution, *i.e.*, we aim at approximating θ_l^* for $l = 1, \ldots, L$, such that,

$$\theta_1^* = \arg\min_{\theta\in\mathbb{R}^d}\text{Var}[Y_1^{\theta_1}] \text{ and } \theta_l^* = \arg\min_{\theta\in\mathbb{R}^d}\text{Var}[Y_l^{\theta_l} - Y_{l-1}^{\theta_l}]. \tag{2.1.11}$$

Further, pertinent to the discussion carried out in [2] and another application of Girsanov's Theorem, the above problem can be reformulated as,

$$\begin{aligned}\theta_1^* &= \arg\min_{\theta_1\in\mathbb{R}^d}\mathbb{E}\left[G(X_1)^2e^{-\langle\theta_1,W_T^1\rangle+\frac{1}{2}|\theta_1|^2T}\right]\\ \theta_l^* &= \arg\min_{\theta\in\mathbb{R}^d}\mathbb{E}\left[\frac{m^{l-1}}{(m-1)T}\left(G(X_l) - G(X_{l-1})\right)^2 e^{-\langle\theta_l,W_T^l\rangle+\frac{1}{2}|\theta_l|^2T}\right].\end{aligned} \tag{2.1.12}$$

We present below the two algorithms, namely, the sample average approximation and

stochastic approximation, in order to approximate the θ_l's as the solution to (2.1.12).

2.1.1 Sample Average Approximation

The sample average approximation deals with approximating the above expectations using $\widetilde{N}_l$ sample paths. More specifically,

$$\mathbb{E}\left[G(X_1)^2 e^{-\langle\theta_1, W_T^1\rangle+\frac{1}{2}|\theta_1|^2 T}\right] \approx \frac{1}{\widetilde{N}_1}\sum_{k=1}^{\widetilde{N}_1} G\left(X_1^k\right)^2 e^{-\langle\theta_1, W_T^{1,k}\rangle+\frac{1}{2}|\theta_1|^2 T} \equiv \mathbb{V}_1, \tag{2.1.13}$$

and,

$$\begin{aligned}&\mathbb{E}\left[\frac{m^{l-1}}{(m-1)T}\left(G(X_l)-G(X_{l-1})\right)^2 e^{-\langle\theta_l, W_T^l\rangle+\frac{1}{2}|\theta_l|^2 T}\right]\\ &\approx \frac{1}{\widetilde{N}_l}\sum_{k=1}^{\widetilde{N}_l}\left(\frac{m^{l-1}}{(m-1)T}\left(G(X_l^k)-G(X_{l-1}^k)\right)^2 \times e^{-\langle\theta_l, W_T^{l,k}\rangle+\frac{1}{2}|\theta_l|^2 T}\right) \equiv \mathbb{V}_l.\end{aligned} \tag{2.1.14}$$

Having approximated the expectation in the minimization problem, the authors used the standard Newton-Raphson algorithm on the functions $\mathbb{V}_1$ and $\mathbb{V}_l$ in order to approximate θ_l^*, for $l = 1, \ldots, L$. In [2] it is proved that if the functional $G(X)$ satisfies the non-degeneracy conditions *i.e.,* $\mathbb{P}\left(G\left(X_T^1\right) \neq 0\right) > 0$ and $\mathbb{P}\left(\left(G\left(X_T^l\right) - G\left(X_T^{l-1}\right)\right) \neq 0\right) > 0$ and have finite second moments, then by [2, Lemma 2.1], $\mathbb{V}_1$ and $\mathbb{V}_l$ are infinitely continuously differentiable. Moreover, both $\mathbb{V}_1$ and $\mathbb{V}_l$ are both strongly convex, thus implying the existence of the unique minimum θ_0^* and θ_l^* as the solution to equations (2.1.12).

2.1.2 Adaptive Stochastic Approximation

Under the stochastic approximation, studied in [3, 10] the aim of determining the optimal change of parameter θ_l^* for $l = 1, \ldots, L$ is carried out using the Robbins-Monro algorithm. Here, we briefly describe the algorithm. Consider a compact convex set $\Theta \subset \mathbb{R}^d$ such that $0 \in \text{int}(\Theta)$. Then the recursive algorithm with projection is defined as follows,

$$\theta_l^{n+1} = \mathbf{Proj}_\Theta\left[\theta_l^n - \lambda_{n+1} H_l(\theta_l^n, Y_l, W_T^l)\right], \tag{2.1.15}$$

where $\mathbf{Proj}_{\Theta}(\theta)$ is the Euclidean projection onto the set Θ. The sequence $(\lambda_n)_{n\geq 1}$ is a decreasing sequence of positive real numbers satisfying,

$$\sum_{n=1}^{\infty}\lambda_n = \infty \text{ and } \sum_{i=1}^{\infty}\lambda_n^2 < \infty. \tag{2.1.16}$$

Also,

$$H_l(\theta_l^n, Y_l, W_T^l) = \left\{ \begin{array}{ll} (\theta_1 T - W_T^1)\left(G(X_1)^2 e^{-\langle\theta_1, W_T^1\rangle + \frac{1}{2}|\theta_1|^2 T)}\right), & l = 1, \\ (\theta_l T - W_T^l)\left[\frac{m^{l-1}}{(m-1)T}\left(G(X_l) - G(X_{l-1})\right)^2 e^{-\langle\theta_l, W_T^l\rangle + \frac{1}{2}|\theta_l|^2 T}\right], & l = 2, \ldots, L. \end{array} \right\} \tag{2.1.17}$$

The algorithm described above is the constrained version of the Robbins-Monro algorithm. The inclusion of the projection operator in the recursive algorithm is to satisfy the NEC. Similar to the discussion carried out in the previous section, if the non-degeneracy conditions are satisfied *i.e.*, $\mathbb{P}\left(G(X_T^1) \neq 0\right) > 0$ and $\mathbb{P}\left(\left(G(X_T^l) - G(X_T^{l-1})\right) \neq 0\right) > 0$, further assuming the finite second moment of $G(X_1)$ and $G(X_l) - G(X_{l-1})$, we can conclude the convergence of the θ_l^*, constructed recursively using equation (2.1.15), for various level of resolutions.

The term adaptive is used in the sense that, the estimation of the optimal importance sampling parameter and the MLMC run simultaneously. The multilevel estimator in this case is given as follows,

$$\widehat{Y}_L^{\theta} = \frac{1}{N_1}\sum_{k=1}^{N_1} Y_1^{k,\theta_1^{k-1}} + \sum_{l=2}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l}\left(Y_l^{k,\theta_l^{k-1}} - Y_{l-1}^{k,\theta_l^{k-1}}\right). \tag{2.1.18}$$

However, for the purpose of practical implementation, one needs to stop the approximation procedure after a finite number of iterations.

Having approximated the θ_l^* for $l = 1, \ldots, L$, we use the multilevel algorithm described by equation (2.1.18) to estimate our expectation. The studies carried out in [3, 10, 52] demonstrate the accuracy of the hybrid importance sampling multilevel algorithm over the standard multilevel algorithm, through a series of numerical examples, where the underlying SDEs are multi-dimensional.

As one can readily observe, the sample average approximation is a deterministic algorithm, whereas the stochastic one is not. The robustness and the stability of the sample average approximation make it a dominant choice for estimating θ_l. However, the algorithm suffers from a low convergence rate, and large memory footprint [52]. Also, the

algorithm does require the calculation of $\nabla\mathbb{V}_l(\theta_l)$, which in turn requires more conditions on the regularity of the payoff $G(X)$, making it less convenient from the practitioner's point of view. On the other hand, the stochastic algorithm does not require the calculation of $\nabla\mathbb{V}_l(\theta_l)$, resulting in less restrictive conditions and making it feasible from the practitioner's point of view. Although the convergence of the stochastic algorithm is faster than the sample average algorithm, it is not robust and is sensitive to the parameter λ_n. We would like to point out that the study discussed above only deals with the Euler MLMC and is restricted to the use of Euler discretization to simulate the underlying SDEs. The interested reader may refer to [10, 52] for a detailed discussion on the convergence results and implementation. We conclude this section by presenting the pseudo-code of the two algorithms discussed.

Algorithm 1: Importance sampling: Sample average approximation

Sample $(X_1^1, X_1^2, \ldots, X_1^{\tilde{N}_1}) \overset{i.i.d}{\sim} X_1$;
Solve $\nabla\mathbb{V}_1(\theta_1) = 0$ using Newton-Raphson method.;
Set $Y = 0$;
for *l = 2:L* **do**
  Sample $(X_l^1, X_{l-1}^1), (X_l^2, X_{l-1}^2), \ldots, (X_l^{\tilde{N}_l}, X_{l-1}^{\tilde{N}_l}) \overset{i.i.d}{\sim} (X_l, X_{l-1})$;
  Solve $\nabla\mathbb{V}_l(\theta_l) = 0$ using Newton-Raphson method.;
end
Sample $(X_1^{1,\theta_1}, X_1^{2,\theta_1}, \ldots, X_1^{N_1,\theta_1}) \overset{i.i.d}{\sim} X_1^{\theta_1}$;
$Y \leftarrow Y + \frac{1}{N_1}\sum_{k=1}^{N_1} G(X_1^{k,\theta_1}) e^{-\langle\theta_1, W_T^1\rangle - \frac{1}{2}|\theta_1|^2 T}$;
for *l = 2:L* **do**
  Sample $(X_l^{1,\theta_l}, X_{l-1}^{1,\theta_l}), \ldots, (X_l^{N_l,\theta_l}, X_{l-1}^{N_l,\theta_l}) \overset{i.i.d}{\sim} (X_l^{\theta}, X_{l-1}^{\theta})$;
  $Y \leftarrow Y + \frac{1}{N_l}\sum_{k=1}^{N_l} (G(X_l^{k,\theta_l}) - G(X_{l-1}^{k,\theta_l})) e^{-\langle\theta_l, W_T^l\rangle - \frac{1}{2}|\theta_l|^2 T}$
end
return Y

From Algorithms 1 and 2, it is evident that the computational cost of the hybrid MLMC Importance sampling method is higher than the standard MLMC algorithms. This increased computational cost is attributed to the extra computation required to solve the optimization problem. However, we would like to point out that the substantial decrease in variance compensates for the increased computational cost, making the hybrid method a more suitable approach.

Algorithm 2: Importance sampling: Stochastic Approximation

Set $\theta = \theta_0$;
Set $Y = 0$;
for $k = 1 : N_1$ **do**
  Sample (X_1, X_1^θ, W_T^1);
  $Y \leftarrow Y + G(X_1^\theta)e^{-\langle\theta, W_T^1\rangle - \frac{1}{2}|\theta|^2 T}$;
  $\theta \leftarrow \mathbf{Proj}_\Theta(\theta - \lambda_k H_1(\theta, Y_1, W_T^1))$;
end
$Y \leftarrow Y/N_1$;
for $l = 2 : L$ **do**
  Set $\theta = \theta_0$;
  Set $S = 0$;
  for $k = 1 : N_l$ **do**
    Sample $(X_l, X_{l-1}, X_l^\theta, X_{l-1}^\theta, W_T^l)$;
    $S \leftarrow S + (G(X_l^{\theta_l}) - G(X_{l-1}^\theta))e^{-\langle\theta, W_T^l\rangle - \frac{1}{2}|\theta|^2 T}$;
    $\theta \leftarrow \mathbf{Proj}_\Theta[\theta - \lambda_k H_l(\theta, Y_l, W_T^l)]$;
  end
  $Y \leftarrow Y + S/N_l$;
end
return Y

2.2 MLMC and Efficient Risk Estimation.

Risk measurement and consequent management is one of the essential components of financial engineering. The computation of risk measures for a financial portfolio is both challenging and computationally intensive, which may be ascribed to computations involving nested expectation, which entails multiple evaluations of the loss to the portfolio, for distinct risk scenarios. Further, the cost of computing loss of portfolio entailing thousands of derivatives becomes progressively expensive with an increase in the size of the portfolio [33]. Value-at-Risk (VaR), Conditional VaR (CVaR), and the likelihood of a large loss are the necessary risk metrics used to estimate the risk of a financial portfolio. At the core of this estimation, is the necessity of evaluating the nested expectation, given by,

$$\varrho = \mathbb{E}\left[H\left(\mathbb{E}[X|Y]\right)\right] \tag{2.2.1}$$

where H is the Heaviside function. More specifically, suppose we need to compute the probability of the expected loss being greater than $L_\varrho \in \mathbb{R}$, *i.e.,* we are interested in the following computation:

$$\varrho = \mathbb{E}\left[H(\mathbb{E}[\Delta|R_\tau] - L_\varrho)\right], \tag{2.2.2}$$

where $\mathbb{E}[\Delta|R_\tau]$ is the expected loss in a risk-neutral world, with R_τ being a possible risk scenario at some short risk (time) horizon τ. Also, Δ is the average loss of many losses incurred from different financial derivatives, depending upon similar underlying assets [33], that is,

$$\Delta = \frac{1}{K}\sum_{i=1}^{K}\Delta_i, \tag{2.2.3}$$

where K is the total number of derivatives and Δ_i is the loss from the i-th derivative. The average is considered to ensure the boundedness of Δ when the portfolio size of K increases. A standard and straightforward way to estimate the nested expectation (2.2.1) is the usage of the Monte Carlo method. This involves, simulating M independent scenarios of the risk parameter R_τ, and for each risk scenario, N total loss samples, which are independent. This method was explored in [41], and an extended analysis was carried out in [39]. The total computational cost to perform the above simulation is $O\left(\max\left(K\epsilon^{-2}, \epsilon^{-3}\right)\right)$ in order to achieve the RMSE of ϵ [33]. In order to handle this issue, we present the ideas studied in [35, 42] under the realm of MLMC.

2.2.1 Adaptive Sampling Multilevel estimator

As mentioned in the previous section, the cost of the standard Monte Carlo to achieve the RMSE of ϵ is $O\left(\epsilon^{-3}\right)$. To improve the computational complexity, the authors in [14] developed an efficient through the adaptation of the sample size required in the inner sampler of Monte Carlo, to the particular outer sampler random variable. Under certain conditions, the authors were able to achieve the $O\left(\epsilon^{-5/2}\right)$ computational complexity to achieve the RMSE of ϵ. Giles in [35] extended this approach to the multilevel framework and was able to achieve $O\left(\epsilon^{-2}|\log\epsilon|^2\right)$ computational cost for a RMSE tolerance ϵ. Before presenting the work initiated by Giles, we put forth a brief review of the studies carried out in [14] and [41].

The authors in [41], estimated the inner expectation of the equation (2.2.1), *i.e.*, $\mathbb{E}[X|Y=y]$, for a given y, using the unbiased Monte Carlo estimator, with N sample paths, as given by,

$$\widehat{\mathcal{Z}}_N(y) = \frac{1}{N}\sum_{n=1}^{N} x_n(y), \tag{2.2.4}$$

where, $\{x_n(y)\}_n$ are the mutually independent samples from the random variable X,

conditioned on $Y = y$. Again, using the Monte Carlo for the outer expectation, we have,

$$\varrho \approx \frac{1}{M} \sum_{m=1}^{M} H\left(\widehat{\mathcal{Z}}_N(y_m)\right), \tag{2.2.5}$$

where $\{y_m\}_m$ are the mutually independent samples from the random variable Y. Further, to bound the RMSE of the estimator (2.2.5), they impose the following assumption, *i.e.*,

Assumption 2.2.1. *Let two random variables $\mathbb{E}[X|Y]$ and $\widehat{\mathcal{Z}}_N$ have the joint density $d_N(y,z)$ and assume that for $i = 0, 1, 2$, $\dfrac{\partial}{\partial y_i} d_N(y,z)$ exists, plus there exists a non-negative function $d_{i,N}$, such that,*

$$\left|\frac{\partial}{\partial y_i} d_N(y,z)\right| \leq d_{i,N}, \text{ for all } N, y, z, \text{ and } \sup_N \int_{-\infty}^{\infty} |z|^q d_{i,N}(z) dz < \infty,$$

for all $0 \leq q \leq 4$.

Under the light of the above assumption, the RMSE of the estimator (2.2.5) is $\mathcal{O}\left(M^{-1/2} + N^{-1}\right)$. Therefore, in order to achieve the RMSE of $\mathcal{O}(\epsilon)$ we need $M = \mathcal{O}(\epsilon^{-2})$ and $N = \mathcal{O}(\epsilon^{-1})$, leading to the total computational complexity of $\mathcal{O}(\epsilon^{-3})$. Authors in [14] developed an adaptive sampling technique to deal with the high computational complexity previously discussed. Their approach was based on the likelihood that an additional sample will result in a negative estimate of $\widehat{\mathcal{Z}}_{N+1}$ having estimated that $\widehat{\mathcal{Z}}_N > 0$ for given Y. More specifically, they showed that,

$$\mathbb{P}\left[\widehat{\mathcal{Z}}_{N+1} \leq 0 | \widehat{\mathcal{Z}}_N\right] \leq \frac{\sigma^2}{\left(N\widehat{\mathcal{Z}}_N(Y) + \mu\right)^2} \approx \frac{\sigma^2}{N^2\mu^2},$$

where $\mu = \mathbb{E}[X|Y]$ and $\sigma^2 = \mathrm{Var}[X|Y]$. Therefore, if $N \geq \dfrac{\epsilon^{-1/2}\sigma}{|\mu|}$, then the probability that $H\left(\widehat{\mathcal{Z}}_N(Y)\right) = H\left(\widehat{\mathcal{Z}}_{N+1}(Y)\right) \approx H\left(\mathbb{E}[X|Y]\right)$ is equal to $1 - \epsilon$. Based on these observations, the authors in [14] introduced two algorithms, the first being based on the minimization of the total number of samples for all inner Monte Carlo samplers with respect to given tolerance ϵ, and the second being iterative, estimating $|\mu|$ and σ after every iteration, for a given value of Y, using N samples further adding more inner samples till $\dfrac{N\mu}{\sigma}$ exceeds some error margin threshold. Under these two algorithms, it was observed

that the overall computational complexity is $\mathcal{O}(\epsilon^{-5/2})$ [35]. The authors in [35] introduced the above algorithms in the realm of multilevel simulation, wherein they used the multilevel estimator in order to achieve an approximation to the outer expectation while making use of the sample size in the inner expectation as the discretization parameter. More specifically,

$$\widetilde{\varrho} := \sum_{l=1}^{L} \frac{1}{M_l} \sum_{m=1}^{M_l} H\left(\widehat{\mathcal{Z}}_{N_l}^{f,l,m}(y^{l,m})\right) - H\left(\widehat{\mathcal{Z}}_{N_{l-1}}^{c,l,m}(y^{l,m})\right), \tag{2.2.6}$$

where,

$$\widehat{\mathcal{Z}}_{N_l}^{f,l,m}(y) = \frac{1}{N_l} \sum_{n=1}^{N_l} x^{f,l,m,n}(y), \tag{2.2.7}$$

with $\{x^{\cdot,l,m,n}(y)\}$ being the i.i.d samples of the random variable X, given $Y = y$. Also, $H\left(\widehat{\mathcal{Z}}_0^{c,1,\cdots}(y)\right) \equiv 0$. Now under the assumptions (2.2.1), it can be proved that [35],

$$\left| \mathbb{E}\left[H\left(\widehat{\mathcal{Z}}_{N_l}(Y)\right) - H\left(\mathbb{E}[X|Y]\right)\right] \right| = O\left(N_l^{-1}\right).$$

Further, under the assumption that there exist constants δ_0 and ρ_0 such that, $\rho(\delta) \leq \rho_0$, for all $\delta \in [0, \delta_0]$ where δ is the random variable with density ρ, the authors in [35] proved that, if X and Y are the two random variables, satisfying the stated assumption, then,

$$\mathrm{Var}\left[H\left(\widehat{\mathcal{Z}}_N(Y)\right) - H\left(\mathbb{E}[X|Y]\right)\right] = O(N^{-1/2}). \tag{2.2.8}$$

The above result determines the strong convergence property necessary to analyze the full potential of the MLMC estimator, in this scenario. However, if $N_l = N_0 2^l$, then with standard MLMC complexity analysis, it is easy to determine that the computational complexity required to achieve RMSE of ϵ, we need $O\left(\epsilon^{-5/2}\right)$ computational complexity. To cater to this high computational demand, even in the framework of MLMC, the authors undertook the adaptive approach developed in [14] and extended it to the framework of MLMC.

Giles extended the studies carried out by authors in [14] to a multilevel paradigm with an aim to reduce the overall computational cost to $O\left(\epsilon^{-2}|\log \epsilon)|^2\right)$. In addition to the assumptions stated above, they further assume the following,

Assumption 2.2.2. *For all* $q \in (2, \infty)$,

$$\sup_y \mathbb{E}[\sigma^{-q}|X - \mathbb{E}[X|Y]|^q|Y = y] < \infty.$$

Thus, under the above-stated assumptions, it was proved in Lemma 2.5 (for the perfect adaptive sampling) and Theorem 2.7 of [35] that if the maximum number of sample paths is restricted to,

$$N = \left\lceil \max\left(O\left(\epsilon^{-1}\right), C^2\frac{\sigma^2}{|\mu|^2}\right)\right\rceil, \tag{2.2.9}$$

then, the further number of sample paths of various levels of resolutions is given by,

$$N_l = \left\lceil N_0 4^l \max\left(2^{-l}, \min\left(1, \left(C^{-1}N_0^{1/2}2^l\frac{|\mu|}{\sigma}\right)^{-r}\right)\right)\right\rceil, \tag{2.2.10}$$

with C being some confidentiality constant and $1 < r < 2 - \dfrac{2}{q}$ for the perfect adaptive sampling and
$1 < r < 2 - \dfrac{\sqrt{4q+1}-1}{q}$ when the values of $|\mu|$ and σ is approximated. Therefore,

$$\text{Var}\left[H\left(\widehat{\mathcal{Z}}_N(Y)\right) - H\left(\mathbb{E}[X|Y]\right)\right] = O\left(2^{-l}\right), \tag{2.2.11}$$

thereby leading to the overall computational complexity of the desired order. In a detailed discussion carried out in Section 4 of [35], it was proved (pertaining to the calculation of VaR and CVaR) that in order to achieve the overall computational cost of $\mathcal{O}(\epsilon)$ RMSE, the required computational complexity is $\mathcal{O}\left(\max(\epsilon^{-2}|\log\epsilon)|, K\epsilon^{-2})\right)$ for the estimation of VaR and CVaR, respectively. The numerical test on a model problem undertaken shows the efficacy of the algorithm constructed. Readers are directed to the referred paper for a detailed discussion of the proofs of the above-stated results. It may be noted that the computational complexity increases with an increase in the portfolio size, K. A random sub-sampling approach, extending it to a multilevel framework, thereby addressing the dependency on the portfolio size, to achieve the desired RMSE was introduced in [33].

2.2.2 Estimation of Probabilities

A study conducted in [42] undertakes a more general problem of developing an efficient numerical method in order to estimate,

$$\mathbb{P}[X \in \Omega] = \mathbb{E}\left[\mathbf{1}_{\{X \in \Omega\}}\right], \tag{2.2.12}$$

for a given tolerance level ϵ. The study performed by the authors goes beyond the applications related to the calculation of the nested expectations. We, however, focus our discussion to the applications pertaining to the estimation of risk. In the study carried out for the one-dimensional setting, the authors consider the problem of estimating the $\mathbb{P}[x > 0] = \mathbb{E}[H(x)]$. In doing so, they have considered an increasingly accurate approximation $\{x_l\}_{l \in \mathbb{N}}$ converging to x, almost surely, as $l \to \infty$. The algorithm proposed in [42] is thus the generalization of the study performed in [35] as discussed in the Subsection 2.2.1. As before, the idea is to approximate $\mathbb{P}[x > 0]$ with $\mathbb{P}[x_L > 0]$, with the choice of L being large enough to control the bias generated by the approximation. Therefore, adapting the idea of the multilevel estimator, the corresponding estimator is,

$$\mathbb{E}\left[H(x)\right] \approx \mathbb{E}\left[H(x_L)\right] = \sum_{l=1}^{L} \mathbb{E}\left[H(x_l) - H(x_{l-1})\right] \approx \sum_{l=1}^{L} \frac{1}{M_l} \sum_{m=1}^{M_l} \left(H\left(x_l^{f,m}\right) - H\left(x_{l-1}^{c,m}\right)\right) \tag{2.2.13}$$

The presented study requires less restrictive conditions than the one discussed in Subsection 2.2.1. Before discussing this further, we put forth the underlying assumption undertaken by the authors and briefly review the consequence of the stated assumption [42]. To begin with, for some $q > 2$, $\beta > 0$ and positive valued random variable σ_l, define,

$$Z_l := \frac{x_l - x}{\sigma_l 2^{-\beta l/2}}, \tag{2.2.14}$$

assuming $\mathbb{E}|Z_l|^q$ is uniformly bounded in $l \geq 0$. Further, let $\delta_l := \dfrac{x_l}{\sigma_l}$.

Assumption 2.2.3. *There exists $\delta > 0, \rho_0 > 0$ such that for all $0 < a \leq \delta$, we have,*

$$\mathbb{P}\left[|\delta_l| < a\right] \leq \rho_0 a,$$

for all $l \geq 0$.

Under the assumptions $\mathbb{E}|Z_l|^q$ is uniformly bounded and Assumption 2.2.3, it was

proven in Proposition 2.3 of [42] that,

$$\mathbb{E}\left[(H(x) - H(x_l))^2\right] \leq c2^{-\left(\frac{q}{q+1}\right)l\beta},$$

which consequently also act as the bound for $\mathbb{E}\left[|H(x_l) - H(x_{l-1})|\right]$. Under the same assumptions, we require that $\beta > 2\left(\frac{q+1}{q}\right)\gamma$, in order to observe ϵ^{-2} computational complexity. Since in most applications (including the one under consideration here), $\beta \leq 2\gamma$ [42]. Therefore, a tighter bound for bias is essential to accurately determine computational complexity. This is achieved by further incorporating the assumption (2.2.1) and also assuming $|\mathbb{E}[|Z_l|]| \leq c_2 2^{l(\beta/2-\alpha)}$, for $\beta/2 \leq \alpha \leq \beta$. Then $\mathbb{E}\left[(H(x) - H(x_l))^2\right] \leq c2^{-l\beta/2}$, and $\mathbb{E}\left[|H(x_l) - H(x_{l-1})|\right] \leq c_3 2^{-\alpha l}$. With these assumptions and bounds, it was shown that the computational complexity C is bounded as described below,

$$C \leq \begin{cases} c_4\epsilon^{-2}, \ \beta > 2\gamma, \\ c_4\epsilon^{-2}\left(\log\epsilon\right)^2, \ \beta = 2\gamma, \\ c_4\epsilon^{-2-\frac{(\gamma-\beta/2)}{\alpha}}, \ 0 < \beta < 2\gamma. \end{cases} \tag{2.2.15}$$

From the above equation, it is quite evident that the discontinuity of the function $H(x)$ affects the computational complexity of the MLMC estimator. The idea proposed in [42] is to use $x_{l+\nu_l}$, (where ν_l is a random, non-negative integer), to approximate x on level l. Further, the refinement is performed between levels $l \leq l + \nu_l \leq l + \lceil \Xi l \rceil$ based on based on $\delta_{l+\nu_l}$, where Ξ is a supplied parameter. The procedure to perform adaptive sampling at level l is given in Algorithm 4, wherein the parameter r determines the strictness of the refinement. Under ideal conditions, a large value of r is desirable in order to observe the maximum benefit for the MLMC complexity. Also, it is important that the refining procedure does not affect the almost sure convergence of $x_{l+\nu_l}$ to x. Further, it is assumed that cost of computing σ_{l+k} is of the order $2^{\gamma(l+k)}$. A comprehensive work analysis carried out in Section 3 of [42] shows the potential of the above algorithm to achieve the computational cost comparable to the standard MLMC, albeit under some additional assumptions. In the context of the application under consideration, *i.e.,* the calculation of the VaR, we take $x = \mathbb{E}[X|Y]$ and consequently, $x_l = \frac{1}{N_l}\sum_{k=1}^{N_l} X^k(Y)$, where, $X^k(Y) \overset{\text{i.i.d}}{\sim} X|Y$ and $N_l = 2^{\gamma^l}$. Using algorithm 4, the refining procedure form level $l+k$ to $l+k+1$ is carried out by adding $(2^\gamma - 1)N_{l+k}$ samples to the already existing N_{l+k}

samples on the level $l+k$. Using the variance of x_l, as the approximation of σ_l^2, on level l, we can write,

$$Z_l = \sqrt{\frac{N_0 N_l}{N_l - 1}} T_{N_l} \tag{2.2.16}$$

where, T_{N_l} is Student's t-statistic with samples $\{X^k(Y) - \mathbb{E}[X|Y]\}_{k=1}^{N_l}$. A thorough analysis undertaken in [42, Section 4], shows how the above procedure satisfies the underlying assumptions for the MLMC computations stated above. Readers can refer to the article for a deeper understanding of the underlying mathematics and proofs.

In conclusion, we would like to highlight specific differences between the algorithm presented above and the one discussed in Subsection 2.2.1 in the refinement procedure. Firstly, the algorithm studied by Giles is tailor-made to improve the computational cost of nested expectation. In contrast, the approach discussed above is more general and is also applicable to the problem of derivative pricing. Unlike the algorithm developed by Giles in [35], which requires generating samples of x_{l+k+1}, independent of x_{l+k}, in the above algorithm, the samples generated for the computation of x_{l+k} can be reused in the refinement to x_{l+k+1}. This allows for acceleration in the refinement procedure. Further, one can observe that the Algorithm 3 returns the number of samples required for the computation of x_{l+k} contrary to the Algorithm 4, which returns the estimate of x as the output of the refinement process. However, both these algorithms suffer from large kurtosis, as all even moments of $H(x_l) - H(x_{l-1})$ are equal, which in turn impacts the robustness of the MLMC algorithm. Therefore, exploring ideas to deal with large kurtosis to obtain a reliable estimate of bias and variance on level l is still an open problem.

2.3 Numerical Illustrations

In this section, we present numerical examples both in the context of option pricing and risk management to demonstrate the efficacy and practicality of the algorithms discussed. As we aim to present the applicability of the discussed algorithm, we intend to keep things simple, working preferably in a single dimension.

Algorithm 3: Adaptive algorithm to deduce N_l [35]

Data: l, y, N_0, r
Result: N_l
Set $N_l := N_0 2^l$;
Set done := False;
while *done* $\neq$ *True* **do**
  if $2N_l \geq N_0 4^l$ **then**
    $N_l \leftarrow N_0 4^l$;
    done $\leftarrow$ *True*;
  end
  else
    generate new, independent inner N_l samples;
    calculate the approximate $|\mu|$ and σ^2;
    given $Y = y$ **if** $N_l \geq N_0 4^l \left(C^{-1} N_0^{1/2} 2^l \frac{|\tilde{\mu}|}{\tilde{\sigma}} \right)^{-r}$ **then**
      done $\leftarrow$ *True*;
    end
    else
      $N_l \leftarrow 2N_l$;
    end
  end
end

Algorithm 4: Adaptive Sampling at level l [42]

Data: $l, r, \Xi, c^* > 0, \gamma, \beta$
Result: Adaptively refined sample $x_{l+\nu_l}$
Set $k = 0$;
Sample (x_l, σ_l);
while $|\delta_{l+k}| < c^* \times 2^{\gamma(\Xi l(1-r)-k)/r}$ *and* $k < \lceil \Xi l \rceil$ **do**
  $(x_{l+k+1}, \sigma_{l+k+1}) \xleftarrow{Refine} (x_{l+k}, \sigma_{l+k})$;
  Compute δ_{l+k+1} given $(x_{l+k+1}, \sigma_{l+k+1})$;
  $k \leftarrow k + 1$;
end
$\nu_l \leftarrow k$;
return $x_{l+\nu_l}$

2.3.1 Option Pricing

Consider a stochastic process $(X_t)_{0\leq t\leq T}$ governed by the one dimensional gBm *i.e.*,

$$dX_t = rX_t dt + \sigma X_t dW_t \tag{2.3.1}$$

where r denotes the risk-free interest rate, σ denotes the volatility, and $(W_t)_{0\leq t\leq T}$ is a one-dimensional standard Brownian motion defined on the probability space $(\Omega, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$. Now, under the probability space $\mathbb{P}_\theta$, the above SDE becomes,

$$dX_t(\theta) = (r + \sigma\theta)X_t(\theta)dt + \sigma X_t(\theta)dW_t., \tag{2.3.2}$$

Further, we intend to estimate the value of the European call option, with payoff the function,

$$G(X_T) = e^{-rT}(X_T - K)_+, \tag{2.3.3}$$

where K is the strike price. In order to perform simulation, we set $X_0 = 80$, $r = 0.06$, $\sigma = 0.4$, $T = 1$ and $K = 100$. We use the Euler discretization scheme in order to simulate the SDE, *i.e.,*

$$X_{n+1} = X_n + rX_n\Delta t_n + \sigma X_n \Delta W_n \tag{2.3.4}$$

where $\Delta t_n = t_{n+1} - t_n$ and $\Delta W_n = W_{n+1} - W_n$. Similarly, for the parametric SDE, we have

$$X_{n+1}(\theta) = X_n(\theta) + (r + \sigma\theta)X_n(\theta)\Delta t_n + \sigma X_n(\theta)\Delta W_n \tag{2.3.5}$$

With all the preludes, we present the numerical results to showcase the efficacy of the two importance sampling algorithms. We have used the formulas from [60] in order to estimate the number of levels and samples per level for the MLMC implementation. Also, we have taken $\Theta = [0, 1]$ and $\gamma_n = \dfrac{1}{n}$ for performing the stochastic approximation, whereas for the deterministic approximation, we have used the formulas from [52] in order to approximate θ. From the practical point of view, we stop both algorithms after a finite number of iterations. For most cases, the number of levels required for the MLMC implementation does not exceed six. Therefore, pre-computing θ for these levels for a given discretization parameter can accelerate the algorithm substantially. The interested reader may refer to `https://bitbucket.org/pefarrell/pymlmc/src/master/` for the implementation of MLMC in option pricing. Further, necessary changes with respect to importance sampling can be incorporated into the suggested package. We present the

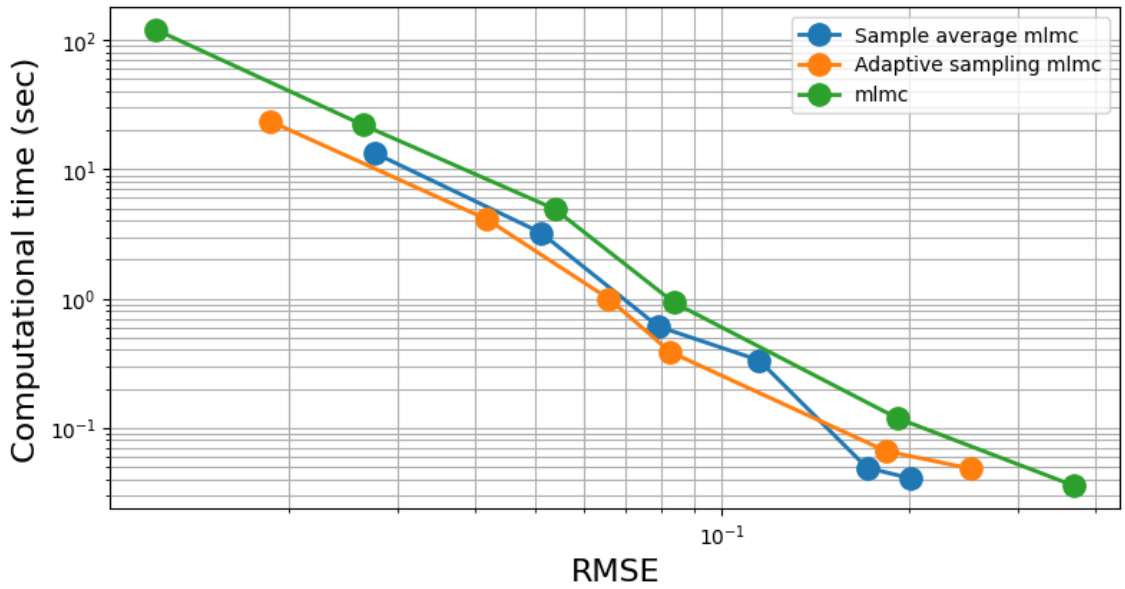


Figure 2.1: Implementation of the algorithms for option pricing

numerical results in Figure 2.1.

2.3.2 Risk Estimation

In this section, we compare the two algorithms over a model problem studied in [35, 42]. Here, we seek to estimate,

$$\varrho = \mathbb{E}\left[H\left(\mathbb{E}[P(Y,Z)] - \mathbb{E}[P(Y_0,Z|Y_0)] - L_\varrho\right)\right], \quad Y, Y_0, Z \overset{i.i.d}{\sim} \mathcal{N}(0,1), \tag{2.3.6}$$

where,

$$P(y,z) = -\tau y^2 - 2\tau^{1/2}(1-\tau)^{1/2}yz - (1-\tau)z^2.$$

The above problem describes a model for a delta-hedged portfolio with a negative Gamma so that the probability of occurrence of a huge loss is very low. Based on the small study performed in Section 4.1 of [35], the above problem can be further modified as $\mathbb{E}[H(\mathbb{E}[X|Y])]$, where

$$X = \tau(Y_0^2 - Y^2) + 2\tau^{1/2}(1-\tau)^{1/2}Y_1 Z - L_\varrho$$

In order to perform our numerical experimentation, we set $\tau = 0.015$ and $L_\varrho = 0.0805$. For the above values $\mathbb{E}[H(\mathbb{E}[X|Y])] \approx 0.0115$. We present in Figure 2.2 a visual com-

parison of the two algorithms discussed concerning risk management. We compare the total number of samples *i.e.*, samples used in the MLMC estimator plus the samples for the adaptive algorithm, for a desired tolerance level. Figure 2.2 gives a pictorial

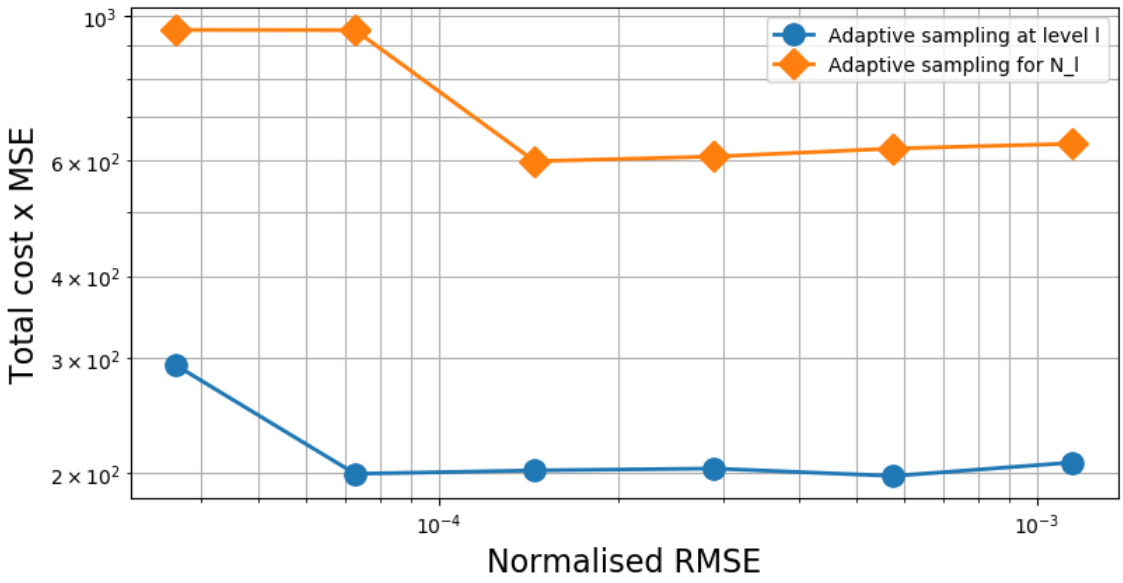


Figure 2.2: Implementation of algorithm for risk estimation

justification to the discussion carried out in Section 2.2. The readers can refer to `https://github.com/JSpence97/mlmc-for-probabilities` for the software package developed by the authors in [42] to implement the algorithms discussed with respect to risk management.

2.4 Summary

In this chapter, we gave a brief overview of the recent trends in the paradigm of the multilevel algorithm concerning the importance sampling in the case of option pricing and an adaptive sampling approach while determining VaR and CVaR for large portfolios. The algorithms discussed serve to improve the computational efficiency of the standard multilevel estimators, each with its merits and shortcomings. As mentioned in Section 2.1, the importance sampling algorithm combined with multilevel estimators significantly decreases variance at various resolution levels. However, this decrease in variance comes at the cost of increased computational complexity in either case and an increase in the sensitivity to approximate the optimal parameter. As for developing MLMC based algorithm for efficient risk estimation discussed in Section 2.2, the adaptive sampling

approach introduced in this paradigm leads to a significant improvement in the overall computational complexity to achieve the desired RMSE. The algorithm discussed in Subsection 2.2.1 suffers from the dependency on the portfolio size, which is curtailed in the algorithm discussed in Subsection 2.2.2. However, both these algorithms suffer from large kurtosis, impacting the robustness of the MLMC estimator. Overall, the presented ideas have substantially contributed to the research and development of the multilevel algorithm for various applications encountered in financial engineering problems.

Chapter 3

Multilevel Richardson-Romberg Extrapolation and Adaptive Importance Sampling for Efficient Simulation

3.1 Introduction

In Chapter 2, we observed how integrating importance sampling in the multilevel framework could lead to a substantial variance reduction while pricing options. These studies capitalize on the formulation of the asymptotic variance that appears in the Central Limit Theorem (CLT) for developing algorithmic procedures to estimate the optimal change of measure parameter. Since the asymptotic results in [4] have been studied in the context of the Euler-Maruyama discretization, therefore the adaptive importance sampling procedure discussed in [2, 52] and in Chapter 2 is limited to its application in the Euler-Maruyama discretization setup. The authors in [39] undertook the asymptotic analysis of the generalized multilevel setup and further proved the applicability of the CLT in the context of path-dependent functionals. Therefore, leveraging these results, we examine the importance sampling algorithm in a generalized setup, extending its applicability to a larger domain of problems. In this regard, we establish the convergence result of the hybrid algorithm and investigate the efficacy of the algorithm in the option pricing domain. Assuming the asset in the option pricing problem follows an underlying SDE, we examine the effect of both Milstein and Euler-Maruyama discretization schemes in our numerical illustration. Furthermore, previous studies have not addressed the compu-

tational challenges associated with this hybrid algorithm. We explore the computational aspects of the hybrid algorithm, where we will observe that the implementation of the hybrid setup can sometimes be challenging. Although the CLTs in [39] have studied both in the context of MLMC and ML2R, we focus our attention on the ML2R setup as, under some circumstances, ML2R intend to perform better than MLMC [60], which we will review in subsection 3.2.

3.2 Multilevel Richardson-Romberg

In [32], authors explored the Richardson extrapolation in the context of both the MLMC and the standard Monte Carlo. As discussed, MLMC on its own performed better than the Richardson extrapolation, however, taken together, they worked even better. This approach was furthered with a comprehensive error analysis in [60]. To this end, we recall the underlying setup where we are interested in estimating the expected payoff value *i.e.,* $\mathbb{E}\left(P(X_T)\right)$, for some $T > 0$, where $(X_t)_{0\leq t\leq T}$ is a process with values in $\mathbb{R}^d$, governed by the following SDE,

$$dX_t = b(X_t)dt + \sum_{j=1}^{q} \sigma_j(X_t)dW_t^j,\ X_0 = x \in \mathbb{R}^d, \tag{3.2.1}$$

where, $W := \begin{pmatrix} W_1 & W_2 & \dots & W_q \end{pmatrix}$ is a q-dimensional Brownian motion on a filtered probability space $(\Omega, (\mathcal{F}_t)_{0\leq t\leq T}, \mathbb{P})$, with $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma_j : \mathbb{R}^d \to \mathbb{R}^d$ being the functions satisfying the assumption 2.1.1, *i.e.,*

$$\forall x, y \in \mathbb{R}^d,\ |b(x) - b(y)| + \sum_{j=1}^{q} |\sigma_j(x) - \sigma_j(y)| < K_{b,\sigma}|x - y|,\ \text{where } K_{b,\sigma} > 0, \tag{3.2.2}$$

where $|\ \cdot\ |$ denotes the usual Euclidean norm. Now, in order to estimate $\mathbb{E}\left(P(X_T)\right)$ one should be able to simulate $(X_t)_{0\leq t\leq T}$. However, except for a handful of cases, where one can devise an analytical or semi-analytical solution, we must resort to discretization schemes to perform the simulation. To begin with, let, $m, l \in \mathbb{N}$ and define $n_l := m^{l-1}$. Further, letting $h_l = \dfrac{T}{n_l}$ as the time step size, we simulate the SDE using some discretization scheme. For instance, using the Euler-Maruyama scheme, we get,

$$X_{n+1}^{n_l} = X_n^{n_l} + b\left(X_n^{n_l}\right) h_l + \sum_{j=1}^{q} \sigma_j(X_n^{n_l}) \Delta W_n^{j,n_l},$$

where, $\Delta W_n^{\cdot,n_l} \sim N(0, \sqrt{h_l})$. Finally, we approximate $\mathbb{E}\left[P\left(X_T\right)\right]$ by $\mathbb{E}\left[P\left(X_T^{n_L}\right)\right]$, for some $L > 0 \in \mathbb{N}$. In [60], the authors proposed a methodology combining the order bias cancellation of Richardson Romberg extrapolation with variance control of MLMC to solve the problem of improving the computational complexity, along with determining the optimal parameters in order to achieve the desired RMSE, with minimum computational effort. This new estimator is popularly known as ML2R. Below, we recall the necessary assumptions that facilitated the construction of the ML2R estimator in [60] and are also necessary for our analysis.

To being with, recall that $\mathfrak{B} \subset (0, +\infty)$ is the set bias parameter such that $\mathfrak{B} \cup \{0\}$ is a compact set and,

$$\frac{\mathfrak{B}}{n} \subset \mathfrak{B}, \quad \forall n \in \mathbb{N}.$$

Further, let $\mathbf{h} \in \mathfrak{B}$, where $\mathbf{h}$ denotes the bias parameter of the coarest level. We now imposed the following assumption,

Assumption 3.2.1 (*Weak Error*). *There exists constants $\alpha > 0$, $\bar{L} \geq 1$ and $(c_l)_{1 \leq l \leq \bar{L}}$ such that,*

$$\mathbb{E}[P(X_T^h)] - \mathbb{E}[P(X_T^0)] = \sum_{k=1}^{\bar{L}} c_k h^{\alpha k} + o(h)$$

Assumption 3.2.2 (*Strong Error*). *There exist constants $\beta > 0$ and $\mathcal{V}_1 \geq 1$, such that,*

$$\|P(X_T^h) - P(X_T^0)\|_2^2 = \mathbb{E}\left[|P(X_T^h) - P(X_T^0)|^2\right] \leq \mathcal{V}_1 h^{\beta}.$$

With the consideration of the above assumptions, the ML2R estimator is defined as,

$$\mathcal{J}_\pi^N := \mathbb{E}\left[P\left(X_T^{n_L}\right)\right] = \frac{1}{N_1} \sum_{k=1}^{N_1} P\left(X_T^{n_1,k}\right) + \sum_{l=2}^{L} \frac{\widetilde{W}_l}{N_l} \sum_{k=1}^{N_l} \left(P\left(X_T^{n_l,k}\right) - P\left(X_T^{n_{l-1},k}\right)\right), \tag{3.2.3}$$

where $\pi = (h, \mu, L)$ (refer Table 3.1) are the optimal parameters obtained as the solution to,

$$(\pi(\epsilon), N(\epsilon)) = \arg\min_{\|\mathcal{J}_\pi^N - \mathcal{J}_0\|_2 \leq \epsilon} \text{Cost}(\mathcal{J}_\pi^N). \tag{3.2.4}$$

where the cost function is given by [39],

$$\text{Cost}(\mathcal{J}_\pi^N) = \frac{N}{h}\sum_{l=1}^{L}\mu_l(n_{l-1}+n_l).$$

Further, in the above equation, N_l is the number of sample paths on level l, and $\widetilde{W}_l$ are the weights given by,

$$\widetilde{W}_l = \sum_{j=l}^{L} w_l,\ l = 1,\dots,L, \tag{3.2.5}$$

where $\mathbf{w} = (w_l)_{1\le l\le L}$ is the solution to the Vandermonde system $V\mathbf{w} = e_1$, with the Vandermonde matrix being defined by,

$$V = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & n_2^{-\alpha} & \dots & n_L^{-\alpha} \\ \vdots & \vdots & \dots & \vdots \\ 1 & n_2^{-\alpha(L-1)} & \dots & n_L^{-\alpha(L-1)} \end{pmatrix} \tag{3.2.6}$$

Here, α is the weak error rate as defined above. The interested reader may refer to [60] for the construction of the optimal parameters, as the closed solution to equation (3.2.4). Here, we tabulate the explicit values of these parameters required to achieve the RMSE ϵ, with the following constants being used:

(A) $\lambda = \sqrt{\dfrac{\mathcal{V}_1}{\text{Var}\left(P(X_T^0)\right)}}$ and $\widetilde{c}_\infty = \lim\limits_{L\to\infty}|c_L|^{1/\alpha} \in (0,\infty)$.

(B) $\underline{C}_{m,\beta} = \dfrac{1+m^{\beta/2}}{\sqrt{1+m^{-1}}}$ and $\overline{C}_{m,\beta} = \left(1+m^{\beta/2}\right)\sqrt{1+m^{-1}}$.

m	$m_l = m^{l-1}, \quad l = 1, \ldots, L$
$L(\epsilon)$	$\left\lceil \frac{1}{2} + \frac{\log(\widetilde{c}_\infty^{1/\alpha}\mathbf{h})}{\log(m)} + \sqrt{\left(\frac{1}{2} + \frac{\log(\widetilde{c}_\infty^{1/\alpha}\mathbf{h})}{\log(m)}\right)^2 + \frac{2\log(A/\epsilon)}{\alpha\log(m)}} \right\rceil$
$h(\epsilon)$	$\frac{\mathbf{h}}{\left\lceil \mathbf{h}(1+2\alpha L)^{\frac{1}{2\alpha L}} \widetilde{c}_\infty^{1/\alpha} \epsilon^{-1/(\alpha L)} m^{-(L-1)/2} \right\rceil}$
$\mu(\epsilon)$	$\mu_1 = q^*(1+\lambda h^{\beta/2})$ $\mu_l = q^* \lambda h^{\beta/2} \underline{C}_{m,\beta} \lvert\widetilde{W}_l(L,m)\rvert \, m^{-(1-\beta)(l-1)/2}, \quad l = 2, \ldots, L$ with q^* s.t. $\sum_{l=1}^{L} \mu_l = 1$
$N(\epsilon)$	$\left(1 + \frac{1}{2\alpha L}\right) \frac{\mathrm{Var}(P(X_T^0)) \left(1 + \lambda h^{\beta/2} + \lambda h^{\beta/2} \overline{C}_{m,\beta} \sum_{l=2}^{L} \lvert\widetilde{W}_l(L,m)\rvert m^{(1-\beta)(l-1)/2}\right)}{\epsilon^2 q^*}$

Table 3.1: Optimal parameters for the ML2R estimator [39]

As discussed in Chapter 1, the estimator (3.2.3) is highly effective whenever the strong order of convergence *i.e.,* $\beta \leq 1$, as it achieves $\mathcal{O}\left(\epsilon^{-2}\log(1/\epsilon)\right)$ for $\beta = 1$ and $\mathcal{O}\left(\epsilon^{-2} e^{\frac{1-\beta}{\sqrt{\alpha}}\sqrt{2\log(1/\epsilon)\log(M)}}\right)$ for $\beta < 1$, contrary to $\mathcal{O}\left(\epsilon^{-2}\log(1/\epsilon)^2\right)$ and $\mathcal{O}\left(\epsilon^{-2-\frac{1-\beta}{\alpha}}\right)$, respectively, which is achieved by the standard MLMC. This improvement in the computational cost inspires us to extend the importance sampling framework to the ML2R setup.

Finally we recall the Central Limit Theorems for the ML2R estimator proved in [39], as it forms the backbone for the construction of the importance sampling estimator. For the sake of brevity, we let,

$$Y(h) \coloneqq \left(\frac{h}{m}\right)^{\frac{-\beta}{2}} \left(P\left(X_T^{h/m}\right) - P\left(X_T^h\right)\right) \text{ and } Y_l \coloneqq Y\left(\frac{\mathbf{h}}{m^{l-2}}\right).$$

Further, let,

$$Z_l = P\left(X_T^{h/m^{l-1}}\right) - P\left(X_T^{h/m^{l-2}}\right) \text{ and } Z_1 = Y(\mathbf{h}) = P(X_T^1).$$

Theorem 3.2.1 (Central Limit Theorem, $\beta > 1$)**.** *Suppose assumption 3.2.2 holds with $\beta > 1$ and that $(Y(h))_{h\in\mathfrak{B}}$ is $\mathcal{L}^2-$uniformly integrable. Let,*

$$\overline{\sigma}_1^2 = \frac{1}{\Sigma}\frac{Var(Y_{\mathbf{h}})}{Var(Y_0)\left(1+\lambda\mathbf{h}^{\frac{\beta}{2}}\right)} \text{ and } \overline{\sigma}_2^2 = \frac{1}{\Sigma}\frac{\mathbf{h}^{\frac{\beta}{2}}\sum\limits_{l\geq 2} m^{\frac{1-\beta}{2}(l-1)}Var(Y_l)}{\sqrt{Var(Y_0)\mathcal{V}_1}\underline{C}_{m,\beta}},$$

with,

$$\Sigma = \left[1+\lambda\mathbf{h}^{\frac{\beta}{2}}\left(1+\overline{C}_{m,\beta}\frac{m^{\frac{1-\beta}{2}}}{1-m^{\frac{1-\beta}{2}}}\right)\right],\ \underline{C}_{m,\beta} = \frac{1+m^{\frac{\beta}{2}}}{\sqrt{1+m^{-1}}},\ \text{and}\ \overline{C}_{m,\beta} = \left(1+m^{\frac{\beta}{2}}\right)\sqrt{1+m^{-1}}.$$

If assumption 3.2.1 holds for $\bar{L} \geq 1$, we have,

$$\frac{\mathcal{J}_\pi^N(\epsilon) - \mathcal{J}_0}{\epsilon} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \overline{\sigma}_1^2 + \overline{\sigma}_2^2),\ \text{as}\ \epsilon \to 0.$$

Proof. Refer to [39] □

Theorem 3.2.2 (Central Limit Theorem, $0 < \beta \leq 1$)**.** *Suppose assumption 3.2.2 hold with $\beta \in (0,1]$ and that $(Y(h))_{h\in\mathfrak{B}}$ is $\mathcal{L}^2-$uniformly integrable. Further assume that, $\lim\limits_{h\to 0}\|Y(h)\|_2^2 = v_\infty(m,\beta)$, and let*

$$\overline{\sigma}^2 = \begin{cases} v_\infty(m,\beta)\left(1+m^{\beta/2}\right)^{-2}\mathcal{V}_1^{-1}, & \text{if } 2\alpha > \beta, \\ \left(v_\infty(m,\beta) - c_1^2(1-m^{\beta/2})^2\right)\left(1+m^{\beta/2}\right)^{-2}\mathcal{V}_1^{-1}, & \text{if } 2\alpha = \beta. \end{cases}$$

If assumption 3.2.1 holds for $\bar{L} \geq 1$, we have,

$$\frac{\mathcal{J}_\pi^N(\epsilon) - \mathcal{J}_0}{\epsilon} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \overline{\sigma}^2),\ \text{as}\ \epsilon \to 0.$$

Proof. Refer to [39] □

3.3 Adaptive Importance Sampling ML2R Algorithm

3.3.1 Setup

Following the same idea discussed in section 2.1 of Chapter 2, we consider a parametric family of stochastic process $(X_t(\theta))_{0\le t\le T}$, with $\theta \in \mathbb{R}^d$, governed by the following SDE,

$$dX_t(\theta) = (b(X_t(\theta)) + \sigma(X_t(\theta))\theta)dt + \sum_{j=1}^{q} \sigma_j(X_t(\theta))dW_t^j,\ \sigma(x) = \begin{pmatrix} \sigma_1(x) & \dots & \sigma_q(x) \end{pmatrix}. \tag{3.3.1}$$

By Girsanov's Theorem, we know that there exists a probability measure $\mathbb{P}_\theta$ equivalent to $\mathbb{P}$ such that,

$$\left.\frac{d\mathbb{P}_\theta}{d\mathbb{P}}\right|_{\mathcal{F}_t} = \exp\left(-\langle\theta, W_t\rangle - \frac{1}{2}|\theta|^2 t\right) \triangleq \mathcal{I}^-(W_t, \theta), \tag{3.3.2}$$

under which, the process $(\theta t + W_t)_{0\le t\le T}$ is a Brownian motion. Therefore,

$$\mathbb{E}_{\mathbb{P}}\left[P(X_T)\right] = \mathbb{E}_{\mathbb{P}_\theta}\left[P(X_T(\theta))\right] = \mathbb{E}_{\mathbb{P}}\left[P(X_T(\theta))\mathcal{I}^-(W_T, \theta)\right]. \tag{3.3.3}$$

For example,

Example 3.3.1. *Suppose, we aim to estimate the expectation*

$$\mathbb{E}_{\mathbb{P}}\left[P(X_T)\right], \quad \textit{where} \quad P(X_T) = 1_{\{X_T>0\}}$$

and X_T follows the SDE,

$$dX_t = dW_t, \quad X_0 = 0,$$

and $(W_t)_{t\ge0}$ is a standard Brownian Motion. Following the parametric version we have,

$$dX_t(\theta) = \theta\, dt + dW_t, \quad X_0 = 0,$$

where $\theta \in \mathbb{R}$. Now as a consequence of Girsanov's theorem, we have,

$$\mathbb{E}_{\mathbb{P}}\left[P(X_T)\right] = \mathbb{E}_{\mathbb{P}}\left[P(X_T(\theta))e^{-\langle W_T,\theta\rangle - \frac{1}{2}|\theta|^2T}\right] = \mathbb{E}_{\mathbb{P}}\left[1_{\{X_T(\theta)>0\}}e^{-\langle W_T,\theta\rangle - \frac{1}{2}|\theta|^2T}\right].$$

3.3.2 Algorithm

As usual, we want to estimate $\mathbb{E}_{\mathbb{P}}[P(X_T)] \approx \mathbb{E}_{\mathbb{P}}[P(X_T^{n_L})]$. In the standard ML2R setup, we have

$$\mathbb{E}_{\mathbb{P}}[P(X_T^{n_L})] = \mathbb{E}_{\mathbb{P}}[P(X_T^{n_1})] + \sum_{l=2}^{L} \widetilde{W}_l \mathbb{E}_{\mathbb{P}}[P(X_T^{n_l}) - P(X_T^{n_{l-1}})] \tag{3.3.4}$$

In the light of equation (3.3.3), we have,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}\left[P\left(X_T^{n_L}\right)\right] &= \mathbb{E}_{\mathbb{P}}\left[P(X_T^{n_1}(\theta_1))\mathcal{I}^{-}(W_T^1, \theta_1)\right] \\ &+ \sum_{l=2}^{L} \widetilde{W}_l \mathbb{E}_{\mathbb{P}}\left[\left(P\left(X_T^{n_l}(\theta_l)\right) - P\left(X_T^{n_{l-1}}(\theta_l)\right)\right)\mathcal{I}^{-}(W_T^l, \theta_l)\right]. \end{aligned} \tag{3.3.5}$$

By applying Monte Carlo method to each level l with N_l samples in equation (3.3.5), we get the Adaptive Importance Sampling ML2R (AISML2R),

$$\begin{aligned} \mathcal{J}_{\pi}^{N,\theta}(\theta_1, \ldots, \theta_L) &= \frac{1}{N_1} \sum_{k=1}^{N_1} P\left(X_{T,\theta_1^{k-1}}^{n_1,k}\right) \mathcal{I}^{-}(W_T^{1,k}, \theta_1^{k-1}) \\ &+ \sum_{l=2}^{L} \frac{\widetilde{W}_l}{N_l} \sum_{k=1}^{N_l} \left(P\left(X_{T,\theta_l^{k-1}}^{n_l,k}\right) - P\left(X_{T,\theta_l^{k-1}}^{n_{l-1},k})\right)\right) \mathcal{I}^{-}(W_T^{l,k}, \theta_l^{k-1}) \end{aligned} \tag{3.3.6}$$

The underlying idea of our setup is to estimate θ_l^* for each level of resolution such that the variance of that level is minimized. In this regard, we formulate and solve the optimization problem on each level to determine θ_l^*. The rest of the discussion in this section deals with formulating the optimization problem and procedure to approximate optimal solutions *i.e.*, θ_l^*.

3.3.3 Optimization Problem

Based on the Theorem 3.2.1 and Theorem 3.2.2 stated above we develop the optimization algorithm for $\beta > 1$ and $\beta \in (0, 1]$.

Case I: $\beta > 1$
In Theorem 3.2.1, let

$$\overline{\sigma}_1^2 = k_1 \mathrm{Var}(Y_{\mathbf{h}}) \text{ and } \overline{\sigma}_l^2 = k_2 m^{\frac{1-\beta}{2}(l-1)} \mathrm{Var}(Y_l),$$

where,

$$k_1 = \frac{1}{\Sigma}\frac{1}{\text{Var}(Y_0)(1+\lambda \mathbf{h}^{\frac{\beta}{2}})} \text{ and } k_2 = \frac{\mathbf{h}^{\frac{\beta}{2}}}{\Sigma\sqrt{\text{Var}(Y_0)\mathcal{V}_1}\underline{C}_{m,\beta}},$$

and therefore,

$$\overline{\sigma}_2^2 = \sum_{l\geq 2} k_l \text{Var}(Y_l) = \sum_{l\geq 2} \overline{\sigma}_l^2.$$

From the practical point of view, it is necessary to use the truncated version of the above summation. In the thorough study carried out in [39], it was proven that for $\beta > 1$,

$$\lim_{L(\epsilon)\to\infty} \sum_{l=2}^{L(\epsilon)} |\widetilde{W}_l^{L(\epsilon)}| m^{\frac{1-\beta}{2}(l-1)} \text{Var}(Y_l) = \sum_{l=2}^{\infty} m^{\frac{1-\beta}{2}(l-1)} \text{Var}(Y_l).$$

Therefore, owing to the above result and motivated by the analysis pertaining to the Central Limit Theorem carried out in [39], we can formulate the problem for $l = 1, \ldots, L(\epsilon)$ as,

$$\begin{aligned}
\theta_l^* &= \arg\min_{\theta_l \in \mathbb{R}^d} \overline{\sigma}_l^2 \\
&= \arg\min_{\theta_l \in \mathbb{R}^d} k_2 m^{\frac{1-\beta}{2}(l-1)} |\widetilde{W}_l^{L(\epsilon)}| \text{Var}\left[\left(Y_l^{\theta}\mathcal{I}^-(W_T^l, \theta_l)\right)^2\right] \\
&= \arg\min_{\theta_l \in \mathbb{R}^d} k_l \mathbb{E}\left[\left(\left(P\left(X_{T,\theta}^{h/m^{l-1}}\right) - P\left(X_{T,\theta}^{h/m^{l-2}}\right)\right)\mathcal{I}^-(W_T^l, \theta_l)\right)^2\right],
\end{aligned} \tag{3.3.7}$$

where,

$$k_l = k_2 m^{\frac{1+\beta}{2}(l-1)} |\widetilde{W}_l^{L(\epsilon)}| \mathbf{h}^{-\beta}. \tag{3.3.8}$$

Using the Girsanov's theorem, we can see that,

$$\begin{aligned}
&\mathbb{E}\left[\left(\left(P\left(X_{T,\theta}^{h/m^{l-1}}\right) - P\left(X_{T,\theta}^{h/m^{l-2}}\right)\right)\mathcal{I}^-(W_T^l, \theta_l)\right)^2\right] \\
&= \mathbb{E}\left[\left(P\left(X_T^{h/m^{l-1}}\right) - P\left(X_T^{h/m^{l-2}}\right)\right)^2 \mathcal{I}^+(W_T^l, \theta_l)\right],
\end{aligned} \tag{3.3.9}$$

where,

$$\mathcal{I}^+(W_T^l, \theta_l) = e^{-\langle W_T^l, \theta_l\rangle + \frac{1}{2}|\theta_l|^2 T}.$$

Similarly for $l = 1$, we have,

$$\theta_1^* = \arg\min_{\theta_1 \in \mathbb{R}^d} k_1 \mathbb{E}\left[\left(Y^2(\mathbf{h})\mathcal{I}^+(W_T^1, \theta_1)\right)\right]. \tag{3.3.10}$$

Case II: $\beta \in (0, 1]$

Based on Theorem 3.2.2, we define v_∞^l be the level l approximation of v_∞, where,

$$v_\infty^l = \left\| Y\left(\frac{h}{m^{l-2}}\right) \right\|_2^2.$$

Therefore, we have,

$$\overline{\sigma}_l^2 = \begin{cases} v_\infty^l(m,\beta)\left(1+m^{\beta/2}\right)^{-2}\mathcal{V}_1^{-1}, \text{ if } 2\alpha > \beta \\ \left(v_\infty^l(m,\beta) - c_1^2(1-m^{\beta/2})^2\right)\left(1+m^{\beta/2}\right)^{-2}\mathcal{V}_1^{-1}, \text{ if } 2\alpha = \beta. \end{cases} \tag{3.3.11}$$

We follow a similar line of development as for $\beta > 1$ to formulate the problem for $\beta \in (0, 1]$. As one can observe from (3.3.11), in order to minimize σ_l^2 on level l, we only need to minimize v_∞^l. Therefore, we formulate the optimization problem as follows,

$$\theta_l^* = \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}\left[Y_l^2 \mathcal{I}^+(W_T^l, \theta_l)\right]. \tag{3.3.12}$$

3.3.4 Stochastic Algorithm

In this subsection, we outline the stochastic approximation algorithm to estimate θ_l^*. Based on the discussion in the previous section, we perform the discretization, such as Euler or Milstein, of the underlying SDE to approximate Y_l's, whereas for stochastic approximation, we resort to the Robbins-Monro algorithm, to approximate the value of $\theta_l^*,\ l = 1, \ldots, L$. The aim is to construct a sequence of $(\theta_l^n)_{n\in\mathbb{N}}$, such that, $\lim_{n\to\infty} \theta_l^n = \theta_l^*$.

Let $\Theta \subset \mathbb{R}^d$ be a compact convex subset such that $\theta \in \text{int}(\Theta)$. Then the sequence is constructed recursively as follows,

$$\theta_l^{n+1} = \mathbf{Proj}_\Theta\left[\theta_l^n - \lambda_{n+1} G_l(\theta_l^n, Y_l, W_T^l)\right], \tag{3.3.13}$$

where $\mathbf{Proj}_\Theta(\theta) = \min_{\theta\in\Theta}|\theta - \theta_0|$. Also $(\lambda_n)_{n\geq 1}$ is a decreasing sequence of positive real numbers satisfying,

$$\sum_{n=1}^{\infty} \lambda_n = \infty \text{ and } \sum_{i=1}^{\infty} \lambda_n^2 < \infty. \tag{3.3.14}$$

where for $\beta > 1$,

$$G_l(\theta_l, Y_l, W_T^l) = \begin{cases} \left(\theta_l T - W_T^l\right) k_l Z_l^2 \mathcal{I}^+(W_T^l, \theta_l), & \text{for } l = 2, \ldots, L, \\ \left(\theta_1 T - W_T^1\right) k_1 Z_1^2(\mathbf{h}) \mathcal{I}^+(W_T^1, \theta_1), & \text{for } l = 1, \end{cases} \tag{3.3.15}$$

and for $\beta \in (0, 1]$,

$$G_l(\theta_l, Y_l, W_T^l) = \begin{cases} \left(\theta_l T - W_T^l\right) Y_l^2 \mathcal{I}^+(W_T^l, \theta_l), & \text{for } l = 2, \ldots, L, \\ \left(\theta_1 T - W_T^1\right) Y^2(\mathbf{h}) \mathcal{I}^+(W_T^1, \theta_1), & \text{for } l = 1. \end{cases} \tag{3.3.16}$$

It may be noted that in the present work, we use the constrained version of the Robbins-Monro algorithm to estimate the approximate value of θ_l^* for various levels of resolutions. We refer interested readers to [18] for a discussion on the Robbins-Monro algorithm.

3.4 Main Results

Below we state the main theoretical result of our discussion, the proofs of which are delayed to subsection 3.4.1.

To begin with, we state the results that establish the existence and uniqueness of optimal change of the measure parameter θ_l^* for various levels of resolution.

Theorem 3.4.1. *Suppose b and σ satisfies assumption (3.2.2). Let P be such that,*

$$\mathbb{P}(X_T \notin D_P) = 0, \text{ where } D_P = \{x \in \mathbb{R} | P \text{ is differentiable at } x\}.$$

Further, if Z_l satisfies the following conditions,

(1) $\mathbb{P}(Z_l \neq 0) > 0$.

(2) $\|Z_l^2\|_p < +\infty$ *for some* $p > 1$.

Then the function $\theta \to \mathbf{v}_l(\theta)$ is $\mathcal{C}^2$ and strictly convex with $\nabla_\theta \mathbf{v}_l(\theta) = \mathbb{E}[G_l(\theta, Z_l, W_T)]$ for all $l \in \mathbb{N}$. Moreover, there exists an unique $\theta^* \in \mathbb{R}^q$ such that $\min_{\theta \in \mathbb{R}^q} \mathbf{v}_l(\theta) = \mathbf{v}_l(\theta^*)$.

The next result establish the convergence of $\theta_l^k \overset{a.s.}{\to} \theta_l^*$ for $l = 1, \ldots, L$.

Theorem 3.4.2. *If* $\theta_l^* = \arg\min_{\theta_l \in \mathbb{R}^q} \mathbf{v}_l(\theta)$ *is such that,* $\nabla_\theta \mathbf{v}_l(\theta_l^*) = 0$ *and* $\theta_l^* \in \Theta$. *Further, assume that for all* $p \geq 2$,

$$\exists \beta > 0, \mathcal{V}_1^{(p)} > 0, \left\| P(X_T^h) - P(X_T) \right\|_p^p = \mathbb{E}\left[\left| P(X_T^h) - P(X_T) \right|^p \right] \leq \mathcal{V}_1^{(p)} h^{\beta p/2}, h \in \mathfrak{B}. \tag{3.4.1}$$

Then $\theta_l^k \xrightarrow{a.s.} \theta_l^*$, *as* $k \to \infty$.

The final result establishes the Strong Law of Large Numbers of the AISML2R estimator.

Theorem 3.4.3. *Let* $(\theta_l^k)_{k\geq 0} \subset \Theta$ *be a family of sequence converging to* $\theta_l^* \in \Theta$ *as* $k \to \infty$. *Suppose Assumption 3.2.1 holds for all* $L \geq 1$ *and for* $p \geq 2$ *assume* $P(X_T) \in \mathcal{L}^p$. *Furthermore assume the* $\mathcal{L}^p$-***strong error*** *rate assumption,i.e.,*

$$\exists \beta > 0, \mathcal{V}_1^{(p)} > 0, \left\| P(X_T^h) - P(X_T) \right\|_p^p = \mathbb{E}\left[\left| P(X_T^h) - P(X_T) \right|^p \right] \leq \mathcal{V}_1^{(p)} h^{\beta p/2}, h \in \mathfrak{B}. \tag{3.4.2}$$

If $(\epsilon_k)_{k\geq 1}$ *is a sequence of positive real numbers such that* $\sum_{k\geq 1} \epsilon_k^p < +\infty$, *then the AISML2R estimator satisfies,*

$$\mathcal{J}_\pi^{N,\theta} \xrightarrow{a.s.} \mathcal{J}_0. \tag{3.4.3}$$

3.4.1 Proof of the Main Result

We now provide detailed proof of the results stated above. We begin our discussion by recalling the results related to the weights $(\widetilde{W}_l)_{l=1,\ldots,L}$, from [39] necessary for our study. Accordingly, let us define,

$$a_l := \frac{1}{\prod_{1\leq k\leq (l-1)} (1 - m^{-k\alpha})}, \; l = 1, \ldots, L,$$

$$b_l := (-1)^l \frac{m^{\frac{-\alpha l(l+1)}{2}}}{\prod_{1\leq k\leq (l-1)} (1 - m^{-k\alpha})}, \; l = 0, \ldots, L.$$

The following result was proved in [39],

Lemma 3.4.1. *Let* $\alpha > 0$. *and the associated weights* $(\widetilde{W}_l)_{l=1,\ldots,L}$, *be as given in* (3.2.5).

(a) $\lim_{l\to+\infty} a_l = a_\infty < +\infty$ *and* $\sum_{l=0}^{+\infty} |b_l| = \widetilde{B}_\infty < +\infty$.

(b) The weights $\widetilde{W}_l$ are uniformly bounded,

$$\forall L \in \mathbb{N}^*, \forall l \in \{1, \ldots, L\},\ |\widetilde{W}_l| \leq a_\infty \widetilde{B}_\infty.$$

(c) For every $\lambda > 0$,

$$\lim_{L \to +\infty} \sum_{l=2}^{L} |\widetilde{W}_l| m^{-\lambda(l-1)} = \frac{1}{m^\lambda - 1}.$$

(d) Let $v_{j\, j \geq 1}$ be a bounded sequence of positive real numbers. Let $\lambda \in \mathbb{R}$ and assume that $\lim_{j \to +\infty} v_j = 1$ when $\lambda > 0$. Then the following limits hold:

$$\sum_{l=2}^{L} |\widetilde{W}_l| m^{\lambda(l-1)} v_l \sim \begin{cases} \sum\limits_{l \geq 2} m^{\lambda(l-1)} v_l < +\infty, \ \textit{for } \lambda < 0 \\ L, \ \textit{for } \lambda = 0, \ \textit{as } L \to +\infty \\ m^{\lambda L} a_\infty \sum\limits_{l \geq 1} |\sum\limits_{k=0}^{l-1} b_k| m^{-\lambda l}, \textit{for } \lambda > 0. \end{cases}$$

With the above results at our disposal, we present a series of results establishing the existence and uniqueness of the optimal parameter θ_l^* on various levels of resolution. For the most part, the proof follows the line of argument similar to that presented in [2, 7, 52]. We start our discussion by proving the following lemma, which is necessary for the existence and uniqueness of results.

Lemma 3.4.2. *Let $p \geq 2$ and assume the $\mathcal{L}^p$-strong error rate assumption, i.e.,*

$$\exists \beta > 0, \mathcal{V}_1^{(p)} > 0, \left\| P(X_T^h) - P(X_T) \right\|_p^p = \mathbb{E}\left[\left| P(X_T^h) - P(X_T) \right|^p \right] \leq \mathcal{V}_1^{(p)} h^{\beta p/2}, h \in \mathfrak{B}. \tag{3.4.4}$$

Then, $\mathbb{E}\left[\sup_{|\theta| \leq c} \left| G_l(\theta, Y_l, W_T) \right| \right] < \infty$ for $l = 2, \ldots, L$ and some constant $c > 0 \in \mathbb{R}$.

Proof. We prove the above lemma for $\beta > 1$. The proof for $\beta \in (0, 1]$ follows a similar line of argument and is relatively easy to prove. Consider,

$$\left| G_l(\theta, Z_l, W_T) \right| = \left| (\theta T - W_T) k_l Z_l^2 e^{-\langle \theta, W_T \rangle + \frac{1}{2}|\theta|^2 T} \right|.$$

Then, for $c > 0$,

$$\sup_{|\theta|\leq c}\left|G_l(\theta, Z_l, W_T)\right| = \sup_{|\theta|\leq c}\left|(\theta T - W_T)k_l Z_l^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}\right|,$$
$$\leq (cT + |W_T|)k_l Z_l^2 e^{c|W_T| + \frac{1}{2}|c|^2 T}.$$

Taking expectation on both sides and applying Hölder's inequality for all $p \geq 2$, we get,

$$\mathbb{E}\left[\sup_{|\theta|\leq c}\left|G_l(\theta, Y_l, W_T)\right|\right] \leq e^{\frac{c^2}{2}T}\left\|e^{c|W_T|}(cT + |W_T|)\right\|_{\frac{p}{p-1}}\left\|k_l Z_l^2\right\|_p.$$

It is clear that $\left\|e^{c|W_T|}(cT + |W_T|)\right\|_{\frac{p}{p-1}}$ is bounded. As for $\|k_l Z_l^2\|_p$ we fallback to the *$\mathcal{L}^p$-strong error* assumption. Therefore, we have,

$$\|k_l Z_l^2\|_p = k_l \|Z_l\|_{2p}^2 \leq K(m, \beta, h, p) k_l m^{-\beta(l-1)}, \tag{3.4.5}$$

where,

$$K(m, \beta, h, p) = \left(\mathcal{V}_1^{(2p)}\right)^{\frac{1}{p}}\left(1 + m^{\frac{\beta}{2}}\right)^2 h^\beta. \tag{3.4.6}$$

The boundedness of $k_l m^{-\beta(l-1)}$ can be derived from the definition of k_l and from *(b)* and *(c)* of Lemma 3.4.1. Hence, we can conclude that, $\mathbb{E}\left[\sup_{|\theta|\leq c}\left|G_l(\theta, Y_l, W_T)\right|\right]$ is bounded for $l = 2, \ldots, L$. □

We now prove the result pertaining to the existence and uniqueness of optimal parameters on various levels of resolution.

Proof of Theorem 3.4.1. To prove the proposition, we refer to the proof in [2, 52], in our context. Here we discuss the proof for $l \geq 2$ and $\beta > 1$. For $l = 1$ and $\beta \in (0, 1]$, the proofs are relatively easy and can be reproduced in a similar way. To begin with, one can observe that $\theta \to k_l(Z_l)^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}$ is a continuously infinitely differentiable function with respect to θ, and $\dfrac{\partial}{\partial\theta^j}(k_l(Z_l)^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}) = k_l(Z_l)^2(\theta^j T - W_T^j)e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}$. Therefore, the first derivative of the map $\theta \to k_l(Z_l)^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}$ is equal to $G_l(\theta, Z_l, W_T)$. As we have already seen in Lemma 3.4.2 that the $\sup_l \mathbb{E}[\sup_{|\theta|\leq c}|G_l(\theta, Z_l, W_T)|]$ is bounded, therefore by Lebesgue's theorem we can conclude that $\mathbf{v}_l(\theta)$ is $\mathcal{C}^1(\mathbb{R}^q)$, with $\nabla_\theta \mathbf{v}_l(\theta) = \mathbb{E}[G_l(\theta, Z_l, W_T)]$ for all $l \in \mathbb{N}$. A similar line of argument also proves that $\mathbf{v}_l(\theta)$ is $\mathcal{C}^2(\mathbb{R}^q)$.

The Hessian of $\mathbf{v}_l(\theta)$ is given as follows,

$$Hess(\mathbf{v}_l(\theta)) = \mathbb{E}\left[((\theta T - W_T)(\theta T - W_T)^* + TI_q)\, k_l(Z_l)^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}\right].$$

Since, $\mathbb{P}(Z_l \neq 0) > 0$, therefore, for all $u \in \mathbb{R}^q \setminus \{0\}$, $u^T Hess(\mathbf{v}_l(\theta))u > 0$. Hence, we can conclude that $\mathbf{v}_l(\theta)$ is strictly convex. As a consequence, there exists a minimum $\theta^* \in \mathbb{R}^q$, such that $\min_{\theta\in\mathbb{R}^q} \mathbf{v}_l(\theta) = \mathbf{v}_l(\theta^*)$. Further, since the unique minimum is attained for a finite value of θ, it is enough to prove that $\lim_{|\theta|\to+\infty} \mathbf{v}_l(\theta) = +\infty$. This can be proved using Fatou's lemma as,

$$+\infty = \mathbb{E}\left[\liminf_{|\theta|\to+\infty} k_l(Z_l)^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}\right] \leq \liminf_{|\theta|\to+\infty} \mathbb{E}\left[k_l(Z_l)^2 e^{-\langle\theta, W_T\rangle + \frac{1}{2}|\theta|^2 T}\right].$$

This completes the proof. □

Proof of Theorem3.4.2. To prove the above result we follow the assertion made in Theorem A.1 [57] which suggest that in order to prove $\theta_l^k \to \theta_l^*$, where the sequence $(\theta_l^k)_{k\geq 1}$ is constructed through a constrained version of the Robbins Monro, we need to verify two conditions, namely,

(1) $\forall\ \theta \neq \theta_l^*,\ \langle\nabla_\theta \mathbf{v}_l(\theta), \theta - \theta_l^*\rangle > 0.$

(2) Non explosion condition: $\exists\ C > 0$ such that $\forall\theta \in \Theta,\ \mathbb{E}\left[|G_l(\theta, Z_l, W_T)|^2\right] < C(1 + |\theta^2|)$.

As we know, $\nabla_\theta \mathbf{v}_l(\theta_l^*) = 0$ and in the previous proposition it was proven that v_l is convex, therefore as a consequence, we prove that,

$$\forall\theta \neq \theta_l^*,\ \langle\nabla_\theta \mathbf{v}_l(\theta), \theta - \theta_l^*\rangle > 0. \tag{3.4.7}$$

For the non-explosion condition, we use the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[|G_l(\theta, Z_l, W_T)|^2\right] \leq e^{|\theta|^2 T}\left(\mathbb{E}[k_l^4 Z_l^8]\right)^{\frac{1}{2}}\left(\mathbb{E}[|e^{-\langle\theta, W_T\rangle}(\theta T - W_T)|^4]\right)^{\frac{1}{2}}. \tag{3.4.8}$$

Under the assumption and following the similar line of argument of Lemma 3.4.2, it is easy to prove that there exists a constant $C > 0$, such that,

$$\mathbb{E}\left[|G_l(\theta, Z_l, W_T)|^2\right] \leq e^{|\theta|^2 T} C\left(\mathbb{E}[|e^{-\langle\theta, W_T\rangle}(\theta T - W_T)|^4]\right)^{\frac{1}{2}}. \tag{3.4.9}$$

Further using the fact the $\theta \in \Theta$, we can deduce that,

$$\sup_{\theta\in\Theta} \mathbb{E}\left[|G_l(\theta, Z_l, W_T)|^2\right] < \infty, \tag{3.4.10}$$

, which in turn concludes the non-explosion condition. This proves the almost sure convergence of θ_l^k to θ_l^* as $k \to \infty$. □

We now prove the Strong Law of Large Numbers *i.e.* Theorem 3.4.3. To this end, we will assume the following notation,

$$\widetilde{\mathcal{J}}^1_{\theta,\pi} := \frac{1}{N_1}\sum_{k=1}^{N_1} P(X^{n_1,k}_{T,\theta_1^{k-1}})\mathcal{I}^-(W_T^{1,k}, \theta_1^{1,k-1}) - \mathbb{E}\left[P(X_T^{n_1})\right] \text{ and } \widetilde{\mathcal{J}}^2_{\theta,\pi} := \sum_{l=2}^{L}\frac{\widetilde{W}_l}{N_l}\sum_{k=1}^{N_l}\widetilde{Y}^k_{l,\theta_l^{k-1}},$$

where we set,

$$\widetilde{Y}_{l,\theta_l} = \left(P(X^{n_l}_{T,\theta_l}) - P(X^{n_{l-1}}_{T,\theta_l})\right)\mathcal{I}^-(W_T^l, \theta_l) - \mathbb{E}\left[P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right], \tag{3.4.11}$$

and

$$\widetilde{Y}_{1,\theta_1} = \left(P(X^{n_1}_{T,\theta_1})\right)\mathcal{I}^-(W_T^1, \theta_1) - \mathbb{E}\left[P(X_T^{n_1})\right]. \tag{3.4.12}$$

Therefore, we have,

$$\mathcal{J}_\pi^{N,\theta} - \mathcal{J}_0 = \widetilde{\mathcal{J}}^1_{\theta,\pi} + \widetilde{\mathcal{J}}^2_{\theta,\pi} + \mathbb{E}[P(X_T^L)] - \mathcal{J}_0.$$

A thorough analysis carried out in section 4.2 of [39] shows that the last term in the equation converges to zero as $\epsilon \to \infty$. We start our discussion by proving the following lemma.

Lemma 3.4.3. *Let $p \geq 2$ and $|\theta| \leq c$. Then there exist a positive constant $K_1(m, \beta, p, c)$ such that,*

$$\|\widetilde{Y}_{l,\theta}\|_p^p \leq K_1(m, \beta, p, c)m^{-\beta p(l-1)/2},\ l = 2, \ldots, L.$$

Proof. By Minkowski's Inequality, we have,

$$\begin{aligned}\left(\mathbb{E}\left[|\widetilde{Y}_{l,\theta})|^p\right]\right)^{1/p} &\leq \left\|\left(P(X_{T,\theta}^{n_l}) - P(X_{T,\theta}^{n_{l-1}})\right)\mathcal{I}^-(W_T^l,\theta)\right\|_p + \left|\mathbb{E}\left[P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right]\right|, \\ &\leq \underbrace{\left\|\left(P(X_{T,\theta}^{n_l}) - P(X_{T,\theta}^{n_{l-1}})\right)\mathcal{I}^-(W_T^l,\theta)\right\|_p}_{\text{I}} + \underbrace{\left\|\left[P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right]\right\|_p}_{\text{II}}.\end{aligned} \tag{3.4.13}$$

In order to bound (I) of (3.4.13), we apply the Holder's Inequality,

$$\begin{aligned}\left\|\left(P(X_{T,\theta}^{n_l}) - P(X_{T,\theta}^{n_{l-1}})\right)\mathcal{I}^-(W_T^l,\theta)\right\|_p^p &= \mathbb{E}\left[\left|\left(P(X_{T,\theta}^{n_l}) - P(X_{T,\theta}^{n_{l-1}})\right)\mathcal{I}^-(W_T^l,\theta)\right|^p\right], \\ &= \mathbb{E}\left[\left|\left(P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right)\right|^p \left(\mathcal{I}^+(W_T^l,\theta)\right)^{p-1}\right], \\ &\leq \left(\mathbb{E}\left[\left|\left(P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right)\right|^{p^2}\right]\right)^{\frac{1}{p}} \left\|\left(\mathcal{I}^+(W_T^l,\theta)\right)^{p-1}\right\|_{\frac{p}{p-1}}, \\ &\leq e^{\frac{(p^2-1)}{2}c^2T}\left\|P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right\|_{p^2}^p.\end{aligned} \tag{3.4.14}$$

Therefore, from the above analysis, we have,

$$\left\|\left(P(X_{T,\theta}^{n_l}) - P(X_{T,\theta}^{n_{l-1}})\right)\mathcal{I}^-(W_T^l,\theta)\right\|_p \leq e^{\frac{(p^2-1)}{2p}c^2T}\left\|P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right\|_{p^2}. \tag{3.4.15}$$

Further for (II) of (3.4.13), the assumption (3.4.2) yields,

$$\left\|\left[P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right]\right\|_p^p \leq \mathcal{V}_1^{(p)}(1+m^{\beta/2})^p \mathbf{h}^{\beta p/2} m^{-\beta p(l-1)/2}. \tag{3.4.16}$$

Now from (3.4.13) and (3.4.15), we get,

$$\left(\mathbb{E}\left[|\widetilde{Y}_{l,\theta})|^p\right]\right)^{1/p} \leq \left(1 + e^{\frac{(p^2-1)}{2p}c^2T}\right)\left\|P(X_T^{n_l}) - P(X_T^{n_{l-1}})\right\|_{p^2}. \tag{3.4.17}$$

Combining above inequality and (3.4.16) yields,

$$\|\widetilde{Y}_{l,\theta}\|_p^p \leq K_1(m,\beta,p,c)m^{-\beta p(l-1)/2}, \tag{3.4.18}$$

where,

$$K_1(m,\beta,p,c) = \left(1 + e^{\frac{(p^2-1)}{2p}c^2T}\right)\left(\mathcal{V}_1^{(p^2)}\right)^{1/p}(1+m^{\beta/2})^p \mathbf{h}^{\beta p/2}.$$

□

Proposition 3.4.1. *Let $p \geq 2$ and $\theta_l^k \in \Theta$ for $l = 2,\ldots,L$ and $k \in \mathbb{N}^*$. Then there exists a constant $K_2(m,\beta,p,c)$, such that,*

$$\mathbb{E}\left[|\widetilde{\mathcal{J}}_{\theta,\pi}^{2,\epsilon}|^p\right] \leq K_2(m,\beta,p,c)\epsilon^p \text{ for some constant } c > 0. \tag{3.4.19}$$

Proof. We start our discussion by observing that as $\theta_l^k \in \Theta$, there exists a positive constant c such that $|\theta_l^k| \leq c,\ \forall k \in \mathbb{N}$. Now, we define the following filtration $(\mathcal{F}_{T,k})_{k\geq 1}$, where $\mathcal{F}_{T,k} := \sigma(W_{t,j},\ j \leq k, t \leq T)$. With this filtration, one can readily observe that $\sum_{k=1}^{j} \widetilde{Y}_{l,\theta_l^{k-1}}^k$ is a martingale with respect to $\mathcal{F}_{T,j}$. Now consider the following definition,

$$s_l := \sum_{k=1}^{N_l} \widetilde{Y}_{l,\theta_l^{k-1}}^k, \text{ for } l = 2,\ldots,L. \tag{3.4.20}$$

Since $\sum_{k=1}^{j} \widetilde{Y}_{l,\theta_l^{k-1}}^k$ is $\mathcal{F}_{T,j}$ martingale, therefore by Rosenthal's inequality [43], we have,

$$\|s_l\|_p^p \leq C_p \left\{ \mathbb{E}\left[\sum_{k=1}^{N_l} \mathbb{E}[(\widetilde{Y}_{l,\theta_l^{k-1}})^2 | \mathcal{F}_{T,k-1}]\right]^{p/2} + \sum_{k=1}^{N_l} \mathbb{E}\left[|\widetilde{Y}_{l,\theta_l^{k-1}}|^p\right] \right\}. \tag{3.4.21}$$

Now, from the previous Lemma, one can easily conclude that,

$$\sum_{k=1}^{N_l} \mathbb{E}\left[|\widetilde{Y}_{l,\theta_l^{k-1}}|^p\right] \leq N_l K_1 m^{-\beta p(l-1)/2}. \tag{3.4.22}$$

As for the first term, we refer to [43, Theorem A.8], to obtain,

$$\mathbb{E}\left[\sum_{k=1}^{N_l}\mathbb{E}\left[(\widetilde{Y}_{l,\theta_l^{k-1}})^2|\mathcal{F}_{T,k-1}\right]\right]^{p/2} \leq A_p\mathbb{E}\left[\sum_{k=1}^{N_l}(\widetilde{Y}_{l,\theta_l^{k-1}})^2\right]^{p/2} = A_p\left\|\sum_{k=1}^{N_l}(\widetilde{Y}_{l,\theta_l^{k-1}})^2\right\|_{p/2}^{p/2}$$
$$\leq A_p\left(\sum_{k=1}^{N_l}\|(\widetilde{Y}_{l,\theta_l^{k-1}})^2\|_{p/2}\right)^{p/2} \leq A_p(N_l)^{p/2}K_1m^{-\beta p(l-1)/2}, \tag{3.4.23}$$

where the last inequality is the consequence of the previous Lemma. Therefore, we have for $p \geq 2$,

$$\|s_l\|_p^p \leq C_p\left\{A_p(N_l)^{p/2}K_1m^{-\beta p(l-1)/2} + N_lK_1m^{-\beta p(l-1)/2}\right\} \leq C_p\left\{2A_p(N_l)^{p/2}K_1m^{-\beta p(l-1)/2}\right\}. \tag{3.4.24}$$

Now, let $K_p \coloneqq C_pA_pK_1$. Therefore, we have,

$$\mathbb{E}[|s_l|^p] \leq \left(K_p(N_l)^{p/2}m^{-\beta p(l-1)/2}\right). \tag{3.4.25}$$

Now, consider the following,

$$\mathbb{E}\left[|\widetilde{\mathcal{J}}_{\theta,\pi}^2|^p\right] = \mathbb{E}\left[\left|\sum_{l=2}^{L}\frac{\widetilde{W}_l}{N_l}s_l\right|^p\right]. \tag{3.4.26}$$

As one can observe that, $(s_l)_{l\geq 2}$ are independent random variable in l, therefore, by a version of Rosenthal's inequality [43], we have,

$$\mathbb{E}\left[\left|\sum_{l=2}^{L}\frac{\widetilde{W}_l}{N_l}s_l\right|^p\right] \leq \delta_p\left\{\left(\sum_{l=2}^{L}\mathbb{E}\left[\frac{\widetilde{W}_l}{N_l}s_l\right]^2\right)^{p/2} + \sum_{l=2}^{L}\mathbb{E}\left[\left|\frac{\widetilde{W}_l}{N_l}s_l\right|^p\right]\right\}$$
$$\leq 2\delta_p\left(\sum_{l=2}^{L}\mathbb{E}\left|\frac{\widetilde{W}_l}{N_l}s_l\right|^2\right)^{p/2}. \tag{3.4.27}$$

For $p = 2$, we have,

$$\mathbb{E}\left[|s_l|^2\right] \leq K_2(N_l)m^{-\beta(l-1)}.$$

Therefore,

$$\mathbb{E}\left[\left|\sum_{l=2}^{L}\frac{\widetilde{W}_l}{N_l}s_l\right|^p\right] \leq 2\delta_p\left(\sum_{l=2}^{L}\frac{|\widetilde{W}_l|^2}{N_l}K_2m^{-\beta(l-1)}\right)^{p/2}. \tag{3.4.28}$$

We know that, $N_l = \lceil N\mu_l \rceil \geq N\mu_l$, and as a result we have,

$$\frac{1}{N_l} \leq \frac{1}{N\mu_l},\ l = 1, \ldots, L.$$

Further, owing to the expression for μ_l, we have,

$$\frac{|\widetilde{W_l}|}{\mu_l} \leq \frac{1}{\lambda \mathbf{h}^{\frac{\beta}{2}} \underline{C}_{m,\beta} q^*} m^{\frac{(\beta+1)(l-1)}{2}},\ l = 2, \ldots, L.$$

Since, $\sup_{l \in (1,\ldots,L), L \geq 1} |\widetilde{W_l}| \leq a_\infty \widetilde{B}_\infty$ [39], therefore, combining everything we get,

$$\mathbb{E}\left[\left|\sum_{l=2}^{L} \frac{\widetilde{W_l}}{N_l} s_l\right|^p\right] \leq \widetilde{K}\left(\frac{1}{N}\sum_{l=2}^{L} m^{\frac{(1-\beta)(l-1)}{2}}\right)^{p/2}, \tag{3.4.29}$$

where, $\widetilde{K} = 2\delta_p \left(\frac{a_\infty \widetilde{B}_\infty K_2}{\lambda \mathbf{h}^{\frac{\beta}{2}} \underline{C}_{m,\beta} q^*}\right)^{\frac{p}{2}}$. Now as a consequence of Lemma 4.5 in [39], with $\bar{\epsilon} \to 0$, we have,

$$\forall\ \epsilon \in (0, \bar{\epsilon}],\ \frac{1}{N} \leq \frac{2}{C_\beta}\epsilon^2 \begin{cases} 1, & \text{if } \beta > 1, \\ L^{-1}, & \text{if } \beta = 1. \\ m^{-\frac{1-\beta}{2}L}, & \text{if } \beta < 1. \end{cases} \tag{3.4.30}$$

Moreover, it is easy to prove that,

$$\sum_{l=2}^{L} m^{\frac{(1-\beta)(l-1)}{2}} \leq \begin{cases} \frac{1}{1-m^{\frac{1-\beta}{2}}}, & \text{if } \beta > 1, \\ L, & \text{if } \beta = 1, \\ \frac{m^{\frac{1-\beta}{2}L}}{m^{\frac{1-\beta}{2}}-1}, & \text{if } \beta < 1. \end{cases} \tag{3.4.31}$$

With all the preceding discussion, we have,

$$\mathbb{E}\left|\sum_{l=2}^{L} \frac{\widetilde{W_l}}{N_l} s_l\right|^p \leq K_2\left(m, \beta, p, c\right)\epsilon^p, \tag{3.4.32}$$

where,

$$K_2\left(m,\beta,p,c\right)=\widetilde{K}\left(\frac{2}{C_\beta}\right)^{p/2}\begin{cases}(1-m^{\frac{1-\beta}{2}})^{-p/2}, & \text{if } \beta>1,\\ 1, & \text{if } \beta=1,\\ (m^{\frac{(1-\beta)}{2}}-1)^{-p/2}, & \text{if } \beta<1.\end{cases} \tag{3.4.33}$$

□

With the above results in our hands, we are ready to prove the Strong Law of Large Numbers.

Proof of Theorem 3.4.3. We start our proof by proving, $\widetilde{\mathcal{J}}^{2,\epsilon_k}_{\theta,\pi}\xrightarrow{a.s.}0$ as $k\to\infty$. Clearly, as a consequence of our assumption on $(\epsilon_k)_{k\geq 1}$, we have,

$$\sum_{k\geq 1}\mathbb{E}\left[|\widetilde{\mathcal{J}}^{2,\epsilon_k}_{\theta,\pi}|^p\right]<+\infty.$$

Hence, by Beppo-Levi's Theorem, $\sum_{k\geq 1}|\widetilde{\mathcal{J}}^{2,\epsilon_k}_{\theta,\pi}|^p<+\infty\ a.s.$, and as an implication we have $\widetilde{\mathcal{J}}^{2,\epsilon_k}_{\theta,\pi}\xrightarrow{a.s.}0$ as $k\to+\infty$. We now turn our attention to proving $\widetilde{\mathcal{J}}^{1,\epsilon_k}_{\theta,\pi}\xrightarrow{a.s.}0$ as $k\to\infty$. It is clear that as $k\to\infty$, $\epsilon_k\to 0$, which in turns implies $N_1\to\infty$. As one can observe that, $\sum_{k=1}^{j}\widetilde{Y}^1_{k,\theta_1^{k-1}}$ is $\mathcal{F}_{T,j}$ martingale. Therefore, as an application of Rosenthal's inequality for $p=2$, we have,

$$\begin{aligned}\mathbb{E}\left[|\widetilde{\mathcal{J}}^{1,\epsilon_k}_{\theta,\pi}|^2\right]&\leq\frac{C_2}{N_1^2}\left\{\mathbb{E}\left[\sum_{k=1}^{N_1}\mathbb{E}\left[\left(\widetilde{Y}^1_{k,\theta_1^{k-1}}\right)^2\middle|\mathcal{F}_{T,k-1}\right]\right]\right\},\ \text{where } C_2 \text{ is a constant.}\\ &=\frac{C_2}{N_1^2}\left\{\sum_{k=1}^{N_1}\left(\mathbb{E}\left[P(X_T^{n_1})^2\mathcal{I}^+(\theta_1^{k-1},W_T)\right]-\left[\mathbb{E}[P(X_T^{n_1})]\right]^2\right)\right\},\end{aligned} \tag{3.4.34}$$

where the last equality is the consequence of $X^1_{T,k}$ being independent of $\mathcal{F}_{T,k-1}$ and θ^1_{k-1} being $\mathcal{F}_{T,k-1}$ measurable. Further, due to Grisanov's theorem, we introduce a couple of random variables X^1_T and W_T independent of $\mathcal{F}_T=\cup_{k\geq 1}\mathcal{F}_{T,k}$, justifying the last equality. Further, as $\theta_1^{k-1}\in\Theta$, therefore there exists a $c>0$ such that $|\theta_1^{k-1}|\leq c$ for $k\in\mathbb{N}$. As a

consequence, $\sup_{k\in\mathbb{N}}|P(X_T^{n_1})^2\mathcal{I}^+(\theta_1^{k-1},W_T)| \leq P(X_T^{n_1})^2 e^{c|W_T|+\frac{c^2}{2}T}$. Now for $p \geq 2$,

$$\mathbb{E}\left[P(X_T^{n_1})^2 e^{c|W_T|+\frac{c^2}{2}T}\right] \leq \left\|P(X_T^{n_1})^2\right\|_p \left\|e^{c|W_T|+\frac{c^2}{2}T}\right\|_{\frac{p}{p-1}} < +\infty.$$

Therefore, as consequence of Theorem 3.4.2 and Lebesgue theorem, we obtain that,

$$\lim_{k\to\infty}\mathbb{E}[P(X_T^{n_1})^2\mathcal{I}^+(\theta_1^{k-1},W_T)] = \mathbb{E}[P(X_T^{n_1})^2\mathcal{I}^+(\theta_1^{*},W_T)].$$

Now as the application of Cesaro's lemma in equation (3.4.34) one can easily conclude that, $\widetilde{\mathcal{J}}_{\theta,\pi}^{1,\epsilon_k} \xrightarrow{a.s.} 0$ as $k \to +\infty$. □

We now look at the numerical illustrations to demonstrate the efficacy of the discussed algorithm and also assess the practicality of the algorithm in real-life applications.

3.5 Numerical Illustration

In this section, we consider the problem of option pricing, where we try to estimate the European and Lookback call option prices. Further, we assume that one-dimensional SDE drives the stochastic process to keep things simple. Accordingly, we use Milstein and the Euler-Maruyama discretization scheme to simulate the underlying stochastic process. Before discussing the implementation procedure, we provide a mathematical description of the discretization schemes. Consider the general one-dimensional SDE on the probability space $(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$,

$$dX_t = r(X_t,t)dt + \sigma(X_t,t)dW_t. \tag{3.5.1}$$

The Milstein discretization of the above equation is given by,

$$X_{n+1} = X_n + r_n h + \sigma_n \Delta W_n + \frac{1}{2}\sigma_n'\sigma_n\left((\Delta W_n)^2 - h\right). \tag{3.5.2}$$

In the above equation, h is the uniform time-step, $r_n = r(X_n,t_n)$, $\sigma_n = \sigma(X_n,t_n)$ and $\sigma_n' = \sigma'(X_n,t_n)$, with $t_n \coloneqq nh$. However, under the parametric change of measure, with

θ as the parameter and the underlying SDE is given by,

$$dX_t(\theta) = r(X_t(\theta), t)dt + \sigma(X_t(\theta), t)dB_t, \tag{3.5.3}$$

where, $B_t := W_t + \theta t$, we use the following discretization schemes in order to simulate the SDE in the probability space $\left(\Omega, \{\mathcal{F}_t^\theta\}_{t\geq 0}, \mathbb{P}_\theta\right)$,

$$\begin{aligned} X_{n+1}^\theta &= X_n^\theta + r(X_n^\theta, t_n)h + \sigma(X_n^\theta, t_n)\Delta B_n + \frac{1}{2}\sigma'(X_n^\theta, t_n)\sigma(X_n^\theta, t_n)\left((\Delta B_n)^2 - h\right), \\ &= X_n^\theta + \left(r(X_n^\theta, t_n) + \theta\sigma(X_n^\theta, t_n)\right)h + \sigma(X_n^\theta, t_n)\Delta W_n, \\ &+ \frac{1}{2}\sigma'(X_n^\theta, t_n)\sigma(X_n^\theta, t_n)\left((\Delta W_n + \theta h)^2 - h\right). \end{aligned}$$

As for the Euler scheme, we have,

$$X_{n+1}^\theta = X_n^\theta + \left(b(X_n^\theta, t_n) + \theta\sigma(X_n^\theta, t_n)\right)h + \sigma(X_n^\theta, t_n)\Delta W_n. \tag{3.5.4}$$

We start our computation by determining the optimal structural parameters required to perform the simulations. In order to do so, we perform a pre-simulation to approximate the value of $\mathcal{V}_1$, λ and $\mathrm{Var}(P(X_T^0))$, necessary for the computation of the optimal structural parameters. For computing $\mathcal{V}_1$ we refer to the formula presented in [60], *i.e.,*

$$\mathcal{V}_1 = \left(1 + m_{\max}^{-\beta/2}\right)^{-2} \mathbf{h}^{-\beta} \| P\left(X_T^h\right) - P\left(X_T^{h/m_{\max}}\right)\|_2^2. \tag{3.5.5}$$

Here we set $m_{\max} = 10$. As for $\mathrm{Var}(P(X_T^0))$, we perform small prior simulations and empirically calculate the variance. Further, the value of λ is calculated as $\lambda = \sqrt{\dfrac{\mathcal{V}_1}{\mathrm{Var}(P(X_T^0))}}$. We use the above values to calculate the structural parameters using formulas from Table 3.1 and perform the AISML2R and ML2R simulation. The computation of the structural parameter requires the desired RMSE ϵ as input. Therefore, we calculate these parameters for $\epsilon = 2^{-k}$, where $k = 1, \ldots, 5$. Also, throughout our computation, we consider the refinement factor of $m = 4$. For the purpose of estimating the optimal θ_l^* for each level of resolution, we use the algorithm described in subsection 3.3.4. In all the numerical experiments carried out below, we consider $\lambda_n = \dfrac{1}{(n+1)}$ and $\Theta := [-10, 10] \subset \mathbb{R}$. For practical purposes, we stop the stochastic approximation procedure after finite iterations, say I. Further, to stabilize the convergence of the algorithm, we employ Rupert and Po-

liak averaging principle (see e.g. [70]), *i.e.,* instead of using θ_l^k, we use $\widetilde{\theta}_l^k = \frac{1}{k+1}\sum_{i=0}^{k}\theta_l^i$, on level l. In order to study our results, we perform $R = 50$ repetition of our experiment and observe the bias *i.e.* $\widetilde{\mu}$, and variance *i.e.*, $\widetilde{\sigma^2}$ for a given computational budget, using the formulas describes below.

$$\widetilde{\mu} = \frac{1}{R}\sum_{j=1}^{R}(\mathcal{J}_{\pi,j}^{N,\theta} - \mathcal{J}_0),$$

$$\widetilde{\sigma}^2 = \frac{1}{R}\sum_{j=1}^{R}\left[\sum_{l=1}^{L}\left(\frac{1}{N_l^j(N_l^j-1)}\left(\sum_{k=1}^{N_l^j}\widetilde{W}_l(Z_l^j) - \widetilde{W}_l\frac{1}{N_l^j}\sum_{k=1}^{N_l^j}Z_l^j\right)^2\right)\right]$$

where,

$$Z_l = \left(P\left(X_T^{h/m^{l-1}}\right) - P\left(X_T^{h/m^{l-2}}\right)\right)\mathcal{I}^-(W_T^l, \widetilde{\theta}_l^I),$$

consequently we have,

$$\text{RMSE} = \sqrt{\widetilde{\mu}^2 + \widetilde{\sigma}^2} \tag{3.5.6}$$

Moreover, we estimate computational cost as,

$$\text{cost}(\mathcal{J}_\pi^{N,\theta}) = \frac{N}{h}\sum_{l=1}^{L}\mu_l(n_{l-1}+n_l) + I\sum_{l=1}^{L}(n_{l-1}+n_l), \tag{3.5.7}$$

where $I = 0$ for standard ML2R simulation. Finally, to assess the effectiveness of the hybrid procedure, we evaluate the Improvement Factor (IF), defined as,

$$\text{IF} := \frac{\widetilde{\sigma}_{ML2R}^2 \times \text{cost}(\mathcal{J}_\pi^N)}{\widetilde{\sigma}_{AISML2R}^2 \times \text{cost}(\mathcal{J}_\pi^{N,\theta})} \tag{3.5.8}$$

Before, delving into the numerical examples, we discuss the computational aspect of importance sampling as opposed to standard simulation. Based on the cost formulation for the adaptive simulation, it is evident that importance sampling is more expensive than standard simulation. Since we stop the stochastic approximation after I iterations, a pertinent question is how many iterations are enough and whether it is necessary to undergo the importance sampling procedure on a particular level. In [10], it was observed that increasing the number of stochastic approximation iterations from $I = 1000$ to $I = 15000$ did not result in any substantial decrease in the overall variance of the

estimator. A similar observation is made in our numerical experiments, where we observe that performing the importance sampling procedure on every level of resolution is not always necessary. For instance, Figure 3.1 and 3.2 represent variance reduction achieved with $I = 500$ on various levels of resolution for the European option pricing problem studied in subsection 3.5.1. From both Figures, it is evident that under the Milstein scheme, *i.e.* $\beta > 1$, one does not see significant variance reduction after we achieve $\log(\text{Var}[(P_l - P_{l-1})]) = \mathcal{O}(10^{-3})$. Observing this and recalling that for $\beta > 1$, the majority of computation is concentrated at the coarser level, undergoing importance sampling at finer levels adds to the computational cost without resulting in substantial variance reduction. In light of these observations, in our numerical experiments, we do not undergo the importance sampling procedure on a level if the $\log(\text{Var}[(P_l - P_{l-1})]) = \mathcal{O}(10^{-3})$ on that level.

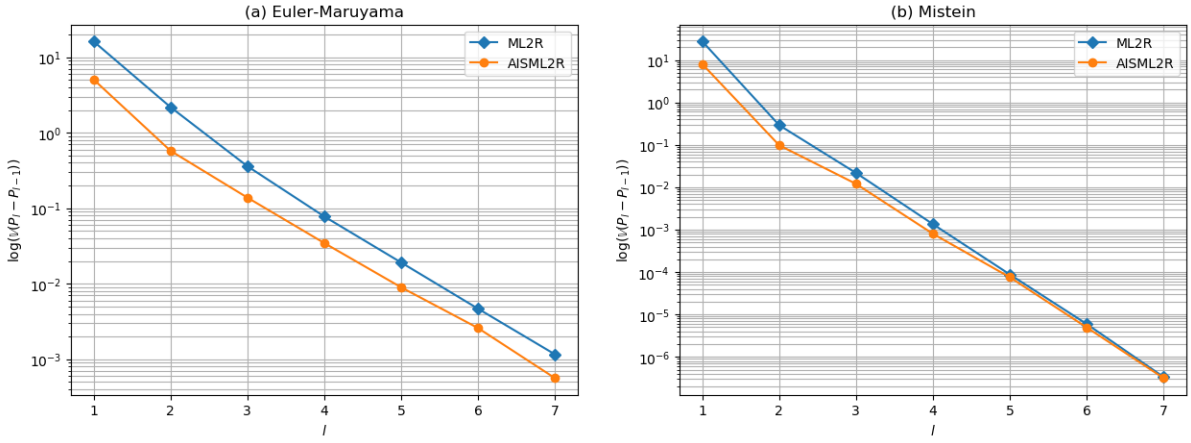


Figure 3.1: $\sigma = 0.2$: (a) Euler-Maruyama Approximation- $\log(\text{Var}[(P_l - P_{l-1})])$ as a function of levels.(b) Milstein Approximation- $\log(\text{Var}(P_l - P_{l-1})])$ as a function of levels.

3.5.1 European Option

The payoff function in the case of the European call option is described below,

$$P(X_T) = e^{-rT}(X_T - K)_+. \tag{3.5.9}$$

For the practical implementation, we have considered $X_0 = 80$, $r = 0.06$, $T = 1$, $K = 100$ and perform our experiments for $\sigma = \{0.2, 0.6\}$. In Figures 3.3 and 3.4, we compare the RMSE achieved by ML2R and AISML2R obtained under Milstein and Euler-Maruyama schemes for σ equals to 0.2 and 0.6 respectively. For the importance sampling

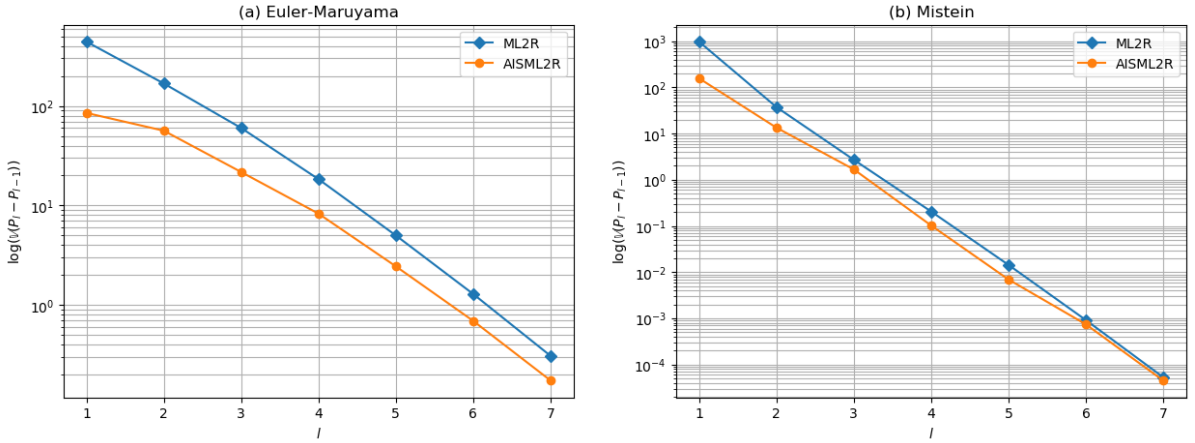


Figure 3.2: $\sigma = 0.6$: (a) Euler-Maruyama Approximation- $\log(\text{Var}[(P_l - P_{l-1})])$ as a function of levels.(b) Milstein Approximation- $\log(\text{Var}(P_l - P_{l-1})])$ as a function of levels.

procedure, we conduct $I = \min\{500, Nl\}$ iteration on a level if $\log(\text{Var}[(P_l - P_{l-1})]) \geq \mathcal{O}(10^{-2})$ on that level. As we observe in each of the cases, adaptive sampling leads to substantial variance reduction with almost the same computational cost. Finally, in Table 3.2, we demonstrate the preponderance of the AISML2R over ML2R for various values of ϵ.

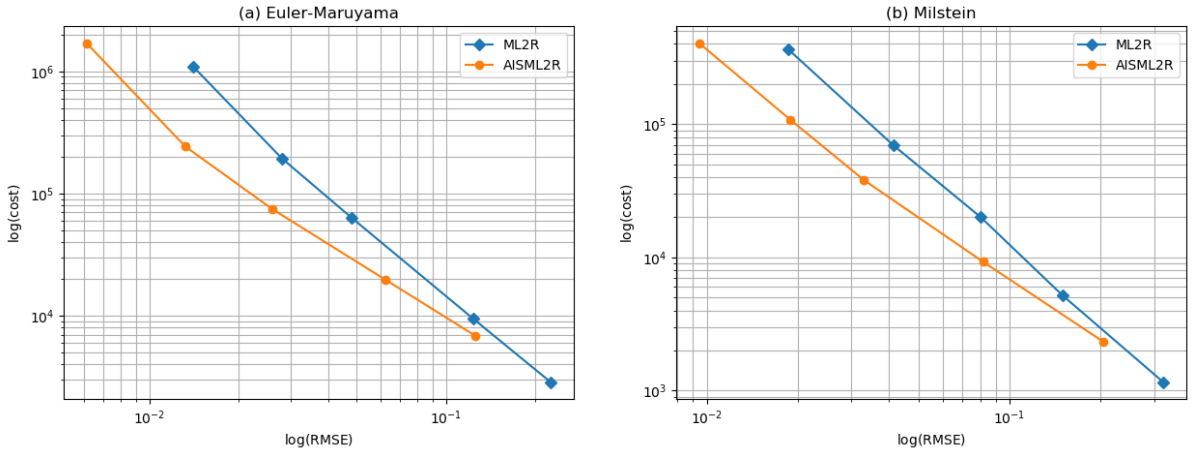


Figure 3.3: $\sigma = 0.2$: (a) Euler-Maruyama Approximation- Computational Cost as a function of RMSE log-log scale.(b) Milstein Approximation- Computational Cost as a function of RMSE log-log scale.

$\sigma = 0.6$ (Euler)							$\sigma = 0.2$ (Euler)					
	AISML2R		ML2R				AISML2R		ML2R			
ϵ	variance	cost	variance	cost	IF		ϵ	variance	cost	variance	cost	IF
5.000e-1	9.473e-3	3.099e+5	4.349e-2	1.959e+5	2.90		5.000e-1	1.522e-2	6.890e+3	4.949e-2	2.866e+3	1.35
2.500e-1	3.664e-3	7.610e+5	1.261e-2	6.735e+5	3.05		2.500e-1	3.857e-3	1.969e+4	1.525e-2	9.442e+3	1.90
1.250e-1	5.340e-4	5.149e+6	2.163e-3	3.940e+6	3.10		1.250e-1	6.730e-4	7.436e+4	2.297e-3	6.332e+4	2.91
6.250e-2	2.938e-4	9.767e+6	8.136e-4	1.049e+7	2.97		6.250e-2	1.740e-4	2.418e+5	7.568e-4	1.932e+5	3.48
3.125e-2	5.700e-5	1.085e+8	1.526e-4	1.089e+8	2.68		3.125e-2	3.770e-5	1.696e+6	1.950e-4	1.094e+6	3.35
$\sigma = 0.6$ (Milstein)							$\sigma = 0.2$ (Milstein)					
	AISML2R			ML2R				AISML2R		ML2R		
ϵ	variance	cost	variance	cost	IF		ϵ	variance	cost	variance	cost	IF
5.000e-1	1.405e-2	1.164e+5	7.076e-2	7.261e+4	3.14		5.000e-1	4.121e-2	2.328e+3	1.027e-1	1.158e+3	1.24
2.500e-1	4.270e-3	3.500e+5	2.467e-2	1.978e+5	3.26		2.500e-1	6.653e-3	9.280e+3	2.225e-2	5.185e+3	1.87
1.250e-1	8.067e-4	1.856e+6	4.526e-3	1.123e+6	3.40		1.250e-1	1.090e-3	3.803e+4	5.941e-3	2.004e+4	2.87
6.250e-2	4.343e-4	2.910e+6	9.686e-4	6.046e+6	4.63		6.250e-2	3.167e-4	1.079e+5	1.714e-3	6.919e+4	3.47
3.125e-2	6.660e-5	2.441e+7	2.039e-4	3.054e+7	3.83		3.125e-2	8.880e-5	4.059e+5	3.474e-4	3.620e+5	3.49

Table 3.2: Comparison of AISML2R and ML2R for European Option

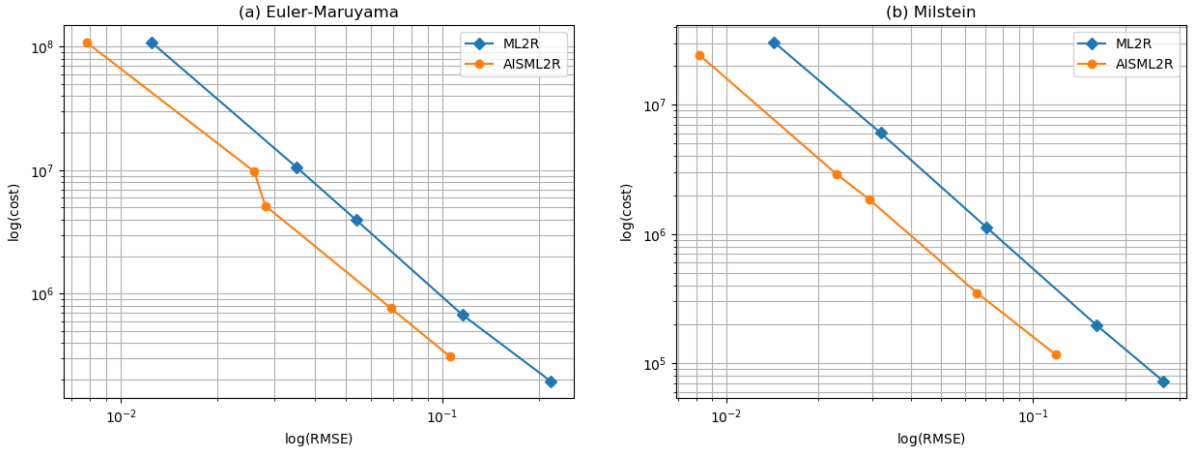


Figure 3.4: $\sigma = 0.6$: (a) Euler-Maruyama Approximation- Computational Cost as a function of RMSE log-log scale.(b) Milstein Approximation- Computational Cost as a function of RMSE log-log scale.

3.5.2 Lookback Option

For our next example, we consider a partial Lookback call option defined in the following way,

$$P(X_T) = e^{-rT}\left(X_T - \zeta \min_{t\in[0,T]} X(t)\right)_+, \text{ where } \zeta \geq 1. \tag{3.5.10}$$

The parameters for the purpose of the practical implementation are $X_0 = 10$, $r = 0.05$, $T = 1$ and $\zeta = 1$. In this case, we perform our experiment with Euler-Maruyama discretization for $\sigma = \{0.4, 0.6\}$. The value of $\alpha = 0.5$ and $\beta = 1$ for the payoff functional defined above [60]. In Figure 3.5, we compare the RMSE achieved by AISML2R and ML2R, and in Table 3.3, we demonstrate the variance reduction achieved via importance sampling for various values of ϵ.

$\sigma = 0.4$					
	AISML2R		ML2R		
ϵ	variance	cost	variance	cost	IF
5.000e-1	2.736e-2	3.909e+3	4.950e-2	1.762e+3	0.82
2.500e-1	6.583e-3	1.170e+4	1.159e-2	7.771e+3	1.17
1.250e-1	1.949e-3	6.465e+4	3.184e-3	4.722e+4	1.19
6.250e-2	5.207e-4	1.942e+5	8.537e-4	1.773e+5	1.50
3.125e-2	2.310e-4	4.126e+5	3.560e-4	4.139e+5	1.54

$\sigma = 0.6$					
	AISML2R		ML2R		
ϵ	variance	cost	variance	cost	IF
5.000e-1	1.746e-2	1.209e+4	4.176e-2	7.141e+3	1.41
2.500e-1	5.217e-3	2.923e+4	1.086e-2	2.772e+4	1.97
1.250e-1	1.853e-3	1.685e+5	2.832e-3	1.903e+5	1.73
6.250e-2	3.853e-4	7.247e+5	9.148e-4	5.914e+5	1.94
3.125e-2	1.700e-4	1.637e+6	3.310e-4	1.637e+6	1.95

Table 3.3: Comparison of AISML2R and ML2R for Lookback Option

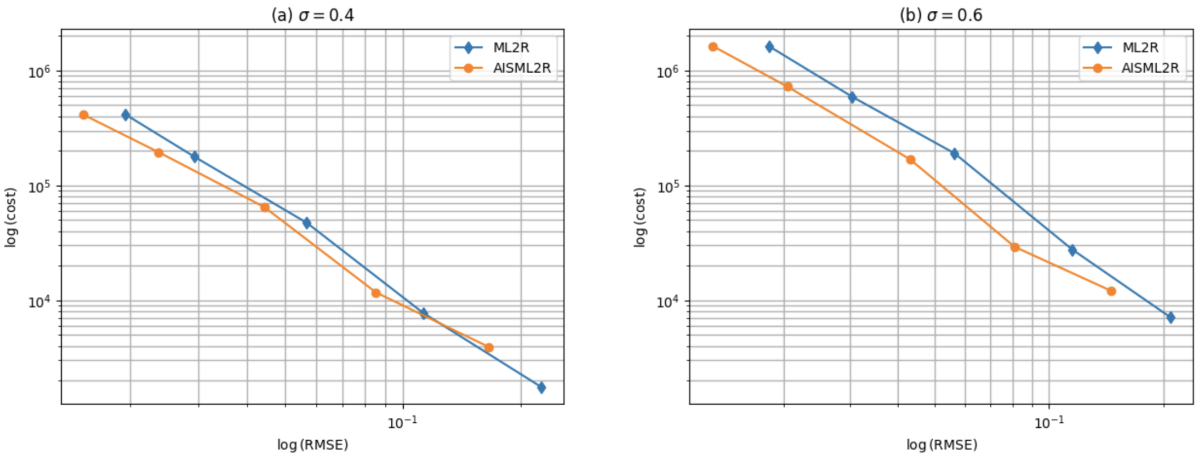


Figure 3.5: Euler-Maruyama Approximation- Computational Cost as a function of RMSE log-log scale.

3.6 Summary

In this chapter, we presented a novel combination of ML2R and Importance sampling algorithm developed upon the studies carried out in [2, 3, 7, 52, 60]. We proved the existence and uniqueness of the optimal change of measure parameter θ_l^* for for every level of resolution l and proved the asymptotic results illustrating the convergence of $\theta_l^k \to \theta_l^*$ as $k \to \infty$. We further proved the Strong Law of Large Numbers for the AISML2R estimator. In Section 3.5, we undertook numerical experiments where we observed the preponderance of AISML2R both in the context of the European Option and the Lookback Option.

Chapter 4

Non-Asymptotic Analysis of Sample Average Approximation with Biased Sampling

4.1 Introduction

In this chapter, we focus on an important numerical procedure grounded in Monte Carlo principles popularly known as *Sample Average Approximation* or SAA. The primary discussion in this chapter revolves around the effect of biased sampling in the stochastic optimization paradigm. Accordingly, we begin our discussion with an overview of the SAA framework.

4.1.1 Sample Average Approximation

The Sample Average Approximation (SAA) method is a numerical algorithm designed to address optimization problems where the input data is uncertain. Specifically, it targets approximating the optimal solution for optimization problems, formulated as follows:

$$\min_{x \in \mathcal{X}} \{F(x) := \mathbb{E}[f(x, \zeta)]\} \tag{4.1.1}$$

Here, $\mathcal{X} \subseteq \mathbb{R}^d$ is assumed to be a finite-dimensional compact set, and $f(\cdot, \zeta)$ denotes the cost function, with ζ being a random vector. Typically, this optimization doesn't lend itself to analytical solutions, necessitating a Monte Carlo-based approach for ap-

proximation. The SAA method, known for its simplicity and robustness, has become a preferred tool among practitioners. SAA operates by solving an approximation of the original problem, defined as:

$$\min_{x\in\mathcal{X}}\left\{F_N(x) := \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_k)\right\} \tag{4.1.2}$$

where, $\zeta_1,\ldots,\zeta_N$ represent independent and identically distributed (i.i.d) samples of the random vector ζ, drawn from its distribution. Extensive literature has documented the convergence of the approximate solution to the optimal one, with numerous references discussing the convergence and providing comprehensive surveys on SAA. In this domain, the seminal work in [54] demonstrated that if $f(\cdot,\zeta)$ is Lipschitz continuous, SAA requires a computational complexity of $\mathcal{O}((d+\gamma)\epsilon^{-2}\log(\epsilon^{-1}))$ to achieve an ϵ-optimal solution with probability ϵ^{γ}, for some $\gamma > 0$. Similar results were obtained for constrained stochastic programs and two-stage stochastic optimization problems in subsequent studies [79].

However, existing studies inherently assumed that the Monte Carlo estimator is unbiased *i.e.*,

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_k)\right] = F(x) \;\; \forall\, x\in\mathcal{X}. \tag{4.1.3}$$

This chapter explores the impact on the computational complexity of the SAA procedure due to bias introduced in the estimator while sampling the random variable from its approximate distribution. The primary motivation for studying the SAA in the biased framework bears testimony to the fact that in most practical scenarios, the sampling of the random variable from its exact distribution is not always possible.

For example, in the financial engineering paradigm, CVaR estimation can be formulated as a stochastic optimization problem. Mathematically speaking, if ζ denotes the loss of a portfolio and α is the required confidence interval, then by Rockafellar-Uryasev [74] representation we have,

$$\mathrm{CVaR}_{\nu}(\zeta) := \min_{x\in[\vartheta^*,\vartheta^{**}]}\left\{x + \frac{1}{1-\nu}\mathbb{E}[(\zeta-x)_+]\right\}. \tag{4.1.4}$$

where $(y)_+ = \max\{y,0\}$, and

$$\vartheta^* = \inf\{x\in\mathbb{R} : \mathbb{P}(\zeta\le x)\ge\nu\} \text{ and } \vartheta^{**} = \sup\{x\in\mathbb{R} : \mathbb{P}(\zeta\le x)\le\nu\}. \tag{4.1.5}$$

In the context of CVaR estimation, if we assume that the underlying asset driving the loss function is modelled on a stochastic differential equation, then using the numerical approximation technique to approximate the loss would induce bias in the Monte Carlo estimation, affecting the performance of the SAA procedure. Moreover, following the discussion in Chapter 2, the scenario simulation approach leads to a nested expectation framework, thereby inducing bias in the Monte Carlo approximation of the objective function.

Additionally, portfolio selection problems in finance [47], robust supervised learning in computer vision and speech recognition [49], reinforcement learning in policy evaluation [49], all belong to a class of conditional stochastic optimization problem, defined as,

$$\min_{x\in\mathcal{X}} \left\{ F(x) := \mathbb{E}_{\zeta}\left[f\left(\mathbb{E}_{\eta|\zeta}[g_{\eta}(x,\zeta)] \right) \right] \right\}. \tag{4.1.6}$$

In the above setup, approximating the inner expectation using a Monte Carlo procedure introduces a bias in estimating the expectation. As the problem belonging to the class of stochastic composition optimization has been a long-standing challenge in science and engineering, extensive research studying SAA to solve them is available in the literature, see, e.g. [24, 29, 83, 5]. However, for the most part, the primary focus of their discussion is towards studying the asymptotic properties of the estimator, deriving central limit formulae [24] and establishing the rate of convergence [29]. The study performed in [49] shows that the computational complexity required for solving the conditional stochastic optimization problem is $\mathcal{O}((d+\gamma)\epsilon^{-4}\log(\epsilon^{-1}))$, if the cost function is Lipschtiz continuous and is $\mathcal{O}((d+\gamma)\epsilon^{-3}\log(\epsilon^{-1}))$, if the cost function is smooth, where the increase in the computational complexity can be attributed to the Monte Carlo approximation of the inner expectation, which in turn induces bias in the estimator. To our knowledge, a generalized study on the biased approximation of the random variable in the SAA paradigm has not been explored. We intend to fill this gap in this work.

4.2 Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the complete probability space. In this space, let us consider the following stochastic optimization problem,

$$\min_{x\in\mathcal{X}} \{F(x) = \mathbb{E}[f(x,\zeta)]\} \tag{4.2.1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a finite-dimensional compact set, ζ is a random vector whose distribution is supported on the set $\Theta \in \mathbb{R}^s$, and $f\colon \mathcal{X} \times \Theta \to \mathbb{R}$ is the cost function. Since we will be working in the bias framework, we recall $\mathfrak{B}$ to be the set of bias parameters with $\mathfrak{B} \cup \{0\}$ being the compact set and

$$\forall m \in \mathbb{N}, \quad \frac{\mathfrak{B}}{m} \subset \mathfrak{B}. \tag{4.2.2}$$

We let ζ_h be an approximation of the random variable ζ for some $h \in \mathfrak{B}$ defined on the same probability space. Also, for a given $x \in \mathcal{X}$, we have the random variable $f(x, \zeta_h)$ and $f(x, \zeta)$ such that $\mathbb{E}[f(x, \zeta_h)] \to \mathbb{E}[f(x, \zeta)]$ as $h \to 0$. Moreover, throughout our discussion, we assume that for all $x \in \mathcal{X}$, $f(\cdot, \zeta)$ is Borel-measurable in ζ and is also Lipschitz continuous, *i.e.*,

Assumption 4.2.1. *For all $x_1, x_2 \in \mathcal{X}$ and for all $h \in \mathfrak{B} \cup \{0\}$,*

$$|f(x_1, \zeta_h) - f(x_2, \zeta_h)| \leq L_f |x_1 - x_2| \tag{4.2.3}$$

where L_f is the Lipschitz constant for any given ζ_h.

Lastly, we assume the existence of a unique solution to the optimization problem (4.2.1). Below, we present the definitions that would be relevant throughout our analysis and discussion.

Definition 4.2.1. *Let $\mathfrak{p}^* := \min_{x \in \mathcal{X}} \{F(x) = \mathbb{E}[f(x, \zeta)]\}$, then $x_\epsilon \in \mathcal{X}$ is said to the ϵ-optimal solution if*

$$F(x_\epsilon) \leq \mathfrak{p}^* + \epsilon \tag{4.2.4}$$

Definition 4.2.2. *For $v \in (0, 1)$ and $|\cdot|$ be the l_2 norm on $\mathcal{X}$, $\{x_k\}_{k=1}^{Q(v, |\cdot|, \mathcal{X})}$ is said to be a v-net of $\mathcal{X}$ if*

- *$x_k \in \mathcal{X}$ for all $k \in \{1, \ldots, Q(v, |\cdot|, \mathcal{X})\}$.*
- *$\forall x \in \mathcal{X}$ there exists $k(x) \in \{1, \ldots, Q(v, |\cdot|, \mathcal{X})\}$, such that $|x - x_{k(x)}| \leq v$.*

Below is an illustrative example that provides an intuitive explanation of covering number.

Example 4.2.1. *Let $\mathcal{X} = [0, 1] \subset \mathbb{R}$ and consider $v = 0.2$. A v-net for $\mathcal{X}$ is a finite subset $\{x_k\}_{k=1}^{Q}$ such that for every $x \in \mathcal{X}$, there exists x_k in the net such that,*

$$|x - x_k| \leq v.$$

In this example, we can choose the v-net as:

$$\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\},$$

which is a set of evenly spaced points with spacing $v = 0.2$. This implies that every point in $[0, 1]$ is within distance 0.2 of some point in the set. The covering number is:

$$Q(v, |\cdot|, [0, 1]) = 6.$$

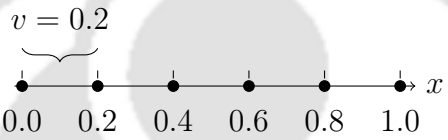


Figure 4.1: A v-net for $\mathcal{X} = [0, 1]$ with $v = 0.2$.

In empirical process theory, $Q(v, |\cdot|, \mathcal{X})$ is considered as the covering number, with ball size v, on the space $\mathcal{X}$. The following result gives an upper bound of the covering number.

Lemma 4.2.1. *Let $\mathcal{X}$ has a finite diameter i.e., $\mathbb{D}(\mathcal{X}) < \infty$, then for any $v \in (0, 1)$, there exists a v-net of $\mathcal{X}$ and size of that v-net is bounded, i.e., $Q(v, |\cdot|, \mathcal{X}) \leq \mathcal{O}((\mathbb{D}(\mathcal{X})/v)^d)$.*

Moreover, the Cramér's large deviation theorem will frequently used throughout our analysis, and we state it here as a lemma based on the discussion in [49, 75].

Lemma 4.2.2. *Let $X_1, \ldots, X_N$ be i.i.d samples of zero-mean random variable X with finite variance σ^2. For any $\epsilon > 0$, it holds that,*

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N} X_j \geq \epsilon\right) \leq \exp(-NI(\epsilon)),$$

where $I(\epsilon)$ is the rate function defined as $I(\epsilon) := \sup_{t \in \mathbb{R}}\{t\epsilon - \log(M(t))\}$, with $M(t) := \mathbb{E}[e^{tX}]$ being the moment generating function of X. Further, for any $\delta > 0$, there exists $\epsilon_1 > 0$, such that for any $\epsilon \in (0, \epsilon_1)$, $I(\epsilon) \geq \dfrac{\epsilon^2}{(2+\delta)\sigma^2}$.

We present below the pertinent results, that would be important for our RMSE analysis. To begin with, let us denote by.

$$\mathfrak{F} := \{f(x,\cdot) - \mathbb{E}[f(x,\cdot)] : x \in \mathcal{X}\} \tag{4.2.5}$$

the class of centered cost function indexed on $x \in \mathcal{X}$. Under the assumption that the function $f(x,\cdot)$ is Lipschitz with respect to x, it is fairly easy to observe that the centered cost function $f(x,\cdot) - \mathbb{E}[f(x,\cdot)]$ is also Lipschitz albeit with a larger Lipschitz constant. Further, let $l^\infty(\mathcal{X})$ be a metric space of all bounded functions from $\mathcal{X}$ to $\mathbb{R}$ endowed with the supremum norm, *i.e.*, $\|f-g\|_\infty := \sup_{x\in\mathcal{X}}|f(x)-g(x)|$ for all $f, g \in l^\infty(\mathcal{X})$. If we define,

$$\mathbb{F}_N(\cdot) = \sqrt{N}\left(\frac{1}{N}\sum_{k=1}^{N}(f(\cdot,\bar{\zeta}^k) - \mathbb{E}[f(\cdot,\bar{\zeta})])\right) \tag{4.2.6}$$

then $\mathbb{F}_N$ is an empirical process indexed on the decision set $\mathcal{X}$ and $\mathbb{F}_N \in l^\infty(\mathcal{X})$.

Since our study heavily relies on moment bounds, we present the results below that provide the necessary bounds. As the concept of covering numbers and bracketing numbers are relevant to these results, we present the definition of the bracketing numbers, where the covering number is already defined above.

Definition 4.2.3 (Bracketing numbers [85])**.** *Given two functions f_1 and f_2, the bracket $[f_1, f_2]$ is the set of all functions f with $f_1 \le f \le f_2$. An $v-$bracket is a bracket $[f_1, f_2]$ such that $\|f_1 - f_2\| < v$. The bracketing number $Q_{[]}(v, \mathfrak{F}, |\cdot|)$ is the minimum number of $v-$brackets needed to cover $\mathfrak{F}$.*

Below is an illustrative example that provides an intuitive explanation of bracketing number.

Example 4.2.2. *Let $\mathfrak{F} = \{f_a(x) = ax : a \in [0,1]\}$ be a class of linear functions over $x \in [0,1]$. To compute the bracketing number $Q_{[]}(v, \mathfrak{F}, \|\cdot\|_\infty)$, consider dividing the interval $[0,1]$ into $\lceil \frac{1}{v} \rceil$ subintervals of width v. Each bracket is defined as,*

$$[f_{a_k}, f_{a_k+v}], \quad \text{where } f_{a_k}(x) = a_k x, \quad a_k = kv, \quad k = 0, 1, \ldots, \left\lfloor \frac{1}{v} \right\rfloor.$$

Since $\|f_{a_k+v} - f_{a_k}\|_\infty = v$, each bracket has width at most v, and all functions in $\mathcal{F}$ are covered. Therefore,

$$Q_{[]}(v, \mathcal{F}, \|\cdot\|_\infty) \le \left\lceil \frac{1}{v} \right\rceil.$$

The next result gives an upper bound for the bracketing number.

Lemma 4.2.3 (Theorem 2.7.1 [85] and Lemma EC.7 [56]). *If the cost function $f(x,\bar{\zeta})$ is Lipschtiz with respect to x, and the decision space $\mathcal{X} \subseteq \mathbb{R}^d$ is compact, then for any $v > 0$*

$$Q_{[]}(4v\|L_f\|_2, \mathfrak{F}, \|\cdot\|_2) \leq Q(v, |\cdot|, \mathcal{X}) \tag{4.2.7}$$

Further, because of Lemma 4.2.1, we have an upper bound of the bracketing numbers. The following two results provide the relevant moment bound necessary for our analysis.

Lemma 4.2.4 (Lemma EC.9 [56]). *Let $\bar{f}(x,\bar{\zeta}) := \sup_{x\in\mathcal{X}} |f(x,\bar{\zeta}) - \mathbb{E}[f(x,\bar{\zeta})]|$. Then, for all N, we have*

$$\sqrt{N}\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{N}\sum_{k=1}^{N} f(x,\bar{\zeta}^k) - \mathbb{E}[f(x,\bar{\zeta})]\right|\right] \leq C\|\bar{f}(x,\bar{\zeta})\|_2 \int_0^1 \sqrt{1+\log(Q_{[]}(v\|\bar{f}(x,\bar{\zeta})\|_2, \mathfrak{F}, \|\cdot\|_2))} \tag{4.2.8}$$

Lemma 4.2.5 (Lemma EC.10 [56]). *For any $p \geq 2$ it holds that*

$$\sqrt{N}\left(\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{N}\sum_{k=1}^{N} f(x,\bar{\zeta}^k) - \mathbb{E}[f(x,\bar{\zeta})]\right|^p\right]\right)^{\frac{1}{p}} \leq$$
$$C\left(\sqrt{N}\mathbb{E}\left[\sup_{x\in\mathcal{X}}\left|\frac{1}{N}\sum_{k=1}^{N} f(x,\bar{\zeta}^k) - \mathbb{E}[f(x,\bar{\zeta})]\right|\right]\right) + C\left(N^{\frac{1}{p}-\frac{1}{2}}\|\bar{f}(x,\bar{\zeta})\|_p\right) \tag{4.2.9}$$

In Lemmas 4.2.4 and 4.2.5, C is a universal constant.

4.3 Monte Carlo SAA

In this section, we undertake a comprehensive study of the Monte Carlo SAA in the biased sampling framework. We begin our discussion by establishing the Uniform Convergence result followed by the sample complexity analysis to obtain $\epsilon-$optimal solution. We also derive a RMSE bound and determine the sample complexity required to render RMSE of $\mathcal{O}(\epsilon)$.

4.3.1 Uniform Convergence and Sample Complexity

This section starts by stating the assumptions necessary for our discussion.

Assumption 4.3.1. *For all $h \in \mathfrak{B}$, assume that $\sup_{x\in\mathcal{X}}|\mathbb{E}[f(x,\zeta_h) - f(x,\zeta)]| \leq c_1 h^\alpha$ for some $\alpha > 0$.*

Assumption 4.3.2. $\sigma^2 := \sup_{h\in\mathfrak{B}} \mathbb{E}[\sup_{x\in\mathcal{X}} |f(x,\zeta_h)|^2] < \infty$.

The above two assumptions illustrate the boundedness of the bias and variance, respectively. In the literature, these assumptions are commonly used for computational complexity analysis. Now, recall the optimization problem (4.1.1), *i.e.*,

$$\min_{x\in\mathcal{X}} \{F(x) = \mathbb{E}[f(x,\zeta)]\}, \tag{4.3.1}$$

the Monte Carlo SAA approximation of the above problem is given as,

$$\min_{x\in\mathcal{X}} \left\{ F_h^N(x) := \frac{1}{N}\sum_{j=1}^{N} f(x,\zeta_h^j) \right\}. \tag{4.3.2}$$

Suppose we let x^* and x_h^* be optimal solutions to (4.1.1) and (4.3.2) respectively, then we are interested in determining the probability of x_h^* being an ϵ-optimal solution to (4.1.1). More specifically, we intend to determine $\mathbb{P}(F(x_h^*) - F(x^*) \leq \epsilon)$ for some $\epsilon > 0$. In order to do so, we begin by establishing the uniform convergence property based on concentration inequality. For the proof of the next result, we follow the approach in [49, 79]

Theorem 4.3.1. *(Uniform Convergence) Suppose assumptions 4.2.1, 4.3.1, and 4.3.2 are satisfied. Also assume that there exists a neighbourhood $\mathcal{N}_\mathfrak{B}$ of zero such that for any $x \in \mathcal{X}$, and for all $u \in \mathcal{N}_\mathfrak{B}$ we have,*

$$\sup_{h\in\mathfrak{B}} \mathbb{E}\left[e^{u(f(x,\zeta_h)-\mathbb{E}[f(x,\zeta_h)])}\right] < \infty. \tag{4.3.3}$$

Then for any $\delta > 0$, there exists an $\epsilon_1 > 0$ such that for all $\epsilon \in (0,\epsilon_1)$ and for all $h \in \mathfrak{B}$, satisfying $c_1 h^\alpha \leq \epsilon/4$, we have,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}|F_h^N(x) - F(x)| > \epsilon\right) \leq \mathcal{O}(1)\left(\frac{4L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d \exp\left(\frac{-N\epsilon^2}{16(\delta+2)\sigma^2}\right). \tag{4.3.4}$$

Proof. We begin with the construction of a v-net in order to get rid of the supremum over x. For this consider a v-net on $\mathcal{X}$ such that, $v = \dfrac{\epsilon}{4L_f}$ and thus $Q \leq \mathcal{O}(1)\left(\dfrac{4L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d$.

Further, by invoking the Lipschitz continuity of $f(\cdot, \zeta)$, we have,

$$|f(x, \zeta_h) - f(x_k, \zeta_h)| \leq \frac{\epsilon}{4} \text{ and } |F(x) - F(x_k)| \leq \frac{\epsilon}{4}. \tag{4.3.5}$$

Therefore, for any $x \in \mathcal{X}$ we have,

$$|F_h^N(x) - F(x)| \leq \frac{\epsilon}{2} + |F_h(x_k) - F(x_k)| \leq \frac{\epsilon}{2} + \max_{k=1,\ldots,Q} |F_h(x_k) - F(x_k)|.$$

Consequently, we have,

$$\begin{aligned} \mathbb{P}\left(\sup_{x \in \mathcal{X}} |F_h^N(x) - F(x)| > \epsilon \right) &\leq \mathbb{P}\left(\max_{k=1,\ldots,Q} |F_h^N(x_k) - F(x_k)| > \frac{\epsilon}{2} \right) \\ &\leq \sum_{k=1}^{Q} \mathbb{P}\left(|F_h^N(x_k) - F(x_k)| > \frac{\epsilon}{2} \right). \end{aligned} \tag{4.3.6}$$

In order to examine the $\mathbb{P}\left(|F_h^N(x_k) - F(x_k)| > \frac{\epsilon}{2} \right)$, we define the random variable $Z_h^j(k)$ as $Z_h^j(k) := f(x_k, \zeta_h^j) - F(x_k)$, and let $\mathbb{E}[Z_h(k)]$ be its respective expectation. It is easy to observe that $Z_h^j(k) - \mathbb{E}[Z_h(k)]$ is zero-mean random variable. Now if $0 < \mathbb{E}[Z_h(k)] \leq c_1 h^\alpha \leq \epsilon/4$, then,

$$\begin{aligned} \mathbb{P}\left(F_h^N(x_k) - F(x_k) > \frac{\epsilon}{2} \right) &= \mathbb{P}\left(\frac{1}{N} \sum_{j=0}^{N} Z_h^j(k) > \frac{\epsilon}{2} \right) \\ &\leq \mathbb{P}\left(\frac{1}{N} \sum_{j=0}^{N} (Z_h^j(k) - \mathbb{E}[Z_h(k)]) > \frac{\epsilon}{4} \right). \end{aligned} \tag{4.3.7}$$

Now, as a consequence of equation (4.3.3), Lemma 4.2.2, and Assumption 4.3.2, we have,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{N} \sum_{j=0}^{N} (Z_h^j(k) - \mathbb{E}[Z_h(k)]) > \frac{\epsilon}{4} \right) &\leq \exp\left(\frac{-N\epsilon^2}{16(\delta+2)\mathbb{V}[Z_h(k)]} \right) \\ &\leq \exp\left(\frac{-N\epsilon^2}{16(\delta+2)\sigma^2} \right). \end{aligned} \tag{4.3.8}$$

Similarly, if $0 < -\mathbb{E}[Z_h(k)] \leq c_1 h^\alpha \leq \epsilon/4$, then,

$$\mathbb{P}\left(\frac{1}{N}\sum_{j=0}^{N}(\mathbb{E}[Z_h(k)] - Z_h^j(k)) > \frac{\epsilon}{4}\right) \leq \exp\left(\frac{-N\epsilon^2}{16(\delta+2)\sigma^2}\right). \tag{4.3.9}$$

Finally, putting everything together, we have,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}|F_h^N(x) - F(x)| > \epsilon\right) \leq \mathcal{O}(1)\left(\frac{4L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d \exp\left(\frac{-N\epsilon^2}{16(\delta+2)\sigma^2}\right).$$

□

The following corollary is the immediate consequence of the above result,

Corollary 4.3.1. *Let assumptions of Theorem 4.3.1 hold, and additionally assume $c_1 h^\alpha \leq \epsilon/8$, then for any $\delta > 0$, there exists $\epsilon_1 > 0$ such that for all $\epsilon \in (0, \epsilon_1)$,*

$$\mathbb{P}\left(F(x_h^*) - F(x^*) > \epsilon\right) \leq \mathcal{O}(1)\left(\frac{8L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d \exp\left(\frac{-N\epsilon^2}{64(\delta+2)\sigma^2}\right). \tag{4.3.10}$$

Proof. To begin with, observe that $F_h^N(x_h^*) - F_h^N(x^*) \leq 0$. Now,

$$\begin{aligned}\mathbb{P}\left(F(x_h^*) - F(x^*) \geq \epsilon\right) &= \mathbb{P}\left([F(x_h^*) - F_h^N(x_h^*)] + [F_h^N(x_h^*) - F_h^N(x^*)] + [F_h^N(x^*) - F(x^*)] \geq \epsilon\right) \\ &\leq \mathbb{P}\left(F(x_h^*) - F_h^N(x_h^*) \geq \epsilon/2\right) + \mathbb{P}\left(F_h^N(x^*) - F(x^*) \geq \epsilon/2\right).\end{aligned} \tag{4.3.11}$$

Invoking Theorem 4.3.1 with the condition $c_1 h^\alpha \leq \epsilon/8$, we get the desired result. □

Now, let us assume the cost of generating a single sample of $Y_h(x)$ is $\bar{\eta}/h$ with $\bar{\eta}$ begin some proportionality constant. As the simulation requires generating N samples, the total computational cost will be $\mathcal{C} := N\bar{\eta}/h$. Also, let the probability of the solution to the Monte Carlo SAA problem being ϵ-optimal to the original problem defined in (4.1.1) be at least $(1 - \epsilon^\gamma)$ for some $\gamma > 0$. The next result illustrates the computational complexity required to achieve an ϵ-optimal solution.

Theorem 4.3.2. *(Computational Complexity) Let $\epsilon < 1/e$ and $\gamma > 0$. Then under the assumption stated in Corollary 4.3.1, the computational complexity for achieving the ϵ-optimal solution with the probability at least $1 - \epsilon^\gamma$ is $\mathcal{O}\left((d+\gamma)\log(\epsilon^{-1})\epsilon^{-\left(2+\frac{1}{\alpha}\right)}\right)$.*

Proof. Since we require $\mathbb{P}\left(F(x_h^*) - F(x^*) \leq \epsilon\right) \leq 1 - \epsilon^\gamma$, we need $\mathbb{P}\left(F(x_h^*) - F(x^*) > \epsilon\right) < \epsilon^\gamma$. Observe that if we take,

$$N = \left\lceil \frac{\mathcal{O}(1)64(\delta+2)\sigma^2}{\epsilon^2}\left[d\log\left(\frac{8L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right) + \log\left(\frac{1}{\epsilon^\gamma}\right)\right]\right\rceil, \tag{4.3.12}$$

then

$$\mathcal{O}(1)\left(\frac{8L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d \exp\left(\frac{-N\epsilon^2}{64(\delta+2)\sigma^2}\right) \leq \epsilon^\gamma, \tag{4.3.13}$$

and as a consequence of corollary 4.3.1, we have the desired probability. Now, the computational cost is defined as $\mathcal{C} = N\bar{\eta}/h$. Therefore, taking,

$$N = \frac{\mathcal{O}(1)64(\delta+2)\sigma^2}{\epsilon^2}\left[d\log\left(\frac{8L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right) + \log\left(\frac{1}{\epsilon^\gamma}\right)\right] + 1, \tag{4.3.14}$$

we get,

$$\mathcal{C} = \bar{\eta}\left(\frac{\mathcal{O}64(1)(\delta+2)\sigma^2}{\epsilon^2 h}\left[d\log\left(\frac{8L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right) + \log\left(\frac{1}{\epsilon^\gamma}\right)\right] + \frac{1}{h}\right). \tag{4.3.15}$$

If $h = c_2\epsilon^{1/\alpha}$, so that $c_1 h^\alpha \leq \epsilon/8$ (required to achieve the bound in Corollary 4.3.1), then for $\epsilon < \frac{1}{e}$, observe that,

$$\frac{1}{h} < c_3(\gamma+d)\log(\epsilon^{-1})\epsilon^{-\left(2+\frac{1}{\alpha}\right)}, \tag{4.3.16}$$

where, $c_3 = \frac{1}{c_2}$. Further, observe that, for $\epsilon < \frac{1}{e}$,

$$d\log\left(\frac{8L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right) + \log\left(\frac{1}{\epsilon^\gamma}\right) < c_4(d+\gamma)\log(\epsilon^{-1}), \tag{4.3.17}$$

where $c_4 = (1 + \max\{0, \log(8L_f\mathbb{D}(\mathcal{X}))\})$. Finally, putting everything together, we get

$$\mathcal{C} < \bar{\eta}\left(\frac{\mathcal{O}(1)64(\delta+2)\sigma^2}{\epsilon^2 h}c_4(d+\gamma)\log(\epsilon^{-1}) + c_3(\gamma+d)\log(\epsilon^{-1})\epsilon^{-\left(2+\frac{1}{\alpha}\right)}\right). \tag{4.3.18}$$

Substituting $h = c_2\epsilon^{1/\alpha}$ in the above inequality, we get,

$$\mathcal{C} < c_5\left((\gamma+d)\log(\epsilon^{-1})\epsilon^{-\left(2+\frac{1}{\alpha}\right)}\right), \tag{4.3.19}$$

with $c_5 = \bar{\eta}\left(\dfrac{\mathcal{O}(1)64(\delta+2)\sigma^2}{c_2}c_4 + c_3\right)$. □

The above analysis shows the effect bias parameter h and the convergence rate α have on the computational complexity of the Monte Carlo SAA. It is easy to observe that as $\alpha \to \infty$, *i.e.* as the bias converges to zero, the computational complexity tends to $\mathcal{O}\left((\gamma+d)\log(\epsilon^{-1})\epsilon^{-2}\right)$, same as the one observed with an unbiased estimator. The complexity results for the conditional stochastic optimization can also be recreated in the context of the above analysis. For instance, the mean-squared-error analysis performed by in section 3.2 of [49] shows that the order of bias convergence is 1/2, *i.e.* $\alpha = 1/2$. Substituting this α in our above analysis, we obtained the computational complexity similar to the one achieved by authors, *i.e.*, $\mathcal{O}\left((\gamma+d)\log(\epsilon^{-1})\epsilon^{-4}\right)$.

4.3.2 Root Mean Squared Error Analysis

Despite the above analysis elucidating the uniform convergence of Monte Carlo SAA in the biased framework, the practical realization of such integration poses notable challenges. One challenge stems from relatively strong but unavoidable assumptions of the existence of a finite valued moment-generating function in the neighbourhood of zero (see (4.3.3)). This assumption is essential to determine the sample complexity required to achieve an exponential rate of convergence. Although [8] studied the convergence in the almost sure sense, the extension to the biased setup is beyond the scope of this thesis. For now, we direct our attention to RMSE analysis, consequently deriving sample complexity without such strong assumptions.

The next result provides an RMSE bound for the optimal value of the SAA problem defined in equation (4.3.2), defined as,

$$\text{RMSE} := \left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathfrak{p}^* \right\|_2 \tag{4.3.20}$$

where $\mathfrak{p}^*$ is the optimal value of the original SAA problem.

Theorem 4.3.3. *Suppose assumptions 4.2.1, 4.3.1 and 4.3.2 holds. Then for any $N \in \mathbb{N}$.*

$$\left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathfrak{p}^* \right\|_2 \leq c_1 h^\alpha + \mathfrak{c}_3 \frac{\sigma}{\sqrt{N}}. \tag{4.3.21}$$

Proof. Let $\mathfrak{p}^{*,h} = \min_{x\in\mathcal{X}} \mathbb{E}[f(x,\zeta_h)]$. Observe that,

$$\left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathfrak{p}^* \right\|_2 = \left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathfrak{p}^{*,h} + \mathfrak{p}^{*,h} - \mathfrak{p}^* \right\|_2$$
$$\leq \left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathfrak{p}^{*,h} \right\| + \left\| \mathfrak{p}^{*,h} - \mathfrak{p}^* \right\|_2. \quad (4.3.22)$$

For the second term in the above inequality, we have,

$$\left\| \mathfrak{p}^{*,h} - \mathfrak{p}^* \right\|_2 \leq \left(\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\mathbb{E}[f(x,\zeta_h)] - \mathbb{E}[f(x,\zeta)]|^2 \right] \right)^{1/2} \leq c_1 h^{\alpha}. \quad (4.3.23)$$

As for the first term, we have,

$$\left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \min_{x\in\mathcal{X}} \mathbb{E}[f(x,\zeta_h)] \right\|_2 = \left(\mathbb{E}\left[\left(\min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \min_{x\in\mathcal{X}} \mathbb{E}[f(x,\zeta_h)] \right)^2 \right] \right)^{1/2}$$
$$\leq \left(\mathbb{E}\left[\sup_{x\in\mathcal{X}} \left(\left| \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathbb{E}[f(x,\zeta_h)] \right| \right)^2 \right] \right)^{1/2}. \quad (4.3.24)$$

Now for a given $h \in \mathfrak{B}$, we define $\mathbb{F}_N^h$ as,

$$\mathbb{F}_N^h(\cdot) = \sqrt{N} \left(\frac{1}{N}\sum_{k=1}^{N} (f(\cdot,\zeta_h^k) - \mathbb{E}[f(\cdot,\zeta_h)]) \right), \quad (4.3.25)$$

is an empirical process. Then, under the assumptions 4.3.1 and 4.3.2 and a direct application of Lemma 4.2.3,4.2.4 and 4.2.5, we have,

$$\sqrt{\mathbb{E}\left[\sup_{x\in\mathcal{X}} \left(\left| \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathbb{E}[f(x,\zeta_h)] \right| \right)^2 \right]} \leq \mathfrak{c}_3 \frac{\sigma}{\sqrt{N}}, \quad (4.3.26)$$

where $\mathfrak{c}_3 = C\left(\int_0^1 \sqrt{1 + \log(Q_{\square}(v\|\bar{f}(x,\bar{\zeta})\|_2, \mathfrak{F}, \|\cdot\|_2))} + 1 \right) < \infty$. By collating all expressions and reassigning constants, we get the desired result. □

Corollary 4.3.2. *The computational complexity required for* $\left\| \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathfrak{p}^* \right\|_2 \leq \epsilon$ *is* $\mathcal{O}\left(\epsilon^{-\left(2+\frac{1}{\alpha}\right)}\right)$.

Proof. Take $h = \mathcal{O}(\epsilon^{-\frac{1}{\alpha}})$ and $N = \mathcal{O}(\epsilon^{-2})$. Since the computational cost is given as $\frac{N}{h}$, we get the desired result. □

4.4 Multilevel Monte Carlo SAA

The basic idea behind the multilevel extension of the Monte Carlo SAA deals with the multilevel approximation of expectation associated with the minimization problem. To this end, we assume the availability of level l approximation of the random variable ζ, denoted by ζ_l, such that as $l \to \infty$, $\zeta_l \to \zeta$. Therefore, the multilevel extension of the optimization problem is given as,

$$\min_{x\in\mathcal{X}} \left\{ F_L(x) = \sum_{l=1}^{L} \frac{1}{N_l} \sum_{j=1}^{N_l} \left(f(x,\zeta_l^j) - f(x,\zeta_{l-1}^j) \right) \right\}.$$

Let x_L^* solve the optimization problem stated above and x^* solve the original problem (4.1.1). As before, we aim to determine the probability of x_L^* being the ϵ-optimal solution. We conduct an analysis similar to the one performed in Section 4.3.1 to study the uniform convergence and computational complexity in the MLMC paradigm. To this end, let $m > 2$ be the refinement factor, and let,

$$h_l = \frac{\mathbf{h}}{m^{l-1}}, \quad l = 1, \ldots, L, \tag{4.4.1}$$

where $\mathbf{h} \in \mathfrak{B}$ is the coarsest bias parameter that may depend on the problem under consideration. We begin our discussion by stating some technical assumptions,

Assumption 4.4.1. $E_l := \sup_{x\in\mathcal{X}} |\mathbb{E}\left(f(x,\zeta_l) - f(x,\zeta)\right)| \leq c_1 h_l^{\alpha}$.

Assumption 4.4.2. $V_l := \mathbb{E}[\sup_{x\in\mathcal{X}} |f(x,\zeta_l) - f(x,\zeta_{l-1})|^2] \leq c_2 h_l^{\beta}$.

Theorem 4.4.1. *(Uniform Convergence) Suppose assumption 4.2.1, 4.4.1 and 4.4.2 holds. Further for* $h, h' \in \mathfrak{B}$ *assume that there exists a neighbourhood* $\mathcal{N}_{hh'}$ *of zero such*

that for any $x \in \mathcal{X}$, and for all $u \in \mathcal{N}_{hh'}$ we have,

$$\mathbb{E}\left[e^{u(f(x,\zeta_h)-f(x,\zeta_{h'})-\mathbb{E}[f(x,\zeta_h)-f(x,\zeta_{h'})])}\right] < \infty. \tag{4.4.2}$$

Then for any $\delta > 0$, there exists an $\epsilon_1 > 0$ such that for all $\epsilon \in (0, \epsilon_1)$ and for any $r > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}|F_L(x) - F(x)| > \epsilon\right) \leq \mathcal{O}(1)\left(\frac{4(2L+1)L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d \sum_{l=1}^{L} \exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{16m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^\beta}\right). \tag{4.4.3}$$

Proof. We begin with the construction of a v-net in order to get rid of the supremum over x. We first pick a v-net on $\mathcal{X}$ such that, $v = \dfrac{\epsilon}{4(2L+1)L_f}$. Thus $Q \leq \mathcal{O}(1)\left(\dfrac{4(2L+1)L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d$. Further, by invoking the Lipschitz continuity of $h(\cdot,\zeta)$, we have,

$$|f_L(x) - f_L(x_k)| \leq \frac{\epsilon}{4} \text{ and } |F(x) - F(x_k)| \leq \frac{\epsilon}{4} \tag{4.4.4}$$

Therefore, for any $x \in \mathcal{X}$, we have,

$$|F_L(x) - F(x)| \leq \frac{\epsilon}{2} + |F_L(x_k) - F(x_k)| \leq \frac{\epsilon}{2} + \max_{k=1,\ldots,Q}|F_L(x_k) - F(x_k)| \tag{4.4.5}$$

Consequently, we have,

$$\begin{aligned}\mathbb{P}\left(\sup_{x\in\mathcal{X}}|F_L(x) - F(x)| > \epsilon\right) &\leq \mathbb{P}\left(\max_{k=1,\ldots,Q}|F_L(x_k) - F(x_k)| > \frac{\epsilon}{2}\right) \\ &\leq \sum_{k=1}^{Q}\mathbb{P}\left(|F_L(x_k) - F(x_k)| > \frac{\epsilon}{2}\right).\end{aligned} \tag{4.4.6}$$

Let us suppose $\mathbb{E}[F_L(x_k) - F(x_k)] \leq \epsilon/4$, then,

$$\mathbb{P}\left(F_L(x_k) - F(x_k) > \frac{\epsilon}{2}\right) \leq \mathbb{P}\left(F_L(x_k) - F(x_k) - \mathbb{E}[F_L(x_k) - F(x_k)] > \frac{\epsilon}{4}\right). \tag{4.4.7}$$

Now, under the multilevel paradigm, we define the random variable $Z_l^j(k)$ as follows,

$$Z_l^j(k) = \begin{cases} f(x_k,\zeta_1^j) - F(x_k), l = 1 \\ f(x_k,\zeta_l^j) - f(x_k,\zeta_{l-1}^j), l = 2,\ldots,L. \end{cases} \tag{4.4.8}$$

and denote $\mathbb{E}[Z_l(k)]$ as its expectation. Now, with the above notational considerations, we have,

$$\mathbb{P}\bigg(F_L(x_k) - F(x_k) - \mathbb{E}[F_L(x_k) - F(x_k)] > \frac{\epsilon}{4}\bigg) = \mathbb{P}\bigg(\sum_{l=1}^{L}\frac{1}{N_l}\sum_{j=1}^{N_l}(Z_l^j(k) - \mathbb{E}[Z_l(k)]) > \frac{\epsilon}{4}\bigg). \tag{4.4.9}$$

Now, consider the following sets,

$$O = \bigg\{\sum_{l=1}^{L}\frac{1}{N_l}\sum_{j=1}^{N_l}(Z_l^j(k) - \mathbb{E}[Z_l(k)]) > \frac{\epsilon}{4}\bigg\} \text{ and,}$$
$$O_l = \bigg\{\frac{1}{N_l}\sum_{j=1}^{N_l}(Z_l^j(k) - \mathbb{E}[Z_l(k)]) > \frac{\epsilon}{4}\frac{(m^r-1)}{m^r}\frac{1}{m^{r(l-1)}}\bigg\}$$

and further observe that $O \subseteq \bigcup_{l=1}^{L} O_l$. Then, by finite sub-additivity of the probability measure, we have, $\mathbb{P}(O) \leq \sum_{l=1}^{L}\mathbb{P}(O_l)$. As $Z_l - \mathbb{E}[Z_l(k)]$ is a zero-mean random variable, and observing that $\mathbb{V}[Z_l(k)] \leq c_2 h_l^{\beta}$, we have,

$$\mathbb{P}(O_l) \leq \exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{16(m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^{\beta}}\right). \tag{4.4.10}$$

Therefore, we have,

$$\mathbb{P}\bigg(F_L(x_k) - F(x_k) > \frac{\epsilon}{2}\bigg) \leq \sum_{l=1}^{L}\exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{16m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^{\beta}}\right). \tag{4.4.11}$$

Similarly, if $\mathbb{E}[F_L(x_k) - F(x_k)] \geq -\epsilon/4$, then,

$$\mathbb{P}\bigg(F(x_k) - F_L(x_k) > \frac{\epsilon}{2}\bigg) \leq \sum_{l=1}^{L}\exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{16m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^{\beta}}\right), \tag{4.4.12}$$

and hence, we have,

$$\mathbb{P}\bigg(|F_L(x_k) - F(x_k)| > \frac{\epsilon}{2}\bigg) \leq \sum_{l=1}^{L}\exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{16m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^{\beta}}\right). \tag{4.4.13}$$

Putting everything together, we have,

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}|F_L(x)-F(x)|>\epsilon\right)\leq\ \mathcal{O}(1)\left(\frac{4(2L+1)L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d\sum_{l=1}^{L}\exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{16m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^\beta}\right).$$

□

The following corollary is the immediate consequence of the above result.

Corollary 4.4.1. *Let assumptions of Theorem 4.4.1 holds, and additionally assume* $\sup_{x\in\mathcal{X}}|\mathbb{E}[F_L(x)-F(x)]|\leq\epsilon/8$, *then for any* $\delta>0$, *there exists* $\epsilon_1>0$ *such that for all* $\epsilon\in(0,\epsilon_1)$ *and for any* $r>0$,

$$\mathbb{P}\left(F(x_L^*)-F(x^*)>\epsilon\right)\leq\mathcal{O}(1)\left(\frac{8(2L+1)L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d\sum_{l=1}^{L}\exp\left(\frac{-N_l\epsilon^2(m^r-1)^2}{64m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^\beta}\right). \tag{4.4.14}$$

The proof of the above corollary follows the line of argument similar to the proof of corollary 4.3.1 and is therefore skipped.

Now that we have established the convergence results, we pivot our focus towards sample complexity. As previously indicated, let $\bar{\eta}/h_l$ denote the cost associated with generating a single sample of $f(x,\zeta_l)$ at level l. Thus, the computational expenditure for generating N_l samples would amount to $\bar{\eta}N_l/h_l$. Given that we generate samples across levels $l=1,\ldots,L$, the aggregate computational cost is delineated as:

$$\mathcal{C}_{mlmc}^{saa}:=\bar{\eta}\sum_{l=1}^{L}\frac{N_l}{h_l}. \tag{4.4.15}$$

The following result estimates the computational complexity associated with the multi-level estimator.

Theorem 4.4.2. *(Computational Complexity) Let the probability of the solution to the multilevel SAA problem being* ϵ*-optimal to the original problem be at least* $1-\epsilon^\gamma$, *for some* $\gamma>0$. *Further, let the assumptions of Theorem 4.4.1 and corollary 4.4.1 hold. Also, assume the existence of positive constants* α,β, *and* m *with* $\alpha\geq1/2$ *and* $m>2$. *Then, there exists an* $r>0$ *such that the computational complexity associated with achieving*

the ϵ-optimal solution is

$$\mathcal{C}^{saa}_{mlmc} = \begin{cases} \mathcal{O}\left((\gamma+d)\epsilon^{-2}\log(\epsilon^{-1})\right), & \text{for } \beta > 1. \\ \mathcal{O}\left((\gamma+d)\epsilon^{-\left(2+\frac{\log(2)}{\alpha\log(m)}\right)}\log(\epsilon^{-1})\right), & \text{for } \beta = 1. \\ \mathcal{O}\left((\gamma+d)\epsilon^{-\left(2+\frac{1-\beta}{\alpha}\right)}\log(\epsilon^{-1})\right), & \text{for } \beta < 1. \end{cases} \tag{4.4.16}$$

Proof. Following the same line of argument as in Theorem 4.3.2, observe that if

$$N_l = \left\lceil \frac{64m^{2r}m^{2r(l-1)}(\delta+2)c_2h_l^{\beta}}{\epsilon^2(m^r-1)^2}\log\left(\frac{\mathcal{A}}{\epsilon^{\gamma}}(L+1)\right)\right\rceil, \text{ where } \mathcal{A} = \mathcal{O}(1)\left(\frac{8(2L+1)L_f\mathbb{D}(\mathcal{X})}{\epsilon}\right)^d,$$

then $\mathbb{P}\left(F(x_L^*) - F(x^*) > \epsilon\right) \leq \epsilon^{\gamma}$. Therefore, we now have a formulation for the number of samples required on various levels of resolution. As for the number of levels and computational complexity, we separately analyze different cases. We present below a general expression for the computational cost based on the formulation in (4.4.15) that would be relevant throughout our analysis.

$$\mathcal{C}^{saa}_{mlmc} \leq \bar{\eta}\left(\frac{64m^{2r}(\delta+2)c_2}{\epsilon^2(m^r-1)^2}\log\left(\frac{\mathcal{A}}{\epsilon^{\gamma}}(L+1)\right)\sum_{l=1}^{L}m^{2r(l-1)}h_l^{\beta-1} + \sum_{l=1}^{L}\frac{1}{h_l}\right) \tag{4.4.17}$$

To begin with, let,

$$L = \left\lceil \frac{\log(8c_1\mathbf{h}^{\alpha}\epsilon^{-1})}{\alpha\log(m)} \right\rceil, \tag{4.4.18}$$

. Then,

$$L+1 \leq \frac{\log(8c_1\mathbf{h}^{\alpha}\epsilon^{-1})}{\alpha\log(m)} + 2 \leq c_3(\log(\epsilon^{-1})), \tag{4.4.19}$$

where $c_3 = \left(\frac{1}{\alpha\log(m)} + \max\left\{0, \frac{\log(8c_1\mathbf{h}^{\alpha})}{\alpha\log(m)}\right\} + 2\right)$. Also, for the above L, it is easy to observe that,

$$\frac{\epsilon}{8c_1m^{\alpha}} \leq h_L^{\alpha} < \frac{\epsilon}{8c_1}, \tag{4.4.20}$$

thereby satisfying the bias conditions required for uniform convergence. Further,

$$\log\left(\frac{\mathcal{A}}{\epsilon^{\gamma}}(L+1)\right) \leq c_4(\gamma+d)\log(\epsilon^{-1}), \tag{4.4.21}$$

where $c_4 = (3+\log(c_3 16 L_f \mathbb{D}(\mathcal{X})))$. And also, for $\alpha \geq 1/2$ and $\epsilon < 1/e$,

$$\sum_{l=1}^{L}\frac{1}{h_l} < \epsilon^{-2}c_5 < (\gamma+d)\epsilon^{-2}\log(\epsilon^{-1})c_5, \tag{4.4.22}$$

with $c_5 = \dfrac{(8c_1)^{1/\alpha}m^2}{m-1}$. Now all that is left for us to analyse is $\sum_{l=1}^{L} m^{2r(l-1)}h_l^{\beta-1}$. We will handle this case by case.

Case 1 ($\beta = 1$)**:** For $\beta = 1$, we have $\sum_{l=1}^{L} m^{2r(l-1)}h_l^{\beta-1} < \dfrac{m^{2r(L+1)}}{m^{2r}-1}$. Substituting the upper bound of $L+1$, we get,

$$\frac{m^{2r(L+1)}}{m^{2r}-1} \leq (8c_1\mathbf{h}^{\alpha})^{2r/\alpha}\epsilon^{-2r/\alpha}\frac{m^{2r}}{m^{2r}-1}. \tag{4.4.23}$$

Choosing $r = \dfrac{\log(2)}{2\log(m)}$ we get that $\dfrac{m^{2r}}{m^{2r}-1} = 2$. Consequently, we have,

$$\frac{m^{2r(L+1)}}{m^{2r}-1} \leq c_6\epsilon^{-\frac{\log(2)}{\log(m)\alpha}}, \tag{4.4.24}$$

where $c_6 = 2(8c_1\mathbf{h}^{\alpha})^{\frac{\log(2)}{\log(m)\alpha}}$.

Case 2 ($\beta > 1$)**:** For this case we assume $2r < \beta - 1$. Then we have,

$$\begin{aligned}\sum_{l=1}^{L} m^{2r(l-1)}h_l^{\beta-1} &= \mathbf{h}^{\beta-1}\sum_{l=1}^{L}\frac{1}{m^{(\beta-1-2r)(l-1)}} \\ &\leq \mathbf{h}^{\beta-1}(1-m^{-(\beta-1-2r)})^{-1}.\end{aligned} \tag{4.4.25}$$

As before, choosing $r = \dfrac{\beta-1}{4}$, we get,

$$\sum_{l=1}^{L} m^{2r(l-1)}h_l^{\beta-1} \leq (1-m^{-(\frac{\beta-1}{2})})^{-1}\mathbf{h}^{\beta-1}. \tag{4.4.26}$$

Case 3 ($\beta < 1$)**:** In this case choosing $r = \dfrac{1-\beta}{4}$, we have,

$$\begin{aligned}\sum_{l=1}^{L} m^{2r(l-1)} h_l^{\beta-1} &= h_L^{-(1-\beta)} \sum_{l=1}^{L} \frac{1}{m^{(1-\beta-2r)(l-1)}} \\ &< h_L^{-(1-\beta)} (1 - m^{-(1-\beta-2r)})^{-1} \\ &\leq \epsilon^{-(1-\beta)/\alpha} (8c_1)^{(1-\beta)/\alpha} (1 - m^{-(\frac{1-\beta}{2})})^{-1}.\end{aligned} \tag{4.4.27}$$

By collating all the cases, we get the desired result. □

4.4.1 Root Mean Squared Error Analysis

In this section, we undertake the RMSE analysis of the biased SAA problem in the MLMC setting. We start our discussion with a technical lemma that would be essential in the remainder of the study.

Lemma 4.4.1. *Suppose* $0 < \gamma < \infty$ *and* $c, d > 0 \in \mathbb{R}$. *Then for all* $x \in (0, d]$, *there exists a constant* $\mathfrak{K}$ *such that,*

$$x \left(1 + \sqrt{c \log\left(\frac{d}{x}\right)}\right) \leq \mathfrak{K} x^{\frac{1}{1+\gamma}} \tag{4.4.28}$$

Proof. Let $\bar{\gamma} = 1 - \dfrac{1}{1+\gamma}$ then $0 < \bar{\gamma} < 1$. Consider the function,

$$g(x) = \begin{cases} x^{\bar{\gamma}} \left(1 + \sqrt{c \log\left(\frac{d}{x}\right)}\right) & , x \in (0, d] \\ 0 & , x = 0 \end{cases}, \tag{4.4.29}$$

then it is easy to see that $g(x) \geq 0$ is continuous on $[0, d]$. Since $[0, d]$ is compact, we have $x^* \in [0, d]$ such that $g(x^*) = \max\limits_{x \in [0,d]} g(x)$. The results follows by letting $\mathfrak{K} = g(x^*)$.

□

In the multilevel setting, we let, let $g(x, \bar{\zeta}_l) := f(x, \zeta_l) - f(x, \zeta_{l-1})$, and define $\mathfrak{p}^{*,L} := \min_{x \in \mathcal{X}} \mathbb{E}[f(x, \zeta^L)] = \min_{x \in \mathcal{X}} \sum_{l=0}^{L} \mathbb{E}[g(x, \bar{\zeta}_l)]$. Further, we define,

$$\hat{\mathfrak{p}}^{*,L} := \min_{x \in \mathcal{X}} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(x, \bar{\zeta}_l^k), \tag{4.4.30}$$

as the Monte Carlo approximation of $\mathfrak{p}^{*,L}$. The RMSE error in this case is given as,

$$\text{RMSE}_{MLMC} := \left\| \min_{x\in\mathcal{X}} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(x, \bar{\zeta}_l^k) - \mathfrak{p}^* \right\|_2 = \left\| \hat{\mathfrak{p}}^{*,L} - \mathfrak{p}^* \right\|_2. \tag{4.4.31}$$

The following result provides an RMSE error bound for the optimal value obtained by solving MLMC-SAA defined in (4.2.1).

Theorem 4.4.3. *Suppose assumptions 4.2.1, 4.4.1 and 4.4.2 holds. Then for any* $0 < a < 1$, $L \geq 2$.

$$\left\| \hat{\mathfrak{p}}^{*,L} - \mathfrak{p}^* \right\|_2 \leq c_1 h_L^{\alpha} + 2c_2\bar{\mathfrak{c}} \sum_{l=1}^{L} \frac{h_l^{\bar{\beta}/2}}{\sqrt{N_l}} \tag{4.4.32}$$

where $\bar{\beta} = \beta\dfrac{1}{1+a}$.

Proof. To begin with, observe that from triangle inequality, we have,

$$\left\| \hat{\mathfrak{p}}^{*,L} - \mathfrak{p}^* \right\|_2 \leq \left\| \hat{\mathfrak{p}}^{*,L} - \mathfrak{p}^{*,L} \right\|_2 + \left\| \mathfrak{p}^{*,L} - \mathfrak{p}^* \right\|_2. \tag{4.4.33}$$

As before, we have,

$$\left\| \mathfrak{p}^{*,L} - \mathfrak{p}^* \right\|_2 \leq \left(\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\mathbb{E}[f(x,\zeta_L)] - \mathbb{E}[f(x,\zeta)]|^2 \right] \right)^{1/2} \leq c_1 h_L^{\alpha}. \tag{4.4.34}$$

Now let us analyse $\left\| \hat{\mathfrak{p}}^{*,L} - \mathfrak{p}^{*,L} \right\|_2$. Observe that,

$$\begin{aligned}
\left\| \hat{\mathfrak{p}}^{*,L} - \mathfrak{p}^{*,L} \right\|_2 &= \left\| \min_{x\in\mathcal{X}} \left(\sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(x, \bar{\zeta}_l^k) \right) - \min_{x\in\mathcal{X}} \sum_{l=0}^{L} \mathbb{E}[g(x, \bar{\zeta}_l)] \right\|_2 \\
&\leq \left\| \sup_{x\in\mathcal{X}} \left(\left| \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(x, \bar{\zeta}_l^k) - \sum_{l=0}^{L} \mathbb{E}[g(x, \bar{\zeta}_l)] \right| \right) \right\|_2 \\
&\leq \left\| \sup_{x\in\mathcal{X}} \left(\sum_{l=1}^{L} \left| \frac{1}{N_l} \sum_{k=1}^{N_l} (g(x, \bar{\zeta}_l^k) - \mathbb{E}[g(x, \bar{\zeta}_l)]) \right| \right) \right\|_2 \\
&\leq \left\| \sum_{l=1}^{L} \sup_{x\in\mathcal{X}} \left| \frac{1}{N_l} \sum_{k=1}^{N_l} (g(x, \bar{\zeta}_l^k) - \mathbb{E}[g(x, \bar{\zeta}_l)]) \right| \right\|_2 \\
&\leq \sum_{l=1}^{L} \left\| \sup_{x\in\mathcal{X}} \left| \frac{1}{N_l} \sum_{k=1}^{N_l} (g(x, \bar{\zeta}_l^k) - \mathbb{E}[g(x, \bar{\zeta}_l)]) \right| \right\|_2.
\end{aligned} \tag{4.4.35}$$

In order to study $\left\| \sup_{x\in\mathcal{X}} \left| \frac{1}{N_l} \sum_{k=1}^{N_l} (g(x,\bar{\zeta}_l^k) - \mathbb{E}[g(x,\bar{\zeta}_l)]) \right| \right\|_2$, we extract tools from empirical process theory. For a given $l \geq 0$, if we define,

$$\mathbb{F}_{N_l}^l(\cdot) = \sqrt{N_l} \left(\frac{1}{N_l} \sum_{k=1}^{N_l} (g(\cdot,\bar{\zeta}_k^l) - \mathbb{E}[g(\cdot,\bar{\zeta}^l)]) \right), \tag{4.4.36}$$

then $F_{N_l}^l$ is an empirical process. Further, since $f(\cdot,\zeta_l)$ is Lipschitz continuous, we conclude that is $g(\cdot,\bar{\zeta}^l)$ is also Lipschitz. Thereby applying Lemma 3,4 and 5 and under the assumption 4.4.2, we have

$$\begin{aligned} &\sqrt{N_l} \left\| \sup_{x\in\mathcal{X}} \left| \frac{1}{N_l} \sum_{k=1}^{N_l} (g(x,\bar{\zeta}_l^k) - \mathbb{E}[g(x,\bar{\zeta}_l)]) \right| \right\|_2 \leq \\ &C\|\bar{g}(x,\bar{\zeta}^l)\|_2 \int_0^1 \sqrt{1 + \log(Q_{[]}(v\|\bar{g}(x,\bar{\zeta}^l)\|_2, \mathfrak{F}, \|\cdot\|_2))} dv + \|g(x,\bar{\zeta}_l)\|_2, \end{aligned} \tag{4.4.37}$$

where $\bar{g}(x,\bar{\zeta}^l) = \sup_{x\in\mathcal{X}} |g(x,\bar{\zeta}_l) - \mathbb{E}[g(x,\bar{\zeta}_l)]|$. Referring to the calculations in [56] (Proposition EC.3), we have

$$\begin{aligned} &C\|\bar{g}(x,\bar{\zeta}^l)\|_2 \int_0^1 \sqrt{1 + \log(Q_{[]}(v\|\bar{g}(x,\bar{\zeta}^l)\|_2, \mathfrak{F}, \|\cdot\|_2))} dv \leq \\ &C' \left(\|\bar{g}(x,\bar{\zeta}^l)\|_2 + \sqrt{d \log \left(\max \left\{ 3, \frac{12\mathbb{D}(\mathcal{X})L_f}{\|\bar{g}(x,\bar{\zeta}^l)\|_2} \right\} \right)} \min \left(4\mathbb{D}(\mathcal{X})L_f, \|\bar{g}(x,\bar{\zeta}^l)\|_2 \right) \right). \end{aligned} \tag{4.4.38}$$

Now if, $\frac{12\mathbb{D}(\mathcal{X})L_f}{\|\bar{g}(x,\bar{\zeta}^l)\|_2} \leq 3$,then we have,

$$\begin{aligned} &C' \left(\|\bar{g}(x,\bar{\zeta}^l)\|_2 + \sqrt{d \log \left(\max \left\{ 3, \frac{12\mathbb{D}(\mathcal{X})L_f}{\|\bar{g}(x,\bar{\zeta}^l)\|_2} \right\} \right)} \min \left(4\mathbb{D}(\mathcal{X})L_f, \|\bar{g}(x,\bar{\zeta}^l)\|_2 \right) \right) \\ &\leq C'\|\bar{g}(x,\bar{\zeta}^l)\|_2 (1 + \sqrt{d \log(3)}). \end{aligned} \tag{4.4.39}$$

Otherwise, we have,

$$C'\left(\|\bar{g}(x,\bar{\zeta}^l)\|_2+\sqrt{d\log\left(\max\left\{3,\frac{12\mathbb{D}(\mathcal{X})L_f}{\|\bar{g}(x,\bar{\zeta}^l)\|_2}\right\}\right)}\min\left(4\mathbb{D}(\mathcal{X})L_f,\|\bar{g}(x,\bar{\zeta}^l)\|_2\right)\right)$$
$$\leq C'\|\bar{g}(x,\bar{\zeta}^l)\|_2\left(1+\sqrt{d\log\left(\frac{12\mathbb{D}(\mathcal{X})L_f}{\|\bar{g}(x,\bar{\zeta}^l)\|_2}\right)}\right)$$
$$\leq \mathfrak{c}_{\mathbb{D}(\mathcal{X})L_f}(\|\bar{g}(x,\bar{\zeta}^l)\|_2)^{\frac{1}{1+a}}, \tag{4.4.40}$$

where the last inequality is the consequence of Lemma 7 for $0<a<\infty$ and for some constant $\mathfrak{c}_l$. Observe that as a consequence of assumption 4.4.2, we have,

$$\|\sup_{x\in\mathcal{X}}|g(x,\bar{\zeta}_l)-\mathbb{E}[g(x,\bar{\zeta}_l)]|\|_2 \leq \|\sup_{x\in\mathcal{X}}|f(x,\zeta_l)-f(x,\zeta_{l-1})|\|_2+\|\sup_{x\in\mathcal{X}}|\mathbb{E}[f(x,\zeta_l)-f(x,\zeta_{l-1})]|\|_2$$
$$\leq 2c_2h_l^{\beta/2}. \tag{4.4.41}$$

Further, letting $\bar{\mathfrak{c}}=\mathfrak{c}_{\mathbb{D}(\mathcal{X})L_f}$ and collating everything together we have,

$$\left\|\mathfrak{p}^{*,L}-\hat{\mathfrak{p}}^{*,L}\right\|_2\leq 2c_2\bar{\mathfrak{c}}\sum_{l=1}^{L}\frac{h_l^{\bar{\beta}/2}}{\sqrt{N_l}}, \tag{4.4.42}$$

where $\bar{\beta}=\beta\dfrac{1}{1+a}$. Putting everything together, we get the desired result. □

Based on the above formulation of the RMSE error, the following corollary illustrates the sample complexity associated with the MLMC-SAA estimator.

Corollary 4.4.2. *Suppose assumptions 4.2.1, 4.4.1 and 4.4.2 holds. Then, for any $\epsilon<\frac{1}{e}$ the computational complexity required for $\|\hat{\mathfrak{p}}^{*,L}-\mathfrak{p}^*\|_2\leq\epsilon$ is,*

$$\mathcal{C}^{\mathfrak{p}}_{mlmc}=\begin{cases}\mathcal{O}(\epsilon^{-2}), & \text{for } \bar{\beta}>1.\\ \mathcal{O}\left(\epsilon^{-2}\log(\epsilon^{-1})\right), & \text{for } \bar{\beta}=1.\\ \mathcal{O}\left(\epsilon^{-\left(2+\frac{1-\bar{\beta}}{\alpha}\right)}\right), & \text{for } \bar{\beta}<1.\end{cases} \tag{4.4.43}$$

The proof of the above result follows the line of argument similar to the one observed by in [32, 68] and is therefore skipped.

Remark 4.4.1. *Observe that Theorem 4.4.3 is valid for any $a \in (0, \infty)$. However, taking a large value of a would be detrimental to the overall performance of the MLMC estimator as it would lead to a lower order of variance convergence. For $\beta > 1$, taking $a < \beta - 1$ would render $\bar{\beta} > 1$, thereby leading to optimal sample complexity. However, for $\beta \leq 1$, any value of a would lead to $\bar{\beta} < 1$. In this scenario, one could benefit from taking a very small value of a to retain the original order of variance convergence.*

The analysis in this section illustrates how the Multilevel approximation of expectations can yield enhancements in computational complexity. From Theorem 4.4.2, we observe how the MLMC-SAA has better sampling complexity compared to Monte Carlo SAA as one could achieve the unbiased level of performance for $\beta > 1$. Similar results were also observed via the RMSE analysis of the MLMC-SAA. In both the analysis *i.e.*, uniform convergence and RMSE, we see the effect of variance convergence on the computational cost, *i.e.*, introducing the estimator with high order variance convergence can affect the overall performance of the MLMC-SAA. For instance, the variance analysis carried out in [49] shows that $\beta = 1/2$ and as $\alpha = 1/2$, consequently we observe the computational complexity of $\mathcal{O}\left((\gamma + d)\epsilon^{-3}(\log(\epsilon^{-1}))\right)$, similar to the one achieved by a smooth function.

4.5 Optimality Gap Estimator

Another aspect of SAA that is paramount among practitioners is the Optimal Gap estimator. As the name suggests, the primary aim of this estimator is to assess the quality of a candidate solution of the optimization problem (4.1.1). Mathematically, let $\hat{x}$ be a candidate solution. The quality of this solution is assessed using the optimality gap defined as

$$\mathfrak{G}(\hat{x}) := F(\hat{x}) - \mathfrak{p}^*, \tag{4.5.1}$$

where $\mathfrak{p}^* := \min_{x \in X} F(x)$. The Monte Carlo approximation of $\mathfrak{G}(\hat{x})$ is given as,

$$\hat{\mathfrak{G}}_N(\hat{x}) := \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}, \zeta_k) - \min_{x \in \mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x, \zeta_k), \tag{4.5.2}$$

where $\{\zeta_k\}_{1\leq k\leq N}$ are i.i.d realization of the random variable ζ that is common in both the terms in the above equation. In an unbiased realization of the samples, we have that,

$$\mathbb{E}[\hat{\mathfrak{G}}_N(\hat{x})] \geq \mathfrak{G}(\hat{x}). \tag{4.5.3}$$

Therefore, the underlying mechanism is to statistically estimate the upper bound of $\mathfrak{G}(\hat{x})$ by performing M independent estimation of $\hat{\mathfrak{G}}_N(\hat{x})$ and determining the one-sided confidence interval. This approach is well-documented and is readily used in various practical applications. Interested readers may refer to [46, 54, 66] for a detailed discussion of this procedure. We prove a similar result (see Proposition 4.5.1) in the bias sampling setup for the standard Monte Carlo approximation of the objective function. Moreover to quantify the error of the optimality gap estimator, we undertake an RMSE analysis formulating the RMSE bound as a function of the bias parameter h and the number of samples N. We also extend the analysis to the MLMC estimator of the objective function.

Let $\hat{x}$ be a candidate solution obtained by solving an SAA problem. Further let, $\mathfrak{p}^{*,h} = \min_{x\in X} F^h(x)$ and define,

$$\mathfrak{G}^h(\hat{x}) := F_h(\hat{x}) - \mathfrak{p}^{*,h} \text{ and,} \tag{4.5.4}$$

$$\mathfrak{G}^h_N(\hat{x}) := \frac{1}{N}\sum_{k=1}^{N} f(\hat{x}, \zeta_h^k) - \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x, \zeta_h^k). \tag{4.5.5}$$

The following result provides the upper bound of the optimality gap estimator defined in equation (4.5.1).

Proposition 4.5.1. *Let $\hat{x} \in \mathcal{X}$ be a candidate solution to the optimization problem (4.1.1). Suppose the assumption 4.3.1 holds, then for any $h \in \mathfrak{B}$ we have,*

$$\mathfrak{G}(\hat{x}) \leq 2c_1 h^\alpha + \mathbb{E}[G^h_N(\hat{x})]. \tag{4.5.6}$$

Proof. Observe that,

$$\begin{aligned} F(\hat{x}) - \mathfrak{p}^* &= F(\hat{x}) - \mathbb{E}[f(\hat{x}, \zeta_h)] + \mathbb{E}[f(\hat{x}, \zeta_h)] - \mathfrak{p}^{*,h} + \mathfrak{p}^{*,h} - \mathfrak{p}^* \\ &\leq |F(\hat{x}) - \mathbb{E}[f(\hat{x}, \zeta_h)]| + \mathbb{E}[f(\hat{x}, \zeta_h)] - \mathfrak{p}^{*,h} + |\mathfrak{p}^{*,h} - \mathfrak{p}^*| \\ &\leq 2c_1 h^\alpha + \mathbb{E}[f(\hat{x}, \zeta_h)] - \mathfrak{p}^{*,h} \\ &\leq 2c_1 h^\alpha + \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} f(\hat{x}, \zeta_h^k) - \min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x, \zeta_h^k)\right], \end{aligned} \tag{4.5.7}$$

where the second last inequality is due to the assumption 4.3.1. □

The following result gives an RMSE formulation for the estimator defined above.

Theorem 4.5.1 (Monte Carlo SAA)**.** *Suppose assumptions 4.2.1,4.3.1 and 4.3.2 holds. The for $N \in \mathbb{N}$ and $\hat{x} \in \mathcal{X}$,*

$$\|\mathfrak{G}_N^h(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \leq 2c_1 h^\alpha + \mathfrak{c}_2 \frac{\sigma}{\sqrt{N}}. \tag{4.5.8}$$

Proof. Observe that as a consequence of the triangle inequality we have,

$$\|\mathfrak{G}_N^h(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 = \|\mathfrak{G}_N^h(\hat{x}) - \mathfrak{G}^h(\hat{x}) + \mathfrak{G}^h(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \leq \|\mathfrak{G}_N^h(\hat{x}) - \mathfrak{G}^h(\hat{x})\|_2 + \|\mathfrak{G}^h(\hat{x}) - \mathfrak{G}(\hat{x})\|_2. \tag{4.5.9}$$

The second term in the above inequality is bounded as,

$$\|\mathfrak{G}^h(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \leq \|F_h(\hat{x}) - F(\hat{x})\|_2 + \|\mathfrak{p}^{*,h} - \mathfrak{p}^*\|_2. \tag{4.5.10}$$

Now $\|F_h(\hat{x}) - F(\hat{x})\|_2 \leq \sup_{x\in\mathcal{X}} \|\mathbb{E}\left(f(x, \zeta_h) - f(x, \zeta)\right)\| \leq c_1 h^\alpha$, and

$$\|\mathfrak{p}^{*,h} - \mathfrak{p}^*\|_2 \leq \left(\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\mathbb{E}[f(x, \zeta_h)] - \mathbb{E}[f(x, \zeta)]|^2\right]\right)^{1/2} \leq c_1 h^\alpha, \tag{4.5.11}$$

therefore, $\|\mathfrak{G}^h(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \leq 2c_1 h^\alpha$. As for the first term *i.e.*,$\|\mathfrak{G}_N^h(\hat{x}) - \mathfrak{G}^h(\hat{x})\|_2$ observe that,

$$\|\mathfrak{G}_N^h(\hat{x}) - \mathfrak{G}^h(\hat{x})\|_2 \leq \left\|\frac{1}{N}\sum_{k=1}^{N} f(\hat{x}, \zeta_h^k) - \mathbb{E}[f(\hat{x}, \zeta_h)]\right\|_2 + \left\|\min_{x\in\mathcal{X}} \frac{1}{N}\sum_{k=1}^{N} f(x, \zeta_h^k) - \min_{x\in\mathcal{X}} \mathbb{E}[f(x, \zeta_h)]\right\|_2, \tag{4.5.12}$$

For the first term in the above equation, we have,

$$\begin{aligned}\left\|\frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \mathbb{E}[f(x,\zeta_h)]\right\|_2 &= \left(\mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} f(\hat{x},\zeta_h^k) - \mathbb{E}[f(\hat{x},\zeta_h)\right)^2\right]\right)^{\frac{1}{2}} \\ &= \left(\mathbb{E}\left[\left(\frac{1}{N}\sum_{k=1}^{N} Z_k^h(\hat{x})\right)^2\right]\right)^{\frac{1}{2}} \\ &\leq \frac{\mathfrak{c}_2}{\sqrt{N}}\left(\mathbb{E}[(f(\hat{x},\zeta_h) - \mathbb{E}[f(\hat{x},\zeta_h)])^2]\right)^{1/2} \\ &\leq \frac{\mathfrak{c}_2}{\sqrt{N}}\sigma,\end{aligned} \tag{4.5.13}$$

where the last two inequalities are the consequence of [35, Lemma 2.5] and assumption 4.3.2. Further, from the calculations in Theorem 4.4.3, we have,

$$\left\|\min_{x\in\mathcal{X}}\frac{1}{N}\sum_{k=1}^{N} f(x,\zeta_h^k) - \min_{x\in\mathcal{X}}\mathbb{E}[f(x,\zeta_h)]\right\|_2 \leq \mathfrak{c}_3\frac{\sigma}{\sqrt{N}}. \tag{4.5.14}$$

By collating everything together and reassigning constants, we get the desired result. □

In the multilevel paradigm, the optimal gap estimator is defined as

$$\begin{aligned}\mathfrak{G}^L(\hat{x}) &= \mathbb{E}[f(\hat{x},\zeta^L)] - \min_{x\in\mathcal{X}}\mathbb{E}[f(x,\zeta^L)] \\ &= \sum_{l=1}^{L}\mathbb{E}[f(\hat{x},\zeta^l) - f(\hat{x},\zeta^{l-1})] - \min_{x\in\mathcal{X}}\sum_{l=1}^{L}\mathbb{E}[f(x,\zeta^l) - f(x,\zeta^{l-1})]\end{aligned} \tag{4.5.15}$$

with the Monte Carlo approximation being defined as,

$$\hat{\mathfrak{G}}^L(\hat{x}) = \sum_{l=1}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l}(f(\hat{x},\zeta_l^k) - f(\hat{x},\zeta_{l-1}^k)) - \min_{x\in\mathcal{X}}\left(\sum_{l=1}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l}(f(x,\zeta_l^k) - f(x,\zeta_{l-1}^k))\right) \tag{4.5.16}$$

For the sake of notational convenience let $g(x,\bar{\zeta}_l) := f(x,\zeta_l) - f(x,\zeta_{l-1})$, $\mathfrak{p}^{*,L} := \min_{x\in\mathcal{X}}\sum_{l=0}^{L}\mathbb{E}[g(x,\bar{\zeta}_l)]$ and let $\hat{\mathfrak{p}}^{*,L}$ be its monte carlo approximation. Then,

$$\hat{\mathfrak{G}}^L(\hat{x}) = \sum_{l=1}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l} g(\hat{x},\bar{\zeta}_l^k) - \min_{x\in\mathcal{X}}\sum_{l=1}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l} g(x,\bar{\zeta}_l^k). \tag{4.5.17}$$

The following result gives an RMSE formulation for the estimator defined above.

Theorem 4.5.2 (MLMC-SAA). *Suppose assumptions 4.2.1,4.4.1 and 4.4.2 holds. Then for any* $0 < a < 1$, $L \geq 2$ *and any* $\hat{x} \in \mathcal{X}$,

$$\|\hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \leq 2c_1 h_L^{\alpha} + \mathfrak{c}_3 \sum_{l=1}^{L} \frac{h_l^{\bar{\beta}/2}}{\sqrt{N_l}}, \tag{4.5.18}$$

where $\bar{\beta} = \beta \dfrac{1}{1+a}$.

Proof. To begin with, observe that from triangle inequality, we have,

$$\begin{aligned} \|\hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 &= \|\hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}^L(\hat{x}) + \mathfrak{G}^L(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \\ &\leq \|\hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}^L(\hat{x})\|_2 + \|\mathfrak{G}^L(\hat{x}) - \mathfrak{G}(\hat{x})\|_2. \end{aligned} \tag{4.5.19}$$

As before $\|\mathfrak{G}^L(\hat{x}) - \mathfrak{G}(\hat{x})\|_2 \leq 2c_1 h_L^{\alpha}$, therefore, we intend to study, $\|\hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}^L(\hat{x})\|_2$.

$$\begin{aligned} \left\|\hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}^L(\hat{x})\right\|_2 &= \left\| \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(\hat{x}, \bar{\zeta}_l^k) - \sum_{l=1}^{L} \mathbb{E}[g(\hat{x}, \bar{\zeta}_l)] + \mathfrak{p}^{*,L} - \hat{\mathfrak{p}}^{*,L} \right\|_2 \\ &\leq \left\| \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(\hat{x}, \bar{\zeta}_l^k) - \sum_{l=1}^{L} \mathbb{E}[g(\hat{x}, \bar{\zeta}_l)] \right\|_2 + \left\| \mathfrak{p}^{*,L} - \hat{\mathfrak{p}}^{*,L} \right\|_2. \end{aligned} \tag{4.5.20}$$

Now,

$$\begin{aligned} \left\| \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(\hat{x}, \bar{\zeta}_l^k) - \sum_{l=1}^{L} \mathbb{E}[g(\hat{x}, \bar{\zeta}_l)] \right\|_2 &\leq \left\| \frac{1}{N_0} \sum_{k=1}^{N_0} \left(f(x, \zeta_0^k) - \mathbb{E} f(x, \zeta_0) \right) \right\|_2 \\ &+ \left\| \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} \left(f(x, \zeta_l^k) - f(x, \zeta_{l-1}^k) - \mathbb{E}[f(x, \zeta_l) - f(x, \zeta_{l-1})] \right) \right\|_2. \end{aligned} \tag{4.5.21}$$

Let,

$$\mathcal{Z}_l(x) = \begin{cases} f(x, \zeta_0) - \mathbb{E} f(x, \zeta_0), l = 1 \\ f(x, \zeta_l) - f(x, \zeta_{l-1}) - \mathbb{E}[f(x, \zeta_l) - f(x, \zeta_{l-1})], l = 2, \ldots, L \end{cases}, \tag{4.5.22}$$

then it is easy to $\mathcal{Z}_l(x)$ is a zero mean random variable, therefore by Lemma 2.5 in [35], we have,

$$\left\| \frac{1}{N_0} \sum_{k=1}^{N_0} \left(f(x, \zeta_0^k) - \mathbb{E} f(x, \zeta_0) \right) \right\|_2 \leq \frac{\mathfrak{c}_2}{\sqrt{N_0}} \left(\mathbb{E}[\mathcal{Z}_0(x)]^2 \right)^{1/2}$$

$$\left\| \sum_{l=2}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} \left(f(x, \zeta_l^k) - f(x, \zeta_{l-1}^k) - \mathbb{E}[f(x, \zeta_l) - f(x, \zeta_{l-1})] \right) \right\|_2 \leq \mathfrak{c}_2 \sum_{l=2}^{L} \frac{1}{\sqrt{N_l}} \left(\mathbb{E}[\mathcal{Z}_l(x)]^2 \right)^{1/2}. \tag{4.5.23}$$

Consequently, by assumption 4.4.2, we get,

$$\left\| \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(\hat{x}, \bar{\zeta}_l^k) - \sum_{l=1}^{L} \mathbb{E}[g(\hat{x}, \bar{\zeta}_l)] \right\|_2 \leq \sum_{l=1}^{L} \frac{\mathfrak{c}_2}{\sqrt{N_l}} \left(\mathbb{E}[\mathcal{Z}_l(x)]^2 \right)^{1/2} \leq \sum_{l=1}^{L} \frac{c_2 \mathfrak{c}_2}{\sqrt{N_l}} h_l^{\beta/2}. \tag{4.5.24}$$

As for $\left\| \mathfrak{p}^{*,L} - \hat{\mathfrak{p}}^{*,L} \right\|_2$, we have from calculations in Theorem 4.4.3, that,

$$\left\| \mathfrak{p}^{*,L} - \hat{\mathfrak{p}}^{*,L} \right\|_2 \leq 2 c_2 \bar{\mathfrak{c}} \sum_{l=1}^{L} \frac{h_l^{\bar{\beta}/2}}{\sqrt{N_l}}, \tag{4.5.25}$$

where $\bar{\beta} = \beta \frac{1}{1+a}$. Clearly, for any $a > 0$, $\bar{\beta} < \beta$, therefore we have,

$$\left\| \hat{\mathfrak{G}}^L(\hat{x}) - \mathfrak{G}^L(\hat{x}) \right\|_2 \leq \mathfrak{c}_3 \sum_{l=1}^{L} \frac{h_l^{\bar{\beta}/2}}{\sqrt{N_l}}, \tag{4.5.26}$$

for some constant $\mathfrak{c}_3$. Hence, the result follows. □

4.6 Numerical Illustration

In this section, we undertake numerical experimentation to illustrate the impact of biased approximation of ζ on sample average approximation. To this end, we consider a minimization problem in the coherent risk measure paradigms, *i.e.*, Conditional Value at Risk or CVaR. Recall that,

$$\mathrm{CVaR}_\nu(\zeta) := \min_{x \in [\vartheta^*, \vartheta^{**}]} \left\{ x + \frac{1}{1-\nu} \mathbb{E}[(\zeta - x)_+] \right\}, \tag{4.6.1}$$

where ν is the confidence level. In practical scenarios, we do not know ϑ^* and ϑ^{**} apriori. Therefore, we consider large $\bar{M} \in \mathbb{R}$ and solve the optimization problem,

$$\mathrm{CVaR}_\nu(\zeta) \coloneqq \min_{x\in[-\bar{M},\bar{M}]} \left\{ x + \frac{1}{1-\nu}\mathbb{E}[(\zeta - x)_+] \right\}, \tag{4.6.2}$$

consequently, $\mathcal{X} = [-\bar{M}, \bar{M}]$.

For our first example, we consider a portfolio consisting of a single put option where the asset price is driven by a geometric Brownian Motion(gBm) given as,

$$dX_t = \bar{r}X_t dt + \sigma X_t dW_t, \tag{4.6.3}$$

where $\bar{r}$ and σ denote the risk-free rate of return and the volatility, and W_t denotes the standard Brownian Motion. As for the second example, we look into the scenario simulation paradigm developed by in [41] leading to a nested simulation framework. Before we dwell on the numerical simulation, we make certain observations on the function $f(x,\zeta) := x + \frac{1}{1-\nu}(\zeta - x)_+$, and discuss the underlying algorithm in order to undergo numerical experimentation.

To begin with, let ζ_l and $\zeta_{l'}$ be two approximation of the random variable ζ, then,

$$|f(x,\zeta_l) - f(x,\zeta_{l'})| \leq \frac{1}{1-\nu}|\zeta_l - \zeta_{l'}|, \tag{4.6.4}$$

for all $x \in \mathcal{X}$. Therefore,

$$\sup_{x\in\mathcal{X}}|f(x,\zeta_l) - f(x,\zeta_{l'})| \leq \frac{1}{1-\nu}|\zeta_l - \zeta_{l'}|. \tag{4.6.5}$$

Consequently, we have,

$$\sup_{x\in\mathcal{X}} \mathbb{E}[|f(x,\zeta_l) - f(x,\zeta_{l'})|] \leq \frac{1}{1-\nu}\mathbb{E}|\zeta_l - \zeta_{l'}|, \text{ and} \tag{4.6.6}$$

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}}|f(x,\zeta_l) - f(x,\zeta_{l'})|^2\right] \leq \frac{1}{(1-\nu)^2}\mathbb{E}[|\zeta_l - \zeta_{l'}|^2]. \tag{4.6.7}$$

The above two inequalities help us determine α and β, *i.e.* the bias and variance convergence rate, essential for our multilevel simulation. The next step in our simulation is determining the number of samples to achieve a RMSE of ϵ. In the multilevel paradigm,

the theoretical analysis undertaken in the previous section suggests that taking,

$$N_l = \left\lceil \frac{16}{\epsilon^2}(c_2\bar{\mathfrak{c}})^2 h_l^{\frac{\bar{\beta}+2}{3}} \left(\sum_{l=1}^{L} h_l^{\frac{\bar{\beta}-1}{3}} \right)^2 \right\rceil \text{ and } L = \left\lceil \frac{\log(2c_1\mathbf{h}^\alpha \epsilon^{-1})}{\alpha \log(m)} \right\rceil, \tag{4.6.8}$$

would lead to RMSE of $\mathcal{O}(\epsilon)$. However, the formulation requires us to estimate various constants, which, given one is able to calculate, leads to very conservative numbers for N_l, whereas, in the application, we often do not have the computational budget to perform simulations based on these estimates. Moreover, the MLMC approximation of the objective function may lead to a loss of convexity even if the original function $f(x, \zeta)$ is convex with respect to x. Therefore, solving the optimization problem to the global optimum can be a challenge.

To address these challenges, we begin by estimating a candidate solution $\hat{x}$ by solving the Monte Carlo SAA (4.3.2). Following this, we follow the algorithms described below to assess the quality of the candidate solution via the Optimality Gap estimator.

Algorithm 5: MLMC Optimality Gap Estimation

input: Candidate solution $\hat{x}$, Required accuracy ϵ, initial step size $\mathbf{h}$, refinement factor m, rate of bias convergence α, rate of variance convergence β;

Step 1: Estimate $h_l, L, \{N_l\}_{0 \le l \le L}$ using the formulas in Table 4.1;

Step 2: For $l = 1, 2, \ldots, L$, generate $(\zeta_l^1, \zeta_{l-1}^1), (\zeta_l^2, \zeta_{l-1}^2), \ldots, (\zeta_l^{N_l}, \zeta_{l-1}^{N_l})$ independent and identically distributed samples of (ζ_l, ζ_{l-1});

Step 3: Calculate the Optimality Gap via,

$$\hat{\mathfrak{G}}^L(\hat{x}) = \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(\hat{x}, \bar{\zeta}_l^k) - \min_{x \in \mathcal{X}} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{k=1}^{N_l} g(x, \bar{\zeta}_l^k) \tag{4.6.9}$$

Return: $\hat{\mathfrak{G}}^L(\hat{x})$.

m	$m_l = m^{l-1},\ l = 1, \ldots, L$
$L = L^*(\epsilon)$	$1 + \left\lceil \dfrac{\log\left((1+2\alpha)^{\frac{1}{2\alpha}} \left(\frac{\lvert c_1 \rvert}{\epsilon}\right)^{\frac{1}{\alpha}} \mathbf{h} \right)}{\log(m)} \right\rceil$
h	$\dfrac{\mathbf{h}}{\left\lceil \mathbf{h}\,(1+2\alpha)^{\frac{1}{2\alpha}} \left(\frac{\lvert c_1 \rvert}{\epsilon}\right)^{\frac{1}{\alpha}} m^{-L} \right\rceil}$
$\mu = \mu^*(\epsilon)$	$\mu_1(\epsilon) = \dfrac{1}{\mu_\epsilon^\dagger}, \quad \mu_l(\epsilon) = \dfrac{\lambda h^{\frac{\beta}{2}} \left(m_{l-1}^{-1} - m_l^{-1}\right)^{\frac{\beta}{2}}}{\mu_\epsilon^\dagger \sqrt{m_{l-1} + m_l}}, \quad l = 2, \ldots, L,$ with $\mu_\epsilon^\dagger$ s.t. $\sum_{l=1}^{L} \mu_l(\epsilon) = 1$ and $\lambda = \dfrac{V_1(\hat{x})}{V_h(\hat{x})}$.
$N = N^*(\epsilon)$	$\left\lceil \dfrac{\left(1 + \frac{1}{2\alpha}\right) V_h(\hat{x}) q_\epsilon^\dagger \left(1 + \lambda h^{\frac{\beta}{2}} \sum_{l=2}^{L} \left(m_{l-1}^{-1} - m_l^{-1}\right)^{\frac{\beta}{2}} \sqrt{m_{l-1} + m_l}\right)}{\epsilon^2} \right\rceil$
$N_l = N_l^*(\epsilon)$	$\lceil N^*(\epsilon)\mu_l(\epsilon) \rceil$

Table 4.1: Parameters for MLMC-SAA [68]

The reader may refer to the [68, section 9.5.2] for a discussion on the calibration of the parameters $V_1(\hat{x})$,$V_h(\hat{x})$ and c_1 for the candidate solution $\hat{x}$. Below is the algorithm for the Monte SAA optimality gap estimation.

Algorithm 6: Monte Carlo SAA

Input: Candidate solution $\hat{x}$, variance $\sigma_{\hat{x}}^2$, required accuracy ϵ, rate of bias convergence α;

Step 1: $\mathbf{h} = \mathcal{O}(\epsilon^{\frac{1}{\alpha}})$;

Step 2: Estimate $N = \left\lceil \left(1 + \dfrac{1}{2\alpha}\right) \dfrac{\sigma_{\hat{x}}^2}{\epsilon^2} \right\rceil$.;

Step 3: Generate $\zeta_{\mathbf{h}_0}^1, \zeta_{\mathbf{h}}^2, \ldots, \zeta_{\mathbf{h}}^N$ independent and identically distributed sample of the random variable $\zeta_{\mathbf{h}}$;

Step 7: Calculate the Optimality Gap via,

$$\mathfrak{G}_N^{\mathbf{h}}(\hat{x}) := \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}, \zeta_{\mathbf{h}}^k) - \min_{x \in \mathcal{X}} \frac{1}{N} \sum_{k=1}^{N} f(x, \zeta_{\mathbf{h}}^k) \tag{4.6.10}$$

Return: $\mathfrak{G}_N^{\mathbf{h}}(\hat{x})$.

Finally, to assess the quality of our estimator, we estimate the upper bound of the RMSE error of the optimality gap estimator.

4.6.1 Geometric Brownian Motion

For our first example, we shall consider an investment consisting of a short position in a single put option, where the loss is defined as

$$\zeta := (K - X_T)_+ - e^{\bar{r}T} P_0, \tag{4.6.11}$$

with P_0 being the initial price at which the option was sold. We assume the underlying stock X_t follows gBm, *i.e.*,

$$dX_t = \bar{r} X_t dt + \sigma X_t dW_t, \tag{4.6.12}$$

and further cosider $X_0 = 100$, $\bar{r} = 0.05$, $\sigma = 0.2$, $T = 1$, $P_0 = 10.7$ and (strike price) $K = 110$ for our simulation [9]. Moreover, we assume $\nu = 0.95$. In order to undergo our simulation, we discretize the gBm using Euler-Maruyama and Milstein numerical scheme, given as,

$$X_{n+1} = X_n + \bar{r} X_n h + \sigma X_n \Delta W_n \qquad \text{(Euler-Marumaya)} \tag{4.6.13}$$

$$X_{n+1} = X_n + (\bar{r} - \frac{1}{2}\sigma^2) h X_n + \sigma X_n \Delta W_n + \frac{1}{2}\sigma^2 X_n (\Delta W_n)^2 \qquad \text{(Milstein)}, \tag{4.6.14}$$

where $h = \mathbf{h}/m$ is the step size and $\Delta W_n = W_{n+1} - W_n$. Here, we take $m = 4$ as the refinement factor, $\mathbf{h} = T$ for the Milstein scheme, and $\mathbf{h} = T/8$ for the Euler-Maruyama scheme. The value of α and β can be estimated based on equation (27) and (28). Referring to the analysis in [34], we have, $\alpha = 1$ and $\beta = 1$ for Euler-Maruyama scheme and $\alpha = 1$ and $\beta = 2$ for Milstein Scheme. Also, we take $a = 10^{-3}$. Based on these parameters, we undertake our simulation where we perform a 50 independent run of algorithms 6 and 5 to determine the quality of the candidate solution $\hat{x}$. In order to solve the optimization problem in the optimality gap estimation, we use the SciPy optimization package. Tables 4.2 and 4.3 tabulate the results of our experiments. In Figures 4.2 and 4.3, we provide the graphical representation of our results.

		Monte Carlo SAA					MLMC SAA			
ϵ		**h**	RMSE	Cost	$\mathfrak{G}_N^{\mathbf{h}}(\hat{x})$		**h**	RMSE	Cost	$\hat{\mathfrak{G}}^L(\hat{x})$
0.6000		0.15000	1.5388	3.7606e+04	0.088		0.1250	1.2582	6.7592e+04	0.228
0.4000		0.10000	1.1280	1.5303e+05	0.233		0.1250	0.6896	2.2835e+05	0.187
0.2000		0.05000	0.6080	1.1732e+06	0.198		0.1250	0.4128	1.5350e+06	0.171
0.1000		0.02500	0.2823	9.2550e+06	0.182		0.1250	0.2317	5.7067e+06	0.174
0.0800		0.02000	0.2381	1.6829e+07	0.182		0.1250	0.1742	1.0364e+07	0.175

Table 4.2: Euler-Maruyama Approximation: RMSE analysis of the Optimality Gap estimator for a candidate solution $\hat{x} = 23.0710$.

		Monte Carlo SAA					MLMC SAA			
ϵ		**h**	RMSE	Cost	$\mathfrak{G}_N^{\mathbf{h}}(\hat{x})$		**h**	RMSE	Cost	$\hat{\mathfrak{G}}^L(\hat{x})$
0.6000		0.15000	2.8130	2.8579e+04	0.036		1.0000	1.1455	1.8719e+04	0.182
0.4000		0.10000	0.8239	1.3828e+05	0.149		1.0000	0.7716	3.3200e+04	0.179
0.2000		0.05000	0.4046	1.0657e+06	0.156		1.0000	0.3555	2.7956e+05	0.159
0.1000		0.02500	0.2029	9.3160e+06	0.162		1.0000	0.1929	1.0914e+06	0.167
0.0800		0.02000	0.1641	1.8268e+07	0.164		1.0000	0.1716	1.4077e+06	0.164

Table 4.3: Milstein Approximation: RMSE analysis of the Optimality Gap estimator for a candidate solution $\hat{x} = 23.0710$

4.6.2 Nested Simulation

For the second example, we refer to the research carried out in [35, 41], where the authors formulated the estimation of CVaR as a nested expectation problem. Consequently, for our simulation, we define ζ as follows,

$$\zeta := -1 - \mathbb{E}[\phi(Y, Z)|Y], \tag{4.6.15}$$

where, $\phi(y, z) := -\tau y^2 - 2\sqrt{\tau(1-\tau)}yz - (1-\tau)z^2$ and $y, z \in \mathbb{R}$. Also, Y, Z are independent following normal distribution $\mathcal{N}(0, 1)$. The above formulation considers an option with payoff $-W_T^2$ at time $T = 1$. The value of the option at a time t is given by $P(t, y) := \mathbb{E}[-W_T^2|W_t = y]$ and the loss ζ is given as, $\zeta := P(0, 0) - P(\tau, W_\tau)$, where

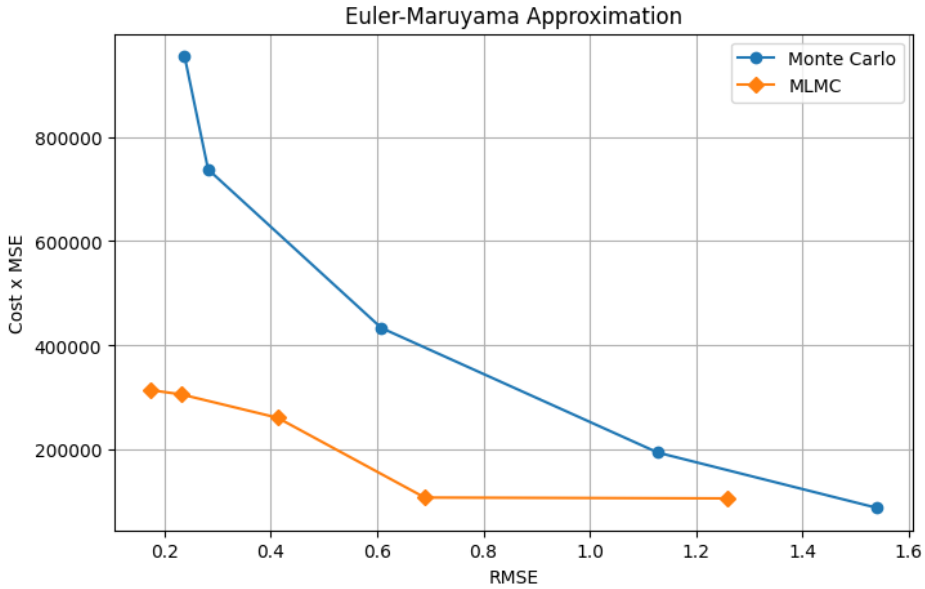


Figure 4.2: Euler-Maruyama Approximation- Comparison of computational efficiency between Monte Carlo-SAA and MLMC-SAA methods in a gBm setting.

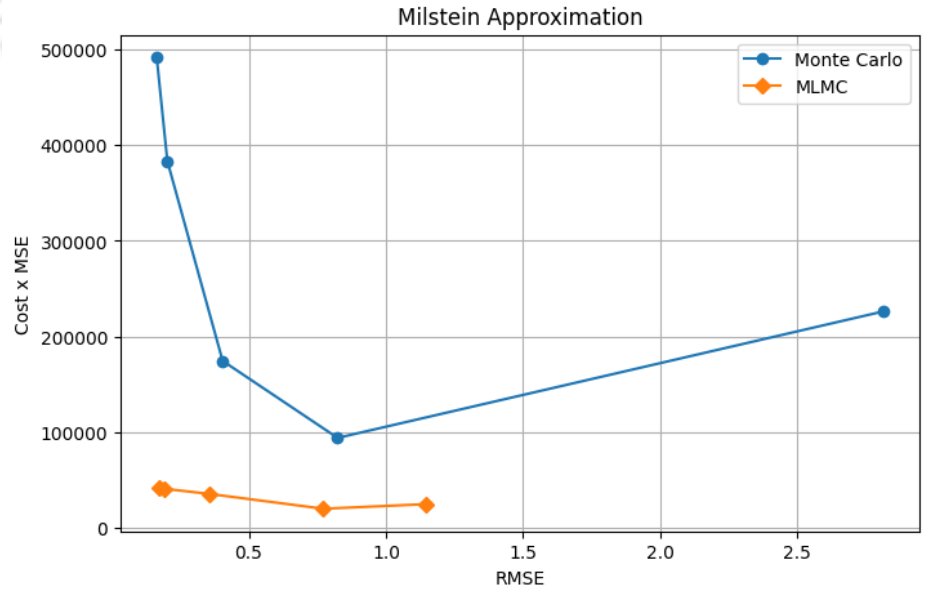


Figure 4.3: Milstein Approximation- Comparison of computational efficiency between Monte Carlo-SAA and MLMC-SAA methods in a gBm setting.

$\tau \in (0,1)$ is the time horizon. For further information on the above formulation and the analytical calculations, refer to [22, 35]. The Monte Carlo approximation of $F(x)$ is based on generating inner and outer samples. Let the bias parameter $h = 1/M$ where M is the number of inner samples and let N denote the number of outer samples; then,

$$F_N^h(x) = \frac{1}{N}\sum_{k=1}^{N}\left(x + \frac{1}{1-\nu}\left(-1 - \frac{1}{M}\sum_{j=1}^{M}\phi(Z_j, Y_k) - x\right)_+\right) \tag{4.6.16}$$

gives the Monte Carlo approximation of $F(x)$. We take $\tau = 0.5$ and $\nu = 0.975$ for our numerical simulation. For the multilevel approximation, let,

$$\hat{E}_{M_l}(Y_k) = -1 - \frac{1}{M_l}\sum_{j=1}^{M_l}\phi(Z_j, Y_k) \tag{4.6.17}$$

and further define,

$$f(x, \hat{E}_{M_l}(Y_k)) := x + \frac{1}{1-\nu}\left(\hat{E}_{M_l}(Y_k) - x\right)_+. \tag{4.6.18}$$

Then $\hat{F}_L(x)$, is given as,

$$\hat{F}_L(x) := \sum_{l=1}^{L}\frac{1}{N_l}\sum_{k=1}^{N_l}\left(f(x, \hat{E}_{M_l}(Y_k)) - f(x, \hat{E}_{M_{l-1}}(Y_k))\right). \tag{4.6.19}$$

To this end, observe that for a given Y, we have,

$$\sup_{x\in\mathbb{R}}\left|f(x, \hat{E}_{M_l}(Y)) - f(x, \hat{E}_{M_{l-1}}(Y_k))\right| \leq \frac{1}{(1-\nu)^2}\left|\hat{E}_{M_l}(Y) - \hat{E}_{M_{l-1}}(Y)\right| \tag{4.6.20}$$

where the right-hand side of the above inequality is independent of x. Therefore, we have

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|f(x, \hat{E}_{M_l}(Y)) - f(x, \hat{E}_{M_{l-1}}(Y_k))\right|^2\right] \leq \frac{1}{(1-\nu)^2}\mathbb{E}\left[\left|\hat{E}_{M_l}(Y) - \hat{E}_{M_{l-1}}(Y)\right|^2\right]. \tag{4.6.21}$$

Now, as a consequence of [68, Proposition 9.2 (a)] and equation (4.6.20) and (4.6.21), we have $\beta = 1$ and $\alpha = 1$ albeit under some regularity assumptions. For multilevel

simulation, we take $\mathbf{h} = 1/64$, *i.e.*, $M = 64$ and take $a = 10^{-3}$. As before, we perform 50 independent simulation and estimate the upper bound of RMSE of the optimality gap estimator. Table 4.4 tabulates the results obtained through our experimentation. Finally, in Figure 4.4, we provide a graphical representation of the computational savings achieved by MLMC-SAA in the nested simulation framework.

		Monte Carlo SAA					MLMC SAA			
ϵ		$\mathbf{h}$	RMSE	Cost	$\mathfrak{G}_N^{\mathbf{h}}(\hat{x})$		$\mathbf{h}$	RMSE	Cost	$\hat{\mathfrak{G}}^L(\hat{x})$
0.5000		0.01562	1.2019	1.9598e+04	0.121		0.01562	1.0927	4.1728e+04	0.120
0.2500		0.00781	0.6407	1.5436e+05	0.050		0.01562	0.7772	1.2877e+05	0.069
0.1250		0.00391	0.3261	1.2210e+06	0.041		0.01562	0.3399	4.9286e+05	0.048
0.0625		0.00195	0.1510	9.2048e+06	0.034		0.01562	0.2471	1.2500e+06	0.037
0.03125		0.00098	0.0784	7.3494e+07	0.035		0.01562	0.1252	4.7751e+06	0.033

Table 4.4: Nested Simulation: RMSE analysis of the Optimality Gap estimator for a candidate solution $\hat{x} = 2.2754$.

From the results presented in this section, it is evident that MLMC approximation of the optimality gap estimator is significantly more efficient than standard Monte Carlo, especially when high accuracy is required. It offers substantial cost savings while maintaining comparable or better accuracy. However, we would like to caution against the use of MLMC for solving convex stochastic programs with biased sampling, especially while using the software package such as CVXPY, due to possible loss of convexity of the objective function with multilevel approximation. In such a scenario, one can, however, approximate the candidate solution via a multilevel estimator and approximate the optimal value by the Monte Carlo SAA using the samples generated in the finest level of the multilevel approximation. Such a method would provide an upper bound for the optimality gap estimator without any additional computational overload.

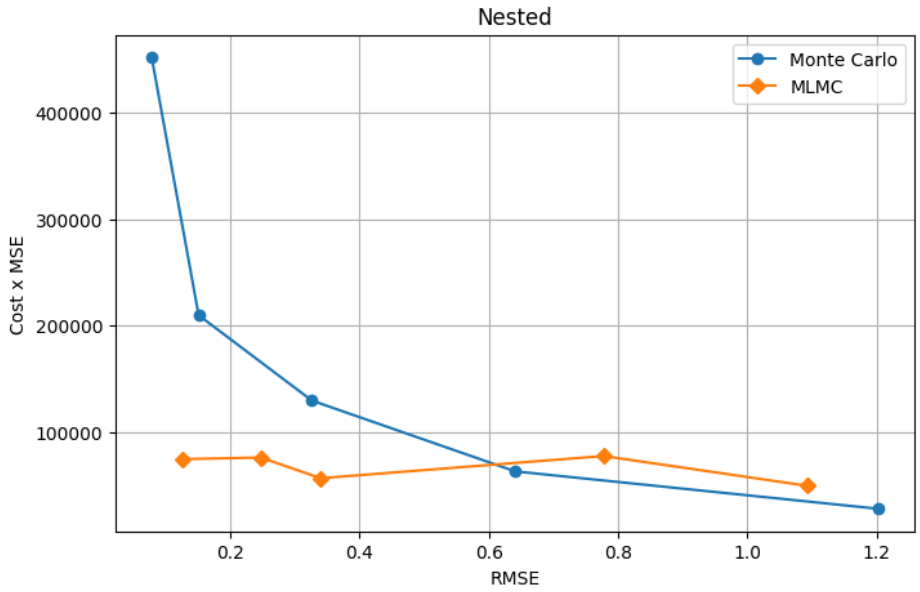


Figure 4.4: Comparison of computational efficiency between Monte Carlo-SAA and MLMC-SAA methods in a nested simulation setting.

4.7 Summary

In this chapter, we looked into the SAA procedure for solving the stochastic optimization problem, where the random variable ζ is sampled from an approximate distribution, introducing bias in the Monte Carlo estimation of the expectation. We extended the conventional SAA setup to the multilevel framework and derived the sample complexity associated with performing the optimization procedure. Following the traditional analysis, we undertook the uniform convergence analysis, establishing the rate of convergence and the sampling complexity results both in the standard and multilevel context. We further analyzed RMSE and derived the sample complexity results to achieve $\epsilon-$RMSE. Finally, we demonstrated the benefits of incorporating MLMC via a series of numerical examples.

Chapter 5

Non-Asymptotic Estimation of Conditional Value-at-Risk via Stochastic Variance Reduced Gradient Langevin Dynamics

5.1 Introduction

In Chapter 4, we dealt with numerical examples pertaining to CVaR estimation for a fixed portfolio where we assumed losses are sampled from a biased distribution. In this chapter, we study an important aspect of risk mitigation that deals with the efficient computation of $\mathrm{CVaR}_\nu(\Phi)$. If the distribution of Φ is tractable, one may be able to analytically formulate the $\mathrm{CVaR}_\nu(\Phi)$. However, as with any practical problems, we often have limited information about Φ, if any. In such scenarios, a conventional way is to use numerical algorithms to compute risk measures efficiently. As before, for efficient numerical computation, one often resorts to the Rockafellar-Uryasev [74] representation, given as,

$$\mathrm{CVaR}_\nu(\Phi) := \inf_{\vartheta \in \mathbb{R}} \left\{ \vartheta + \frac{1}{1-\nu} \mathbb{E}[(\Phi - \vartheta)_+] \right\}, \tag{5.1.1}$$

where $(x)_+ = \max\{x, 0\}$.

With the above formulation, the computation of $\mathrm{CVaR}_\nu(\Phi)$ boils down to solving a stochastic optimization problem. In this regard, there exists a vast literature discussing the numerical aspect of $\mathrm{CVaR}_\nu(\Phi)$ estimation. For instance, one may refer to [9, 13, 21,

30, 51, 68, 81, 76] for estimation of VaR and CVaR via stochastic optimization problem. Additionally, [20, 35, 88, 48] offers insights into the Monte Carlo methods for estimating CVaR.

For our purpose, we consider the regularized version of the Rockafellar-Uryasev expression, given as,

$$\mathrm{CVaR}_\nu(\Phi) \coloneqq \inf_{\vartheta\in\mathbb{R}} \left\{ \vartheta + \frac{1}{1-\nu}\mathbb{E}[(\Phi-\vartheta)_+] + \frac{\eta}{2}|\vartheta|^2 \right\}. \tag{5.1.2}$$

Moreover, the portfolio's total loss, Φ, is influenced both by the loss incurred by each individual financial instrument and the weight assigned to that instrument. Suppose our portfolio comprises $d-1$ financial instruments $X = \{X^1, \ldots, X^{d-1}\}$, where X^i is the $\mathbb{R}$ valued random variable that represents the loss from the i^{th} instrument. Accordingly, we express the overall portfolio loss as,

$$\Phi(w, X) = \sum_{i=1}^{d-1} g_i(w) X^i, \tag{5.1.3}$$

where $w = \{w_1, w_2, \ldots, w_{d-1}\}$ being a vector of parameters and $g_i : \mathbb{R}^{d-1} \to \mathbb{R}$ a function that determines the weight attributed to X_i. Consequently, we can extend the problem of estimation of CVaR for a given loss of Φ to the problem of determining the minimum-CVaR portfolio. To this end, we let $\theta = \{w, \vartheta\}$ be a $d-$dimensional vector. Then,

$$\begin{aligned} \widetilde{\mathrm{CVaR}}_\nu(\Phi) &\coloneqq \inf_{w\in\mathbb{R}^{d-1}} \inf_{\vartheta\in\mathbb{R}} \left\{ \vartheta + \frac{1}{1-\nu}\mathbb{E}[(\Phi(w,X)-\vartheta)_+ + \frac{\eta}{2}|\theta|^2] \right\}, \\ &= \inf_{\theta\in\mathbb{R}^d} \left\{ \vartheta + \frac{1}{1-\nu}\mathbb{E}[(\Phi(w,X)-\vartheta)_+] + \frac{\eta}{2}|\theta|^2 \right\}. \end{aligned} \tag{5.1.4}$$

In order to facilitate easier navigation of the rest of the chapter, we let,

$$f(\theta, X) = \vartheta + \frac{1}{1-\nu}(\Phi(w,X)-\vartheta)_+ + \frac{\eta}{2}|\theta|^2, \tag{5.1.5}$$

and define,

$$F(\theta) \coloneqq \mathbb{E}[f(\theta, X)] = \int_{\mathbb{R}^q} f(\theta, x)\mathbb{P}(dx). \tag{5.1.6}$$

Therefore,

$$\widetilde{\mathrm{CVaR}}_\nu(L) = \min_{\theta\in\mathbb{R}^d} F(\theta). \tag{5.1.7}$$

As stated previously, we often don't have knowledge about the distribution of X. Therefore, a standard approach is to minimize the finite sum approximation of the original minimization problem. To this end, we assume access to $\mathbf{x} = \{X_1, X_2, \ldots, X_N\} \in (\mathbb{R}^q)^N$ and define the finite sum approximation as,

$$F_{\mathbf{x}}(\theta) = \frac{1}{N} \sum_{k=1}^{N} f(\theta, X_k). \tag{5.1.8}$$

Accordingly, our problem becomes,

$$\widetilde{\mathrm{CVar}}_{\nu}^{\mathbf{x}}(L) = \min_{\theta \in \mathbb{R}^d} F_{\mathbf{x}}(\theta). \tag{5.1.9}$$

One of the approaches to solving the stochastic optimization problem defined above is via Stochastic Gradient Langevin Dynamics (SGLD), which derives its foundation from the Langevin SDE defined on $t \in [0, \infty)$, starting with the random variable $\theta_0 \in \mathbb{R}^d$ and is given by,

$$d\theta_t = -\nabla F_{\mathbf{x}}(\theta_t)dt + \sqrt{2\delta^{-1}}dW_t, \tag{5.1.10}$$

where $\delta > 0$ is a constant and $(W_t)_{t \geq 0}$ is a $d-$dimensional Brownian motion. SGLD is essentially the Euler-Maruyama discretization of the SDE (5.1.10) given as,

$$\theta_{n+1} = \theta_n - hv_n(\theta) + \sqrt{2\delta^{-1}}\Delta W_n, \tag{5.1.11}$$

where h is the stepsize and v_n is the conditional unbiased estimator of $\nabla F_{\mathbf{x}}(\theta)$. We observe that a natural candidate for v_n is $\nabla_\theta f(\theta_n, X_{i_{n+1}})$ where $i_{n+1} \sim \mathrm{Unif}\{1, \ldots, N\}$.

The primary interest in Langevin Monte Carlo methods stems from the fact that under favourable conditions, the SDE (5.1.10) leads to a unique invariant measure $\pi_\delta^{\mathbf{x}} \propto e^{-\delta F_{\mathbf{x}}(\theta)}$ which has been proven to concentrate around the minimizer of $F_{\mathbf{x}}(\theta)$, for large enough δ, see, e.g., [50]. Moreover, the recent developments undertaken in [53, 71, 84], where the central motive was to solve the optimization problems that appear in the machine learning paradigms, have prompted researchers to extend this technique to more general stochastic optimization problems. In the context of minimum-CVaR portfolio optimization problems, [21, 76] provided non-asymptotic error bounds for estimation of $\mathrm{CVaR}_\nu(L)$. Additionally, various variants of SGLD, which have been explored in [62, 63, 64], are also applicable for solving minimum-CVaR portfolio optimization problems.

In the context of SGLD, the implementation of the algorithm faces two major chal-

lenges. The first challenge is more general as it deals with the large number of replications required by the algorithm to generate high-precision samples from the target distribution. This issue arises due to the large variance of the stochastic gradient estimator. On the other hand, the second challenge is more specific to the problem under consideration as it stems from the discontinuity of the $\nabla_\theta f(\theta, x) = [\partial_{w_1} f(\theta, x), \ldots, \partial_{w_{d-1}} f(\theta, x), \partial_\vartheta f(\theta, x)]^T$, where,

$$\partial_{w_i} f(\theta, x) = \frac{\partial \Phi(w, x)}{\partial w_j} \mathbf{1}_{\{\Phi(w,x) \geq \vartheta\}} + \eta w_j \text{ and } \partial_\vartheta f(\theta, x) = 1 - \mathbf{1}_{\{\Phi(w,x) \geq \vartheta\}} + \eta\vartheta. \quad (5.1.12)$$

Accordingly, in this chapter, we consider a variant of SGLD, also known as Stochastic Variance Reduced Gradient Langevin Dynamic (SVRG-LD), to address the issue of large variance of the stochastic gradient estimator. In order to tackle the issue of discontinuous gradient, we consider a smooth approximation $F_{\mathbf{x}}^{\varepsilon}(\theta)$ of $F_{\mathbf{x}}(\theta)$ and solve the corresponding optimization problem given as,

$$\widetilde{\mathrm{CVar}}_{\nu}^{\mathbf{x},\varepsilon}(\Phi) = \min_{\theta \in \mathbb{R}^d} F_{\mathbf{x}}^{\varepsilon}(\theta). \quad (5.1.13)$$

In this setup, we undertake a non-asymptotic convergence analysis and provide the error bound for the expected excess risk, which is given as,

$$\mathrm{ER} := \mathbb{E}[F(\theta_n)] - \mathrm{CVaR}_{\nu}(\Phi), \quad (5.1.14)$$

where θ_n is the $n-$th iterate of the SVRG-LD procedure for sampling from $\pi_{\delta}^{\mathbf{x},\varepsilon} \propto e^{-\delta F_{\mathbf{x}}^{\varepsilon}(\theta)}$.

In the machine learning paradigm, the first issue has been explored in the SVRG-LD framework. For example, in [17, 25], authors introduced two types of variance-reduced stochastic gradient algorithms, namely SVRG-LD and SAGA-LD and undertook mean-squared error analysis of the sample path average. In [15], authors provide the convergence results in the 2-Wasserstein distance for sampling from log-concave distribution. These results were further improved in [89]. Additionally, [91] provides convergence results in the 2-Wasserstein distance for non-convex optimization. In the same paradigm, [53] undertook the non-asymptotic analysis of the SVRG-LD in KL-divergence. In our setup, however, we consider a variant of standard SVRG-LD which we describe in Subsection 5.2.2. Also, in our analysis, in addition to estimating the non-asymptotic bound of the 2-Wasserstein distance, we also provide a non-asymptotic bound for the 1-Wasserstein distance in the SVRG-LD framework. Moreover, the challenge of the discontinuous gra-

dient has been addressed in literature by assuming Lipschitz continuity in expectation, see e.g. [16, 63, 64, 65, 76]. These studies have undertaken rigorous mathematical analysis providing non-asymptotic error bounds in 1-Wasserstein and 2-Wasserstein distances. However, the underlying assumption in these setups was the continuity of $\mathcal{L}(X)$. We do not make any such assumption in our framework.

5.2 Preliminaries

In this section, we describe the underlying setup for our analysis and also present the SVRG-LD algorithm that we intend to employ to solve the optimization problem described in (5.1.13).

5.2.1 Setup

Let the function $f : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$ be the Borel measurable function which is non-convex with respect to θ. Further, to address this issue of discontinuous gradient, discussed in Section 5.1, we assume the existence of a smooth approximation $f^\varepsilon(\theta, X)$ of $f(\theta, X)$, such that,

Assumption 5.2.1. *For all $x \in \mathbb{R}^q$,*

$$\sup_{\theta \in \mathbb{R}^d} |f^\varepsilon(\theta, x) - f(\theta, x)| \leq \varepsilon. \tag{5.2.1}$$

Following the setup described in [71], we assume the existence of a collection of probability law $\{\mu_{\mathbf{x}}\}_{\mathbf{x} \in (\mathbb{R}^q)^N}$ on some space $\mathfrak{X}$ such that,

$$\mathbb{E}[\nabla f^\varepsilon(\theta, Y^{\mathbf{x}})] = \nabla F^\varepsilon_{\mathbf{x}}, (\theta) \tag{5.2.2}$$

where $Y^{\mathbf{x}}$ is a random element from $\mathfrak{X}$ with probability law $\mu_{\mathbf{x}}$. In addition, for any $\delta > 0$, we define,

$$\pi^{\mathbf{x},\varepsilon}_\delta(A) := \frac{\int_A e^{-\delta F^\varepsilon_{\mathbf{x}}(\theta)} d\theta}{\int_{\mathbb{R}^d} e^{-\delta F^\varepsilon_{\mathbf{x}}(\theta)} d\theta}, \tag{5.2.3}$$

where $\int_{\mathbb{R}^d} e^{-\delta F^\varepsilon_{\mathbf{x}}(\theta)} d\theta < \infty$. We consider $(\mathcal{G}_n)_{n \in \mathbb{N}}$ to represent the filtration corresponding to the historical progression of information and denote by $\mathcal{G}_\infty := \sigma\left(\bigcup_{n \in \mathbb{N}} \mathcal{G}_n\right)$. We pos-

tulate that $(Y_n^{\mathbf{x}})_{n\in\mathbb{N}}$ is an $\mathcal{G}_n-$adapted process, such that $(Y_n^{\mathbf{x}})_{n\in\mathbb{N}}$ constitutes a sequence of independent and identically distributed random variables, governed by the probability law $\mu_{\mathbf{x}}$. Furthermore, it is presumed throughout this study that the random variable θ_0 (initial condition), $\mathcal{G}_\infty$, and the Brownian increment $(\Delta W_n)_{n\in\mathbb{N}}$ are independent of one another.

5.2.2 Algorithm

As stated in Section 5.1, in this chapter, we intend to implement the SVRG-LD method to solve the optimization problem described in (5.1.13). Detailed in Algorithm 7, the SVRG-LD method deals with computation of the full gradient *i.e.*, $\nabla F_{\mathbf{x}}^{\varepsilon}(\theta_s^v)$ after some fixed number of iterations and use this estimate as the control variate in the gradient estimation in the future iterations. To be more precise, after every $m-$iterations, the full gradient is computed via scanning through the whole data set, *i.e.*,

$$\widetilde{G} = \nabla F_{\mathbf{x}}^{\varepsilon}(\theta_s^v) = \frac{1}{N}\sum_{k=1}^{N} \nabla f^{\varepsilon}(\theta_s, X_k). \tag{5.2.4}$$

Consequently, the SVRG-LD scheme for the next $m-$iterations is given as,

$$\theta_{n+1}^v = \theta_n^v - hv_n + \sqrt{2\delta^{-1}}\Delta W_{n+1} \tag{5.2.5}$$

where,

$$v_n = \nabla f^{\varepsilon}(\theta_n^v, Y_{n+1}^{\mathbf{x}}) - \nabla f^{\varepsilon}(\theta_s^v, Y_{n+1}^{\mathbf{x}}) + \widetilde{G}. \tag{5.2.6}$$

It is easy to observe that $\mathbb{E}[v_n] = \nabla F_{\mathbf{x}}^{\varepsilon}(\theta_n)$, making it an unbiased estimator of $\nabla F_{\mathbf{x}}^{\varepsilon}(\theta_n)$. In contrast to the investigation presented in [25, 53, 89, 91], in which the interval of iterations between consecutive full gradient estimates is fixed, we adopt a strategy of doubling the number of subsequent iterations necessary before computing the next full gradient estimate (see Algorithm 7). This approach thus further contributes to reducing the computational expense associated with full gradient estimation while concurrently preserving the variance reduction characteristic inherent to the underlying method. We conclude this section by introducing certain terminologies that would help navigate the rest of the article.

In this article, we shall denote m as the epoch length, which doubles after every full gradient computation. The iterations at which the complete gradient is evaluated will be

referred to as the snapshot points. Consequently, let $\mathcal{S}^{\nabla} = \{0, s_1, s_2, \cdots : s_1 < s_2 < \dots\}$ represent the set comprising all points where the full gradient is computed. In the context of Algorithm 7, $s_j = 2^{j-1}m_0$ where m_0 is the initial epoch length. Additionally, we shall refer to $\widetilde{G}$ as the snapshot gradient and θ_s^v as the snapshot iterate. Finally, we observe that, for given $n \in \mathbb{N}$, if we let $i_n = \max\{j \in \mathbb{N} : n \geq s_j, s_j \in \mathcal{S}^{\nabla}\}$, then there exists an $\ell_n \in \{0, 1, 2, \dots, 2^{i_n-1}m_0 - 1\}$ such that $n = 2^{i_n-1}m_0 + \ell_n$.

Algorithm 7: Stochastic Variance Reduced Langevin Dynamics

input: step size $h > 0$, total iterations K, initial epoch length m_0, inverse temperature parameter $\delta > 0$, dataset $\mathbf{x} = \{X_1, \dots, X_N\}$;
initialization: $\theta_0 = \theta_0, m = m_0$;
for $n = 0, \dots, K-1$ **do**
  if $n = 0$ **then**
    $\widetilde{G} = \nabla F_{\mathbf{x}}^{\varepsilon}(\theta_0)$;
    $\theta_s^v = \theta_0$;
  end
  else if $n \mod m = 0$ **then**
    $\widetilde{G} = \nabla F_{\mathbf{x}}^{\varepsilon}(\theta_n^v)$;
    $\theta_s^v = \theta_n^v$;
    $m = 2 \times m$;
  end
  sample $Y_{n+1}^{\mathbf{x}} \sim \mu_{\mathbf{x}}$;
  sample $\Delta W_{n+1} \sim \mathcal{N}(0, \sqrt{h})$;
  $v_n = \nabla f^{\varepsilon}(\theta_n^v, Y_{n+1}^{\mathbf{x}}) - \nabla f^{\varepsilon}(\theta_s^v, Y_{n+1}^{\mathbf{x}}) + \widetilde{G}$;
  $\theta_{n+1}^v = \theta_n^v - hv_n + \sqrt{2\delta^{-1}}\Delta W_{n+1}$;
end
return θ_{K-1}^v

5.3 Main Result

In this section, we provide a non-asymptotic error estimate for the SVRG-LD algorithm that characterizes the convergence rate of SVRG-LD and also provides the non-asymptotic bound on the Excess Risk (5.1.14). We first impose the following assumptions that are necessary for the theoretical analysis.

Assumption 5.3.1. *The initial condition θ_0 has a finite second and fourth moment i.e., $\kappa_{0,2} \coloneqq \mathbb{E}[|\theta_0|^2] < \infty$ and $\kappa_{0,4} \coloneqq \mathbb{E}[|\theta_0|^4] < \infty$.*

Assumption 5.3.2. *For each $x \in \mathbb{R}^{d-1}$ we assume,*

$$|\nabla f^{\varepsilon}(\theta_1, x) - \nabla f^{\varepsilon}(\theta_2, x)| \leq L|\theta_1 - \theta_2|, \tag{5.3.1}$$

for all $\theta_1, \theta_2 \in \mathbb{R}^d$, that is f^{ε} is $L-$smooth.

Assumption 5.3.3. *For all $x \in \mathbb{R}^{d-1}$, we assume,*

$$|\nabla f^{\varepsilon}(0, x)| < K_1 \text{ and } |f^{\varepsilon}(0, x)| < K_2. \tag{5.3.2}$$

Assumption 5.3.4. *We assume that $F^{\varepsilon}_{\mathbf{x}}(\theta)$ is (a,b) - dissipative, i.e., for some $a > 0$, $b \geq 0$ and for all $\theta \in \mathbb{R}^d$, we assume that,*

$$\langle \nabla F^{\varepsilon}_{\mathbf{x}}(\theta), \theta \rangle \geq a|\theta|^2 - b. \tag{5.3.3}$$

Let us briefly discuss the assumption before stating the main results of the theorem. The assumptions 5.3.1, 5.3.2, and 5.3.3 are conventional assumptions in the literature concerning non-convex optimization. Assumption 5.3.4, also known as the dissipativity condition, typically used to analyze the ergodicity of the SDE and diffusion approximation, has also become a standard assumption to analyze the convergence Langevein Monte Carlo methods in the non-convex setting [64, 67, 71, 91]. We now state the main result of this article.

Theorem 5.3.1. *Suppose assumptions 5.2.1, 5.3.1, 5.3.2, 5.3.3 and 5.3.4 hold. If $h \leq h^*$, where,*

$$h^* = \min\left\{1, \frac{\min\{a, \sqrt{a}\}}{96(1+L)^2}, \frac{1}{2a}\right\}, \tag{5.3.4}$$

then for $\varepsilon > 0$, $t \in (n, n+1]$ and some constant $\bar{C}_8 > 0$,

$$\mathbb{E}[F(\theta^v_t)] - \mathrm{CVaR}_{\nu}(\Phi) \leq 2\varepsilon + \bar{C}_{1,7} h^{1/4} + \bar{C}_{2,7} e^{-\bar{C}_6 hn} + \frac{\bar{C}_8}{N} + \frac{d}{2\delta} \log\left(\frac{eL}{a}\left(\frac{b\delta}{d} + 1\right)\right), \tag{5.3.5}$$

where $\bar{C}_6$ is given in (5.3.78)*, whereas $\bar{C}_{1,7}$ and $\bar{C}_{2,7}$ are given in* (5.3.84)*.*

We highlight that the analysis undertaken below is inspired by the theoretical analysis carried out in [16, 63, 65, 76]. The very first step of the analysis is to determine the convergence result of the SVRG-LD iterations θ^v_n to the target distribution $\pi^{\mathbf{x},\varepsilon}_{\delta}$ in the Wasserstein distance. In order to do so, an important step is to obtain the moment

estimates for all the processes involved in the convergence analysis. Consequently, we obtain the upper bound for the Wasserstein distance, where we employ the splitting technique discussed in [16, 63]. Finally, in order to obtain the upper bound for excess risk, we rely on the theoretical discussion undertaken in [71]. We emphasize that, from a technical perspective, our analysis and results bear resemblance to those outlined in [76]. However, because of the structural disparity between the SVRG-LD algorithm and the conventional SGLD, the major challenge in our analysis is to derive the second and fourth-moment estimate of the θ_n^v. Moreover, this disparity also requires us to revisit some of the auxiliary results from [76] to complement our setup.

To begin with, we define below the Wasserstein distance as the majority of analysis deals with its estimation.

Definition 5.3.1. *[63] Consider ζ_1 and ζ_2 as two Borel probability measures on the space $\mathbb{R}^d$, each possessing a finite $p-$th moment. The Wasserstein distance of order $p \geq 1$ is then defined as follows:*

$$\mathcal{W}_p(\zeta_1, \zeta_2) \coloneqq \left(\inf_{\zeta \in \Gamma(\zeta_1, \zeta_2)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\theta_1 - \theta_2|^p \zeta(d\theta_1, d\theta_2) \right)^{1/2}, \tag{5.3.6}$$

where $\Gamma(\zeta_1, \zeta_2)$ set of all probability measures ζ on $\mathcal{B}(\mathbb{R}^{2d})$ with marginals ζ_1 and ζ_2.

We recall the Langevin SDE, which is given as,

$$dz_t = -\nabla F_{\mathbf{x}}^{\varepsilon}(z_t)dt + \sqrt{2\delta^{-1}}dW_t, \tag{5.3.7}$$

where $F_{\mathbf{x}}^{\varepsilon}(\theta)$ is the continuous time interpolation of $F_{\mathbf{x}}(\theta)$, $z_0 = \theta_0 \in \mathbb{R}^d$ is the initial conditions and $(W_t)_{t\geq 0}$ is the standard $d-$dimensional Brownian Motion equipped with complete natural filtration $(\mathcal{F}_t)_{t\geq 0}$. Further, we assume that $(\mathcal{F}_t)_{t\geq 0}$ is independent of $\sigma(\mathcal{G}_\infty \cup \sigma(\theta_0))$. Now, following the construction described in [16, 76], we consider a time change version of the SDE stated above. To this end, let $\tilde{z}_t = z_{ht}$ for $t \geq 0$ and for each $h > 0$, define $\widetilde{W}_t \coloneqq W_{ht}/\sqrt{h}$. Moreover, let $(\widetilde{\mathcal{F}}_t)_{t\geq 0} \coloneqq (\mathcal{F}_{ht})_{t\geq 0}$ be the natural filtration of $(\widetilde{W}_t)_{t\geq 0}$. Consequently, we have,

$$d\tilde{z}_t = -h\nabla F_{\mathbf{x}}^{\varepsilon}(\tilde{z}_t)dt + \sqrt{2h\delta^{-1}}d\widetilde{W}_t, \tag{5.3.8}$$

with $\tilde{z}_0 = \theta_0$ being the initial condition.

Let us consider the following continuous time interpolation of the SVRG-LD method,

$$d\xi_t^v = -hv_{\lfloor t \rfloor}dt + \sqrt{2\delta^{-1}h}d\widetilde{W}_t. \quad \xi_0^v = \theta_0, \tag{5.3.9}$$

with $v_{\lfloor t \rfloor}$ begin defined as

$$v_{\lfloor t \rfloor} := \nabla f^{\varepsilon}\left(\xi_{\lfloor t \rfloor}, X_{\lceil t \rceil}\right) - \nabla f^{\varepsilon}\left(\xi_{s_{\lfloor t \rfloor}}, X_{\lceil t \rceil}\right) + \nabla F_{\mathbf{x}}^{\varepsilon}\left(\xi_{s_{\lfloor t \rfloor}}\right). \tag{5.3.10}$$

where,

$$s_{\lfloor t \rfloor} := \max\{s \in \mathcal{S}^{\nabla} : t \geq s\}. \tag{5.3.11}$$

We would like to point out that the $\mathcal{L}(\theta_n^v) = \mathcal{L}(\xi_n^v)$ for all $n \in \mathbb{N}$. We also consider the following auxiliary process $\{\Xi_t^{s,w,h}\}_{t \geq s}$,

$$\mathrm{d}\Xi_t^{s,w,h} = -h\nabla F_{\mathbf{x}}^{\varepsilon}(\Xi_t^{s,w,h})\mathrm{d}t + \sqrt{2\delta^{-1}h}\mathrm{d}\tilde{W}_t \tag{5.3.12}$$

where the initial condition is $\Xi_s^{s,w,h} = w$. We now further denote, $\Xi_{t,v}^{k,h} := \Xi_t^{kT,\xi_{kT}^v,h}$, where $T := \left\lfloor \frac{1}{h} \right\rfloor$. Moreover, for each $p \in [2,\infty) \cap \mathbb{N}$ and for all $\theta \in \mathbb{R}^d$, consider the Lyapunov function $\mathcal{V}_p(\theta) := (1 + |\theta|^2)^{p/2}$. Also, let $\mathfrak{v}_p(x) = (1 + x^2)^{p/2}$ for all $x \geq 0$ and denote by $\mathcal{P}_{\mathcal{V}_p}(\mathbb{R}^d)$ the set of all probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ such that $\int_{\mathbb{R}^d} \mathcal{V}_p(\theta)\mu(d\theta) < \infty$ for all $p \geq 1$.

We now recall the following results, which show that $\mathcal{V}_p(\theta)$ satisfies the geometric drift condition.

Lemma 5.3.1. *Suppose assumption 5.3.4 holds. Then for any $\theta \in \mathbb{R}^d$, $p \in [2,\infty) \cap \mathbb{N}$, we have,*

$$-\langle \nabla \mathcal{V}_p(\theta), \nabla F_{\mathbf{x}}^{\varepsilon}(\theta)\rangle + \Delta \mathcal{V}_p(\theta)\delta^{-1} \leq -C_{\mathcal{V},1}(p)\mathcal{V}_p(\theta) + C_{\mathcal{V},2}(p), \tag{5.3.13}$$

where $C_{\mathcal{V},1}(p) := \frac{ap}{4}$, $C_{\mathcal{V},2}(p) := \frac{3ap\mathfrak{v}_p(K_{\mathcal{V}}(p))}{4}$ *with* $K_{\mathcal{V}}(p) := \left(\frac{1}{3} + \frac{4b}{3a} + \frac{4d}{3a\delta} + \frac{4(p-2)}{3a\delta}\right)^{1/2}$.

Proof. Refer to [16, Lemma 3.5]. □

5.3.1 Moment Estimates

In this sub-section, we present the moment estimates of the process $(\xi_t^v)_{t \geq 0}$ and $(\Xi_{t,v}^{k,h})_{t \geq kT}$.

Lemma 5.3.2. *Suppose assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 hold. Then we have the following:*

(a) For any $0 < h \leq h^*$ *given in* (5.3.4), $n \in \mathbb{N}$ *and* $t \in (n, n+1]$, *we have,*

$$\mathbb{E}[|\xi_t^v|^2] \leq \mathbb{E}[|\theta_0|^2](1-ah)^n(1-ah(t-n)) + \mathring{B}\left(1+\frac{1}{a}\right), \tag{5.3.14}$$

where $\mathring{B} = 2\left(K_1^2 + b + \frac{d}{\delta} + 2L^2\kappa_{0,2}\right)$.

(b) Let $M_1(q) := 2^{2q-2}L^q$, $M_2(q) := 2^{2q-1}L^q$, $M_3(q) := 2^{2q-2}K_1^q$ *and*

$$\widehat{M} := \max\left\{\left(\frac{12M_3(2)+8b}{a}\right)^{1/2}, \left(\frac{8M_3(3)}{a}\right)^{1/3}\right\}.$$

Then for any $0 < h \leq h^*$ *given in* (5.3.4), $n \in \mathbb{N}$ *and* $t \in (n, n+1]$, *we have,*

$$\mathbb{E}[|\theta_t^v|^4] \leq (1-ah(t-n))(1-ah)^n\mathbb{E}[|\theta_0|^4] + \mathring{C}\left(1+\frac{1}{a}\right) \tag{5.3.15}$$

where $\mathring{C} = (1+a)\left(4b\widehat{M}^2 + 6M_3(2)\widehat{M}^2 + 4M_3(3)\widehat{M} + M_3(4) + a\kappa_{0,4}\right) + \frac{(9+a)}{a}12d^2\delta^{-2}$

Proof (a). Let $t \in (n, n+1]$ and define $\Delta_{n,t}^v := \xi_n^v - h(t-n)v_n$ with $\Upsilon_{n,t}^h := \sqrt{2h\delta^{-1}}(\widetilde{W}_t - \widetilde{W}_n)$. Also let $s_n = \max\{s \in \mathcal{S}^\nabla : n \geq s\}$. Then we have,

$$\begin{aligned}
\mathbb{E}[|\xi_t^v|^2|\xi_n^v, \xi_{s_n}^v] &= \mathbb{E}[|\Delta_{n,t}^v + \Upsilon_{n,t}^h|^2|\xi_n^v, \xi_{s_n}^v] \\
&= \mathbb{E}[|\Delta_{n,t}^v|^2 + |\Upsilon_{n,t}^h|^2 + 2\langle\Delta_{n,t}^v, \Upsilon_{n,t}^h\rangle|\xi_n^v, \xi_{s_n}^v] \\
&= \mathbb{E}[|\delta_{n,t}^v|^2|\xi_n^v, \xi_{s_n}^v] + \mathbb{E}[|\Upsilon_{n,t}^h|^2|\xi_n^v, \xi_{s_n}^v] \\
&= \mathbb{E}[|\xi_n^v - hv_n(t-n)|^2|\xi_n^v, \xi_{s_n}^v] + \frac{2hd}{\delta}(t-n).
\end{aligned} \tag{5.3.16}$$

Expanding the last term of the above equation, we get,

$$\begin{aligned}\mathbb{E}[|\xi_n^v - hv_n(t-n)|^2|\xi_n^v, \xi_{s_n}^v] &= \mathbb{E}[|\xi_n^v - h(t-n)\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v) + h(t-n)\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v) - hv_n(t-n)|^2|\xi_n^v, \xi_{s_n}^v] \\ &= \mathbb{E}[|\xi_n^v - h(t-n)\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v)|^2|\xi_n^v, \xi_{s_n}^v] \\ &\quad + \mathbb{E}[|h(t-n)\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v) - hv_n(t-n)|^2|\xi_n^v, \xi_{s_n}^v] \\ &\quad + 2\mathbb{E}[\langle \xi_n^v - \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v), \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v) - hv_n(t-n)\rangle|\xi_n^v, \xi_{s_n}^v] \\ &= \underbrace{\mathbb{E}[|\xi_n^v - h(t-n)\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v)|^2|\xi_n^v, \xi_{s_n}^v]}_{A_1} \\ &\quad + \underbrace{\mathbb{E}[|h(t-n)\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v) - hv_n(t-n)|^2|\xi_n^v, \xi_{s_n}^v]}_{A_2}.\end{aligned} \tag{5.3.17}$$

Expanding A_1, we have,

$$\begin{aligned}A_1 &= \mathbb{E}[|\xi_n^v|^2|\xi_n^v, \xi_{s_n}^v] + h^2(t-n)^2\mathbb{E}[|\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v)|^2|\xi_n^v, \xi_{s_n}^v] - 2h(t-n)\mathbb{E}[\langle \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v), \xi_n^v\rangle|\xi_n^v, \xi_{s_n}^v] \\ &\leq |\xi_n^v|^2 + h^2(t-n)^2|\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_n^v)|^2 - 2h(t-n)(a|\xi_n^v|^2 - b) \\ &\leq |\xi_n^v|^2 + 2h^2(t-n)^2(K_1^2 + L^2|\xi_n^v|^2) - 2h(t-n)(a|\xi_n^v|^2 - b) \\ &\leq |\xi_n^v|\left(1 + 2h^2(t-n)^2L^2 - 2h(t-n)a\right) + 2h^2(t-n)^2K_1^2 + 2h(t-n)b.\end{aligned} \tag{5.3.18}$$

Expanding A_2, we have,

$$\begin{aligned}A_2 &= h^2(t-n)^2\mathbb{E}[\left(\nabla f^{\varepsilon}(\xi_n^v, Y_{n+1}^{\mathbf{x}}) - \nabla f^{\varepsilon}(\xi_{s_n}^v, Y_{n+1}^{\mathbf{x}}) - \mathbb{E}[\nabla f^{\varepsilon}(\xi_n^v, Y_{n+1}^{\mathbf{x}}) - \nabla f^{\varepsilon}(\xi_{s_n}^v, Y_{n+1}^{\mathbf{x}})]\right)^2|\xi_n^v, \xi_{s_n}^v] \\ &\leq h^2(t-n)^2\mathbb{E}[\left(\nabla f^{\varepsilon}(\xi_n^v, Y_{n+1}^{\mathbf{x}}) - \nabla f^{\varepsilon}(\xi_{s_n}^v, Y_{n+1}^{\mathbf{x}})\right)^2|\xi_n^v, \xi_{s_n}^v] \\ &\leq h^2(t-n)^2L^2|\xi_n^v - \xi_{s_n}^v|^2 \leq 2h^2(t-n)^2L^2\left(|\xi_n^v|^2 + |\xi_{s_n}^v|^2\right).\end{aligned} \tag{5.3.19}$$

Substituting A_1 and A_2 back in equation (15), we get,

$$\begin{aligned}\mathbb{E}[|\xi_n^v - hv_n(t-n)|^2|\xi_n^v, \xi_{s_n}^v] &\leq |\xi_n^v|^2\left(1+2h^2(t-n)^2L^2-2h(t-n)a\right)+2h^2(t-n)^2K_1^2\\ &\quad+2h(t-n)b+2h^2(t-n)^2L^2\left(|\xi_n^v|^2+|\xi_{s_n}^v|^2\right)\\ &=|\xi_n^v|^2\left(1+4h^2(t-n)^2L^2-2h(t-n)a\right)+2h^2(t-n)^2K_1^2\\ &\quad+2h(t-n)b+2h^2(t-n)^2L^2|\xi_{s_n}^v|^2.\end{aligned} \tag{5.3.20}$$

Now, since $h<1$, we have,

$$\mathbb{E}[|\xi_t^v|^2] \leq \mathbb{E}[|\xi_n^v|^2]\left(1+4h^2(t-n)^2L^2-2h(t-n)a\right)+h(t-n)B+2h^2(t-n)^2L^2\mathbb{E}[|\xi_{s_n}^v|^2], \tag{5.3.21}$$

where $B=2\left(K_1^2+b+\frac{d}{\delta}\right)$. Now let $t\in(0,1]$, therefore $s_0=0$ and $\xi_0^v=\theta_0$(by our assumption). Consequently, we have,

$$\begin{aligned}\mathbb{E}[|\xi_t^v|^2] &\leq \mathbb{E}[|\theta_0|^2]\left(1+4h^2(t-n)^2L^2-2h(t-n)a\right)+h(t-n)\mathring{B}+2h^2(t-n)^2L^2\mathbb{E}[|\theta_0|^2]\\ &=\mathbb{E}[|\theta_0|^2]\left(1+4h^2(t-n)^2L^2-2h(t-n)a\right)+h(t-n)\mathring{B}.\end{aligned} \tag{5.3.22}$$

where $\mathring{B}=B+2L^2\kappa_{0,2}$. Now, since $h\leq h^*$ we have,

$$\begin{aligned}\mathbb{E}[|\xi_t^v|^2] &\leq \mathbb{E}[|\theta_0|^2]\left(1-ah(t-n)\right)+h(t-n)\mathring{B}\\ &\leq \mathbb{E}[|\theta_0|^2]\left(1-ah(t-n)\right)+\mathring{B}\left(1+\frac{1}{a}\right).\end{aligned} \tag{5.3.23}$$

Now, suppose that, for all $k\leq n-1$ and $t\in(k,k+1]$, we assume,

$$\mathbb{E}[|\xi_t^v|^2] \leq \mathbb{E}[|\theta_0|^2](1-ah)^k(1-ah(t-k))+\mathring{B}\left(1+\frac{1}{a}\right). \tag{5.3.24}$$

We prove that the conclusion is true for $t \in (n, n+1]$.

$$\begin{aligned}
\mathbb{E}[|\xi_t^v|^2] &\leq \mathbb{E}[|\xi_n^v|^2]\left(1+4h^2(t-n)^2L-2ah(t-n)\right)+h(t-n)B+2h^2(t-n)^2L^2\mathbb{E}[|\xi_{s_n}^v|^2] \\
&\leq \left(\mathbb{E}[|\theta_0|^2](1-ah)^n+\mathring{B}\left(1+\frac{1}{a}\right)\right)\left(1+4h^2(t-n)^2L-2ah(t-n)\right)+h(t-n)B \\
&+2h^2(t-n)^2L^2\left(\mathbb{E}[|\theta_0|^2](1-ah)^{s_n}+\mathring{B}\left(1+\frac{1}{a}\right)\right) \\
&\leq \mathbb{E}[|\theta_0|^2](1-ah)^n\left(1+4h^2(t-n)^2L-2ah(t-n)\right))+h(t-n)\mathring{B} \\
&+\mathring{B}\left(1+\frac{1}{a}\right)\left(1+6h^2(t-n)^2L-2ah(t-n)\right) \\
&\leq \mathbb{E}[|\theta_0|^2](1-ah)^n(1-ah(t-n))+\mathring{B}\left(1+\frac{1}{a}\right).
\end{aligned} \tag{5.3.25}$$

Therefore, by induction, the result follows. □

Proof (b). As before let $t \in (n, n+1]$, $\Delta_{n,t}^v := \xi_n^v - h(t-n)v_n, \Upsilon_{n,t}^h := \sqrt{2h\delta^{-1}}(\widetilde{W}_t - \widetilde{W}_n)$, and $s_n = \max\{s \in \mathcal{S}^\nabla : n \geq s\}$. Then we have,

$$\begin{aligned}
\mathbb{E}[|\xi_t^v|^4|\xi_n^v, \xi_{s_n}^v] &= \mathbb{E}[|\Delta_{n,t}^v + \Upsilon_{n,t}^h|^4|\xi_n^v, \xi_{s_n}^v] \\
&= \mathbb{E}[(|\Delta_{n,t}^v|^2 + |\Upsilon_{n,t}^h|^2 + 2\langle\Delta_{n,t}^v, \Upsilon_{n,t}^h\rangle)^2|\xi_n^v, \xi_{s_n}^v] \\
&= \mathbb{E}[|\Delta_{n,t}^v|^4 + |\Upsilon_{n,t}^h|^4 + 4(\langle\Delta_{n,t}^v, \Upsilon_{n,t}^h\rangle)^2 + 2|\Delta_{n,t}^v|^2|\Upsilon_{n,t}^h|^2 \\
&+ 4|\Delta_{n,t}^v|^2\langle\Delta_{n,t}^v, \Upsilon_{n,t}^h\rangle + 4|\Upsilon_{n,t}^h|^2\langle\Delta_{n,t}^v, \Upsilon_{n,t}^h\rangle|\xi_n^v, \xi_{s_n}^v] \\
&= \mathbb{E}[|\Delta_{n,t}^v|^4 + |\Upsilon_{n,t}^h|^4 + 2|\Delta_{n,t}^v|^2|\Upsilon_{n,t}^h|^2 + 4(\langle\Delta_{n,t}^v, \Upsilon_{n,t}^h\rangle)^2|\xi_n^v, \xi_{s_n}^v] \\
&\leq \mathbb{E}[|\Delta_{n,t}^v|^4 + |\Upsilon_{n,t}^h|^4 + 6|\Delta_{n,t}^v|^2|\Upsilon_{n,t}^h|^2|\xi_n^v, \xi_{s_n}^v] \\
&\leq (1+ah(t-n))\,\mathbb{E}[|\Delta_{n,t}^v|^4|\xi_n^v, \xi_{s_n}^v] + \left(1+\frac{9}{ah(t-n)}\right)\mathbb{E}[|\Upsilon_{n,t}^h|^4].
\end{aligned} \tag{5.3.26}$$

Also,

$$\begin{aligned}
\mathbb{E}[|\Delta_{n,t}^{v}|^4|\xi_n^v,\xi_{s_n}^v] &= \mathbb{E}[|\xi_n^v - v_n h(t-n)|^4|\xi_n^v,\xi_{s_n}^v] \\
&= \mathbb{E}[\left(|\xi_n^v|^2 + h^2(t-n)^2|v_n|^2 - 2h(t-n)\langle\xi_n^v,v_n\rangle\right)^2|\xi_n^v,\xi_{s_n}^v] \\
&= \mathbb{E}[|\xi_n^v|^4 + h^4(t-n)^4|v_n|^4 + 4h^2(t-n)^2(\langle\xi_n^v,v_n\rangle)^2 + 2h^2(t-n)^2|\xi_n^v|^2|v_n|^2 \\
&- 4h(t-n)|\xi_n^v|^2\langle\xi_n^v,v_n\rangle - 4h(t-n)|v_n|^2\langle\xi_n^v,v_n\rangle|\xi_n^v,\xi_{s_n}^v] \\
&\leq \mathbb{E}[|\xi_n^v|^4 + h^4(t-n)^4|v_n|^4 + 6h^2(t-n)^2|\xi_n^v|^2|v_n|^2 - 4h^3(t-n)^3|\xi_n^v|^2\langle\xi_n^v,v_n\rangle \\
&- 4h^3(t-n)^3|v_n|^2\langle\xi_n^v,v_n\rangle|\xi_n^v,\xi_{s_n}^v] \\
&\leq |\xi_n^v|^4 - 4h(t-n)|\theta_n^v|^2\mathbb{E}[\langle\xi_n^v,v_n\rangle|\xi_n^v,\xi_{s_n}^v] \\
&+ \sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q|\xi_n^v|^{4-q}\mathbb{E}[|v_n|^q|\xi_n^v,\xi_{s_n}^v]
\end{aligned} \tag{5.3.27}$$

Observe that,

$$\begin{aligned}
\mathbb{E}[|v_n|^q|\xi_n^v,\xi_{s_n}^v] &= \mathbb{E}[|\left(\nabla f^\varepsilon(\xi_n^v,Y_{n+1}^{\mathbf{x}}) - \nabla f^\varepsilon(\xi_{s_n}^v,Y_{n+1}^{\mathbf{x}}) + \nabla F_{\mathbf{x}}^\varepsilon(\xi_{s_n}^v)\right)|^q|\xi_n^v,\xi_{s_n}^v] \\
&\leq 2^{q-1}\left(\mathbb{E}[|\nabla f^\varepsilon(\xi_n^v,Y_{n+1}^{\mathbf{x}}) - \nabla f^\varepsilon(\xi_{s_n}^v,Y_{n+1}^{\mathbf{x}})|^q|\xi_n^v,\xi_{s_n}^v] + \mathbb{E}[|\nabla F_{\mathbf{x}}^\varepsilon(\xi_{s_n}^v)|^q|\xi_n^v,\xi_{s_n}^v]\right) \\
&\leq 2^{q-1}L^q|\xi_n^v - \xi_{s_n}^v|^q + 2^{q-1}(K_1 + L|\xi_{s_n}^v|)^q \\
&\leq 2^{2q-2}L^q(|\xi_n^v|^q + |\xi_{s_n}^v|^q) + 2^{2q-2}(K_1^q + L^q|\xi_{s_n}^v|^q) \\
&= 2^{2q-2}L^q|\xi_n^v|^q + 2^{2q-1}L^q|\xi_{s_n}^v|^q + 2^{2q-2}K_1^q \\
&= M_1(q)|\xi_n^v|^q + M_2(q)|\xi_{s_n}^v|^q + M_3(q).
\end{aligned} \tag{5.3.28}$$

Further, using the dissipativity condition we have, $-\mathbb{E}[\langle\xi_n^v,v_n\rangle|\xi_n^v,\xi_{s_n}^v] \leq -(a|\xi_n^v|^2 - b)$.

Therefore, we have,

$$\begin{aligned}\mathbb{E}[|\Delta^v_{n,t}|^4|\xi^v_n,\xi^v_{s_n}] \leq |\xi^v_n|^4\left(1-4h(t-n)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_1(q)\right)+4h(t-n)b|\xi^v_n|^2\\ +\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_2(q)|\xi^v_n|^{4-q}|\xi^v_{s_n}|^q\\ +\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_3(q)|\xi^v_n|^{4-q}.\end{aligned} \tag{5.3.29}$$

For $h<h^*$ and let $t\in(0,1]$, then , we have,

$$\begin{aligned}\mathbb{E}[|\Delta^v_{0,t}|^4|\theta_0] \leq&|\theta_0|^4\left(1-4h(t)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t)^qM_1(q)\right)+4h(t)b|\theta^v_0|^2\\ &+\sum_{q=2}^{4}\binom{4}{q}h^q(t)^qM_2(q)|\theta_0|^{4-q}|\theta_0|^q+\sum_{q=2}^{4}\binom{4}{q}h^q(t)^qM_3(q)|\theta_0|^{4-q}\\ &\leq|\theta_0|^4\left(1-4h(t)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t)^q(M_1(q)+M_2(q))\right)+4h(t)b|\theta^v_0|^2\\ &+\sum_{q=2}^{4}\binom{4}{q}h^q(t)^qM_3(q)|\theta_0|^{4-q}\\ &\leq|\theta_0|^4(1-3hta)+4htb|\theta_0|^2+\sum_{q=2}^{4}\binom{4}{q}h^q(t)^qM_3(q)|\theta_0|^{4-q}\end{aligned} \tag{5.3.30}$$

Let $\mathfrak{S}_{0,\widehat{M}}:=\{\omega\in\Omega:|\theta_0|>\widehat{M}\}$. Consequently, we have,

$$\mathbb{E}[|\Delta^v_{0,t}|^4|\theta_0]=\mathbb{E}[|\Delta^v_{0,t}|^4\mathbb{I}_{\mathfrak{S}_{0,\widehat{M}}}|\theta_0]+\mathbb{E}[|\Delta^v_{0,t}|^4\mathbb{I}_{\mathfrak{S}^c_{0,\widehat{M}}}|\theta_0] \tag{5.3.31}$$

Furthermore, observing that, for $|\theta_0|>\widehat{M}$,

$$-\frac{ah(t)|\theta_0|^4}{2}+4h^3(t)^3M_3(3)|\theta_0|<0, \tag{5.3.32}$$

$$-\frac{ah(t)|\theta_0|^4}{2}+6h^2(t)^2M_3(2)|\theta_0|^2+4h(t)b|\theta_0|^2<0, \tag{5.3.33}$$

we have,

$$\mathbb{E}[|\Delta_{0,t}^{v}|^4\mathbb{I}_{\mathfrak{S}_{0,\widehat{M}}}|\theta_0] \leq (1-2ah(t))|\theta_0^v|^4\mathbb{I}_{\mathfrak{S}_{0,\widehat{M}}} + h^4(t)^4M_3(4)\mathbb{I}_{\mathfrak{S}_{0,\widehat{M}}}, \tag{5.3.34}$$

whereas,

$$\mathbb{E}[|\Delta_{0,t}^{v}|^4\mathbb{I}_{\mathfrak{S}^c_{0,\widehat{M}}}|\theta_0] \leq |\theta_0|^4(1-2ah(t))\mathbb{I}_{\mathfrak{S}^c_{0,\widehat{M}}} \tag{5.3.35}$$

$$+ h(t)\left(4b\widehat{M}^2 + 6M_3(2)\widehat{M}^2 + 4M_3(3)\widehat{M} + M_3(4)\right)\mathbb{I}_{\mathfrak{S}^c_{0,\widehat{M}}}. \tag{5.3.36}$$

Combining both, we get,

$$\mathbb{E}[|\Delta_{0,t}^{v}|^4|\theta_0] \leq (1-2ah(t))|\theta_0^v|^4 + h(t)\hat{b}, \tag{5.3.37}$$

where $\hat{b} \coloneqq \left(4b\widehat{M}^2 + 6M_3(2)\widehat{M}^2 + 4M_3(3)\widehat{M} + M_3(4)\right)$. Consequently, we have,

$$\begin{aligned}\mathbb{E}[|\theta_t^v|^4|\theta_0] &\leq (1+ah(t)2)\left((1-2ah(t))|\theta_0^v|^4 + h(t)\hat{b}\right) + \left(\frac{(ah(t)+9)}{ah(t)}12h^2d^2\delta^{-2}(t)^2\right)\\ &\leq (1-ah(t))\,|\theta_0|^4 + \mathring{C},\end{aligned} \tag{5.3.38}$$

where $\mathring{C} = (1+a)\,(\hat{b}+a\kappa_{0,4}) + \dfrac{(9+a)}{a}12d^2\delta^{-2}$. Therefore, we have,

$$\mathbb{E}[|\xi_t^v|^4] \leq (1-ah(t))\,\mathbb{E}[|\theta_0|^4] + \mathring{C}h(t) \leq (1-ah(t))\,\mathbb{E}[|\theta_0|^4] + \mathring{C}\left(1+\frac{1}{a}\right), \tag{5.3.39}$$

proving that the conclusion is true for $t \in (0,1]$. Let us assume that the conclusion is true for $t \in (k, k+1]$ for all $k \leq n-1$, and we intend to prove that the conclusion is true for $t \in (n, n+1]$. To this end, we have,

$$\begin{aligned}\mathbb{E}[|\Delta_{n,t}^{v}|^4|\xi_n^v,\xi_{s_n}] \leq |\xi_n^v|^4&\left(1-4h(t-n)a + \sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_1(q)\right) + 4h(t-n)b|\xi_n^v|^2\\ &+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_2(q)|\xi_n^v|^{4-q}|\xi_{s_n}|^q + \sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_3(q)|\xi_n^v|^{4-q}.\end{aligned}$$

Suppose , $|\xi_n^v| \geq |\xi_{s_n}^v|$, therefore we have,

$$\begin{aligned}\mathbb{E}[|\Delta_{n,t}^v|^4|\xi_n^v,\xi_{s_n}] &\leq |\xi_n^v|^4\left(1-4h(t-n)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q(M_1(q)+M_2(q))\right)\\ &\quad+4h(t-n)b|\xi_n^v|^2+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_3(q)|\xi_n^v|^{4-q}\\ &\leq |\xi_n^v|^4\left(1-3ha(t-n)\right)\\ &\quad+4h(t-n)b|\xi_n^v|^2+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_3(q)|\xi_n^v|^{4-q}.\end{aligned} \tag{5.3.40}$$

Now, as before, the above inequality can be further bounded as,

$$\mathbb{E}[|\Delta_{n,t}^v|^4|\xi_n^v,\xi_{s_n}^v] \leq (1-2ah(t-n))|\xi_n^v|^4+h(t-n)\hat{b}. \tag{5.3.41}$$

Therefore, following a similar argument, we can conclude that,

$$\mathbb{E}[|\xi_t^v|^4] \leq (1-ah(t-n))\mathbb{E}[|\xi_n^v|^4]+\mathring{C} \leq (1-ah(t-n))(1-ah)^n\mathbb{E}[|\theta_0|^4]+\mathring{C}\left(1+\frac{1}{a}\right). \tag{5.3.42}$$

Now, suppose that $|\xi_n^v| < |\xi_{s_n}^v|$. Then we have,

$$\begin{aligned}\mathbb{E}[|\Delta_{n,t}^v|^4|\xi_n^v,\xi_{s_n}] &\leq |\xi_n^v|^4\left(1-4h(t-n)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_1(q)\right)+4h(t-n)b|\theta_n^v|^2\\ &\quad+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_2(q)|\xi_{s_n}|^4+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_3(q)|\xi_n^v|^{4-q}.\\ &\leq |\xi_n^v|^4\left(1-3h(t-n)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_1(q)\right)\\ &\quad+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^qM_2(q)|\xi_{s_n}|^4+h(t-n)\hat{b}.\end{aligned} \tag{5.3.43}$$

Taking expectations on both sides in the above equation, we obtain,

$$\begin{aligned}
\mathbb{E}[|\Delta_{n,t}^v|^4] &\leq \mathbb{E}[|\xi_n^v|^4]\left(1-3h(t-n)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q M_1(q)\right)\\
&\quad+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q M_2(q)\mathbb{E}[|\xi_s^v|^4]+h(t-n)\hat{b}\\
&\leq\left((1-ah)^n\mathbb{E}[|\theta_0|^4]+\mathring{C}\left(1+\frac{1}{a}\right)\right)(1-3h(t-n)a\\
&\quad+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q(M_1(q)+M_2(q))\Bigg)\\
&\quad+\left(\mathbb{E}[|\theta_0|^4]+\mathring{C}\left(1+\frac{1}{a}\right)\right)\left(\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q(M_1(q)+M_2(q))\right)+h(t-n)\hat{b}\\
&\leq(1-ah)^s\mathbb{E}[|\theta_0|^4]\left(1-3h(t-n)a+\sum_{q=2}^{4}\binom{4}{q}h^q(t-n)^q(M_1(q)+M_2(q))\right)\\
&\qquad+\mathring{C}\left(1+\frac{1}{a}\right)(1-2ha(t-n))+h(t-n)(\hat{b}+a\kappa_{0,4})\\
&\leq(1-ah)^n(1-2ah(t-n))\mathbb{E}[|\theta_0|^4]+\mathring{C}\left(1+\frac{1}{a}\right)(1-2ha(t-n))+h(t-n)(\hat{b}+a\kappa_{0,4})
\end{aligned}\tag{5.3.44}$$

Taking expectation on both sides of equation (5.3.26) and substituting the above inequality, we get,

$$\begin{aligned}
\mathbb{E}[|\xi_t^v|^4] &\leq (1+ah(t-n))\Bigg((1-ah)^n(1-2ah(t-n))\mathbb{E}[|\theta_0|^4]\\
&\quad+\mathring{C}\left(1+\frac{1}{a}\right)(1-2ha(t-n))+h(t-n)(\hat{b}+a\kappa_{0,4})\Bigg)\\
&\quad+\left(1+\frac{9}{ah(t-n)}\right)12h^2d^2\delta^{-2}(t-n)^2\\
&\leq(1-ah(t-n))(1-ah)^n\mathbb{E}[|\theta_0|^4]+\mathring{C}\left(1+\frac{1}{a}\right).
\end{aligned}\tag{5.3.45}$$

Therefore, by induction, the conclusion holds. □

We have the following corollary as the consequence of the above analysis.

Corollary 5.3.1. *Suppose assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 hold. Then for*

$0 < h \leq h^*$ (5.3.4), *and for any* $t \in \mathbb{N}$, *we have,*

1. $\mathbb{E}[\mathcal{V}_2(\xi_t^v)] \leq 1 + (1-ah)^{\lfloor t \rfloor}\mathbb{E}[\mathcal{V}_2(\theta_0)] + \mathring{B}\left(1 + \frac{1}{a}\right)$,

2. $\mathbb{E}[\mathcal{V}_4(\xi_t^v)] \leq 2 + 2(1-ah)^{\lfloor t \rfloor}\mathbb{E}[\mathcal{V}_4(\theta_0)] + 2\mathring{C}\left(1 + \frac{1}{a}\right)$.

Proof. The proof follows directly from the definition of $\mathcal{V}_p(\theta)$ and Lemma 5.3.2. □

The following result provides the moment estimate of the process $(\Xi_{t,v}^{n,h})_{t \geq nT}$.

Lemma 5.3.3. *Suppose assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 hold. Then for* $0 < h \leq h^*$ *given in* (5.3.4), *and for any* $t \geq nT$, *we have,*

$$\begin{aligned} &1.\ \ \mathbb{E}[\mathcal{V}_2(\Xi_{t,v}^{n,h})] \leq e^{-aht/2}\mathbb{E}[\mathcal{V}_2(\theta_0)] + 3\mathfrak{v}_3(K_{\mathcal{V}}(2)) + \mathring{B}\left(1 + \frac{1}{a}\right) + 1 \\ &2.\ \ \mathbb{E}[\mathcal{V}_4(\Xi_{t,v}^{n,h})] \leq 2e^{-aht}\mathbb{E}[\mathcal{V}_4(\theta_0)] + 3\mathfrak{v}_5(K_{\mathcal{V}}(4)) + 2\mathring{C}\left(1 + \frac{1}{a}\right) + 2 \end{aligned} \tag{5.3.46}$$

Proof. The proof follows the same line of argument as in [76] where Lemma 5.3.2 and Lemma 5.3.1 are used to obtain the constants. □

5.3.2 Auxiliary Results

Lemma 5.3.4. *Suppose Assumption 5.3.1, 5.3.2, 5.3.3 and 5.3.4 hold. Then for* $0 < h \leq h^*$ *and any* $t > 0$, *we have,*

$$\mathbb{E}[|\xi_t^v - \xi_{\lfloor t \rfloor}^v|^2] \leq h\left(ae^{-ah\lfloor t \rfloor}\mathbb{E}[\mathcal{V}_2(\theta_0)] + B_1\right) \tag{5.3.47}$$

where $B_1 = a\left(\mathring{B}\left(1 + \frac{1}{a}\right) + \kappa_{0,2} + M_3(2) + 2\delta^{-1}d\right)$.

Proof. Observe that,

$$\begin{aligned} \mathbb{E}[|\xi_t^v - \xi_{\lfloor t \rfloor}^v|^2] &= \mathbb{E}\left[\left| -h\int_{\lfloor t \rfloor}^t v_{\lfloor t \rfloor}du + \sqrt{2h\delta^{-1}}(\tilde{W}_t - \tilde{W}_{\lfloor t \rfloor})\right|^2\right] \\ &\leq h^2\mathbb{E}[|v_{\lfloor t \rfloor}|^2] + 2h\delta^{-1}d \\ &= h^2\mathbb{E}[\mathbb{E}[|v_{\lfloor t \rfloor}|^2|\xi_{\lfloor t \rfloor}^v, \xi_{s_{\lfloor t \rfloor}}^v]] + 2h\delta^{-1}d \\ &\leq h^2M_1(2)\mathbb{E}[|\theta_{\lfloor t \rfloor}^v|^2] + M_2(2)h^2\mathbb{E}[|\theta_{s_{\lfloor t \rfloor}}^v|^2] + hM_3(2) + 2h\delta^{-1}d \end{aligned}$$

$$\leq h^2 M_1(2)\left((1-ah)^{\lfloor t\rfloor}\mathbb{E}[|\theta_0|^2] + \mathring{B}\left(1+\frac{1}{a}\right)\right)$$
$$+ h^2 M_2(2)\left(\mathbb{E}[|\theta_0|^2] + \mathring{B}\left(1+\frac{1}{a}\right)\right) + hM_3(2) + 2h\delta^{-1}d. \tag{5.3.48}$$

Now observing that for $h < h^*$,

$$h(M_1(2)+M_2(2)) \leq a \text{ and } hM_1(2) \leq a, \tag{5.3.49}$$

we have,

$$\mathbb{E}[|\xi_t^v - \xi_{\lfloor t\rfloor}^v|^2] \leq h(a\mathbb{E}[|\theta_0|^2](1-ah)^{\lfloor t\rfloor} + B_1) \leq h\left(ae^{-ah\lfloor t\rfloor}\mathbb{E}[\mathcal{V}_2(\theta_0)] + B_1\right), \tag{5.3.50}$$

where $B_1 = a\left(\mathring{B}\left(1+\frac{1}{a}\right) + \kappa_{0,2} + M_3(2) + 2\delta^{-1}d\right)$. □

Lemma 5.3.5. *Suppose Assumption 5.3.1, 5.3.2, 5.3.3 and 5.3.4 hold. Then for $T \in \mathbb{N}$ and, $t \in [nT, (n+1)T]$, we have,*

$$h^2\mathbb{E}\left[\left|\left(\int_{mT}^t \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u\rfloor}^v) - v_{\lfloor u\rfloor}\right)du\right|^2\right] \leq h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)] + \mathring{B}(a+1) + a\kappa_{0,2}\right) \tag{5.3.51}$$

Proof. Observe that given nT, there exists an $s_T \in \mathcal{S}^{\nabla}$, such that $nT \in [s_T, 2s_T - 1]$. Therefore, we have, $nT = s_T + \bar{L}$, where $\bar{L} \in \{0, 1, \ldots, s_T - 1\}$. Further, we have that, $(n+1)T \in [K_T s_T, 2K_T s_T - 1]$ where $K_T \in \{1, 2, 4, \ldots\}$. Now for a given t there exists, $K_t \in \{1, 2, 4, \ldots, K_T\}$ such that, $t \in [K_t s_T + \ell, K_t s_T + \ell + 1]$ where $\ell + 1 \leq K_t s_T - 1$. Consequently, we have,

$$\begin{aligned}\left(\int_{nT}^t \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u\rfloor}^v) - v_{\lfloor u\rfloor}du\right) &= \sum_{k=\bar{L}}^{s_T-1}\int_{s_T+k}^{s_T+k+1}\left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u\rfloor}^v) - v_{\lfloor u\rfloor}du\right)\\ &+ \sum_{j=2}^{K_t-1}\sum_{k=0}^{js_T-1}\int_{s_Tj+k}^{s_Tj+k+1}\left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u\rfloor}^v) - v_{\lfloor u\rfloor}\right)du\\ &+ \sum_{k=0}^{\ell-1}\int_{s_TK_t+k}^{s_TK_t+k+1}\left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u\rfloor}^v) - v_{\lfloor u\rfloor}\right)du\\ &+ \int_{s_TK_t+\ell}^{t}\left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u\rfloor}^v) - v_{\lfloor u\rfloor}\right)du\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=\bar{L}}^{s_T-1} (\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{s_T+k}) - v_{s_T+k}) + \sum_{j=2}^{K_t-1} \sum_{k=0}^{js_T-1} \left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{s_T j+k}^{v}) - v_{s_T j+k}\right) \\
&+ \sum_{k=0}^{\ell-1} \left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{s_T K_t+k}^{v}) - v_{s_T K_t+k}\right) \\
&+ (t - (s_T K_t + \ell)) \left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{s_T K_t+\ell}^{v}) - v_{s_T K_t+\ell}\right) \\
&= \sum_{k=\bar{L}}^{s_T-1} A_{1k} + \sum_{j=2}^{K_t-1} \sum_{k=0}^{js_T-1} A_{jk} + \sum_{k=0}^{\ell-1} A_{K_t k} + B_{\ell}, \qquad (5.3.52)
\end{aligned}
$$

where $A_{jk} = \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{s_T j+k}^{v}) - v_{s_T j+k}$ and $B_{\ell} = (t - (s_T K_t + \ell)) \left(\nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{s_T K_t+\ell}^{v}) - v_{s_T K_t+\ell}\right)$. Squaring on both sides, expanding the right-hand side and taking expectation, we get,

$$
\begin{aligned}
\mathbb{E}\Big[\Big(\int_{nT}^{t} \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u \rfloor}^{v}) - v_{\lfloor u \rfloor} du\Big)^2\Big] &= \sum_{k=\bar{L}}^{s_T-1} \mathbb{E}\big[|A_{1k}|^2\big] + \sum_{j=2}^{K_t-1} \sum_{k=0}^{js_T-1} \mathbb{E}\big[|A_{jk}|^2\big] + \sum_{k=0}^{\ell-1} \mathbb{E}\big[|A_{K_t k}|^2\big] + \mathbb{E}\big[|B_{\ell}|^2\big] \\
&+ 2 \sum_{\bar{L} \le k < k' \le s_T-1} \mathbb{E}\big[\langle A_{1k}, A_{1k'} \rangle\big] + 2 \sum_{\substack{(j,k),(j',k') \\ (j,k)<(j',k')}} \mathbb{E}\big[\langle A_{jk}, A_{j'k'} \rangle\big] \\
&+ 2 \sum_{0 \le k < k' \le \ell-1} \mathbb{E}\big[\langle A_{K_t k}, A_{K_t k'} \rangle\big] + 2 \sum_{k=\bar{L}}^{s_T-1} \sum_{j=2}^{K_t-1} \sum_{k'=0}^{js_T-1} \mathbb{E}\big[\langle A_{1k}, A_{jk'} \rangle\big] \\
&+ 2 \sum_{k=\bar{L}}^{s_T-1} \sum_{k'=0}^{\ell-1} \mathbb{E}\big[\langle A_{1k}, A_{K_t k'} \rangle\big] + 2 \sum_{k=\bar{L}}^{s_T-1} \mathbb{E}\big[\langle A_{1k}, B_{\ell} \rangle\big] \\
&+ 2 \sum_{j=2}^{K_t-1} \sum_{k=0}^{js_T-1} \sum_{k'=0}^{\ell-1} \mathbb{E}\big[\langle A_{jk}, A_{K_t k'} \rangle\big] + 2 \sum_{j=2}^{K_t-1} \sum_{k=0}^{js_T-1} \mathbb{E}\big[\langle A_{jk}, B_{\ell} \rangle\big] \\
&+ 2 \sum_{k=0}^{\ell-1} \mathbb{E}\big[\langle A_{K_t k}, B_{\ell} \rangle\big]. \qquad (5.3.53)
\end{aligned}
$$

We now observe that, for index $(j,k) \neq (j',k')$, we have,

$$\begin{aligned}\mathbb{E}[\langle A_{jk}, A_{j'k'}\rangle] &= \mathbb{E}[\langle \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j+k}) - v_{s_T j+k}, \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'+k'}) - v_{s_T j'+k'}\rangle] \\ &= \mathbb{E}[\langle \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j+k}) \\ &\qquad - \nabla f(\xi^{v}_{s_T j+k}, Y^{\mathbf{x}}_{s_T j+k+1}) - \nabla f(\xi^{v}_{s_T j}, Y^{\mathbf{x}}_{s_T j+k+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j}), \\ &\qquad \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'+k'}) - \nabla f(\xi^{v}_{s_T j'+k'}, Y^{\mathbf{x}}_{s_T j'+k'+1}) \\ &\qquad - \nabla f(\xi^{v}_{s_T i}, Y^{\mathbf{x}}_{s_T j'+k'+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'})\rangle]. \end{aligned} \tag{5.3.54}$$

Without loss of generality, assume, $s_T j+k < s_T j'+k'$ and let $\mathcal{H}_{sTj'+k'} = \sigma\left(\mathcal{G}_{sTj'+k'} \cup \widetilde{\mathcal{F}}_{sTj'+k'}\right)$, then we get,

$$\begin{aligned}\mathbb{E}[\langle A_{jk}, A_{j'k'}\rangle] &= \mathbb{E}[\langle \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j+k}) - \nabla f(\xi^{v}_{s_T j+k}, Y^{\mathbf{x}}_{s_T j+k+1}) - \nabla f(\xi^{v}_{s_T j}, Y^{\mathbf{x}}_{s_T j+k+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j}), \\ &\qquad \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'+k'}) - \nabla f(\xi^{v}_{s_T j'+k'}, Y^{\mathbf{x}}_{s_T j'+k'+1}) - \nabla f(\xi^{v}_{s_T i}, Y^{\mathbf{x}}_{s_T j'+k'+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'})\rangle] \\ &= \mathbb{E}[\mathbb{E}[\langle \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j+k}) - \nabla f(\xi^{v}_{s_T j+k}, Y^{\mathbf{x}}_{s_T j+k+1}) - \nabla f(\xi^{v}_{s_T j}, Y^{\mathbf{x}}_{s_T j+k+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j}), \\ &\qquad \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'+k'}) - \nabla f(\xi^{v}_{s_T j'+k'}, Y^{\mathbf{x}}_{s_T j'+k'+1}) - \nabla f(\xi^{v}_{s_T j'}, Y^{\mathbf{x}}_{s_T j'+k'+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'})\rangle | \mathcal{H}_{s_T j'+k'}] \\ &= \mathbb{E}[\langle \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j+k}) - \nabla f(\xi^{v}_{s_T j+k}, Y^{\mathbf{x}}_{s_T j+k+1}) - \nabla f(\xi^{v}_{s_T j}, Y^{\mathbf{x}}_{s_T j+k+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j}), \\ &\qquad \mathbb{E}[\nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'+k'}) - \nabla f(\xi^{v}_{s_T j'+k'}, Y^{\mathbf{x}}_{s_T j'+k'+1}) - \nabla f(\xi^{v}_{s_T j}, Y^{\mathbf{x}}_{s_T j'+k_1+1}) + \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{s_T j'}) | \mathcal{H}_{s_T j'+k'}]\rangle \\ &= 0. \end{aligned}$$

By a similar argument, we can prove that the expected value of all the other inner products with non-matching indices is equal to zero. Therefore, we have,

$$\begin{aligned} h^2 \mathbb{E}\left[\left|\left(\int_{mT}^{t} \nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{\lfloor u \rfloor}) - v_{\lfloor u \rfloor}\right) du\right|^2\right] = h^2 \Bigg(&\sum_{k=\bar{L}}^{s_T-1} \mathbb{E}\left[|A_{1k}|^2\right] + \sum_{j=2}^{K_t-1} \sum_{k=0}^{js_T-1} \mathbb{E}\left[|A_{jk}|^2\right] \\ &+ \sum_{k=0}^{\ell-1} \mathbb{E}\left[|A_{K_t k}|^2\right] + \mathbb{E}\left[|B_\ell|^2\right] \Bigg). \end{aligned} \tag{5.3.55}$$

Now for any $k \in \mathbb{N} \cap [nT, (n+1)T]$, we observe,

$$\begin{aligned}\mathbb{E}\left[|\nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{\hat{v}}_{k}) - v_k|^2\right] &\leq \mathbb{E}[|\nabla f(\xi^{v}_{k}, Y^{\mathbf{x}}_{k+1}) - \nabla f(\xi^{v}_{s}, Y^{\mathbf{x}}_{k+1})|^2] \\ &\leq 2L^2(\mathbb{E}[|\xi_k|^2] + \mathbb{E}[|\xi_s|^2]) \\ &\leq 2L^2\left((1-ah)^k \mathbb{E}[|\theta_0|^2] + (1-ah)^s \mathbb{E}[|\theta_0|^2]\right) + 2L^2 \mathring{B}\left(1 + \frac{1}{a}\right). \end{aligned}$$

Therefore,

$$h^2\left(\mathbb{E}\left[|\nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{\hat{v}}_k)-v_k|^2\right]\right) \le h^2\left(2L^2\left((1-ah)^k\mathbb{E}[|\theta_0|^2]+(1-ah)^s\mathbb{E}[|\theta_0|^2]\right)+2L^2\mathring{B}\left(1+\frac{1}{a}\right)\right)$$
$$\le h\left(a(1-ah)^k\mathbb{E}[|\theta_0|^2]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right)$$
$$\le h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right). \tag{5.3.56}$$

Consequently,

$$h^2\mathbb{E}\left[\left|\left(\int_{mT}^{t}\nabla F^{\varepsilon}_{\mathbf{x}}(\xi^{v}_{\lfloor u\rfloor})-v_{\lfloor u\rfloor}\right)du\right|^2\right] \le \sum_{k=L}^{s_T-1}h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right)$$
$$+\sum_{j=2}^{K_t-1}\sum_{k=0}^{js_T-1}h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right)$$
$$+\sum_{k=0}^{\ell-1}h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right)$$
$$+h(t-(s_TK_t+\ell))\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right)$$
$$\le h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)]+\mathring{B}\,(a+1)+a\kappa_{0,2}\right). \quad \square \tag{5.3.57}$$

5.3.3 Proof of Theorem 5.3.1

Now that we have established the moment estimates of all the underlying processes, we provide complete proof of our main theorem in this section. To this end, we recall the semi-metric $w_{1,p}$, which is crucial to the results. For any $p \ge 1$ and $\zeta_1, \zeta_2 \in \mathcal{P}_{\mathcal{V}_p}(\mathbb{R}^d)$, let,

$$w_{1,p}(\zeta_1,\zeta_2) \coloneqq \inf_{\zeta\in\Gamma(\zeta_1,\zeta_2)}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}[\min\{1,|\theta_1-\theta_2|\}](1+\mathcal{V}_p(\theta_1)+\mathcal{V}_p(\theta_2))\zeta(\mathrm{d}\theta_1\mathrm{d}\theta_2) \tag{5.3.58}$$

In [16], it has been established that the analysis of Theorem 5.3.1 relies on the contractivity of the SDE (5.3.7) in $w_{1,2}$ [16]. Accordingly, the following proposition provides the explicit statement of the contraction property.

Proposition 5.3.1. *Let $z'_t \in \mathbb{R}_+$ be the solution of (5.3.7) with initial condition $z'_0 = \theta'_0$ which is independent of $\mathcal{F}_\infty = \sigma\left(\bigcup_{t\ge0}\mathcal{F}_t\right)$ and satisfies $(\mathbb{E}[|\theta'_0|^2])^{1/2}$ is finite. Then under*

the assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4, we have,

$$w_{1,2}(\mathcal{L}(z_t), \mathcal{L}(z'_t)) \leq \mathfrak{c}_2 e^{-\mathfrak{c}_1 t} w_{1,2}(\mathcal{L}(\theta_0), \mathcal{L}(\theta'_0)),$$

where the constants $\mathfrak{c}_1$ *and* $\mathfrak{c}_2$ *are given are given as,*

$$\mathfrak{c}_1 = \min\{\bar{\phi}, C_{\mathcal{V},1}(p), 4C_{\mathcal{V},2}(p)\dot{\varepsilon}C_{\mathcal{V},1}(p)\}/2,$$

where the explicit expressions for $C_{\mathcal{V},1}(p)$ *and* $C_{\mathcal{V},2}(p)$ *can be found in Lemma 5.3.1 and* $\bar{\phi}$ *is given by,*

$$\bar{\phi} = \left(\sqrt{4\pi/L}\bar{b}\exp\left(\left(\bar{b}\sqrt{L}/2 + 2/\sqrt{L}\right)^2\right)\right)^{-1}.$$

Furthermore, any $\dot{\varepsilon}$ *can be chosen which satisfies the following inequality,*

$$\dot{\varepsilon} \leq 1 \wedge \left(8C_{\mathcal{V},2}(p)\sqrt{\pi}/L \int_0^{\tilde{b}} \exp\left(\left(s\sqrt{L}/2 + 2/\sqrt{L}\right)^2\right) ds\right)^{-1},$$

where

$$\tilde{b} = \sqrt{2C_{\mathcal{V},2}(p)/C_{\mathcal{V},1}(p) - 1}, \quad \bar{b} = \sqrt{4C_{\mathcal{V},2}(p)(1 + C_{\mathcal{V},1}(p))/C_{\mathcal{V},1}(p) - 1}.$$

The constant $\mathfrak{c}_2$ *is given as, the ratio* C_{11}/C_{10}*, where* C_{11}, C_{10} *are given explicitly in [16] Lemma 3.26.*

Proof. Refer [16, Lemma 3.26] □

Below, we present a series of results that facilitate the estimation of ER. We would like to point out that the proofs of these results follow directly from the analysis in [16, 76]. However, for the purpose of exposition, we provide the outline of the proofs.

Lemma 5.3.6. *Consider* $n \in \mathbb{N}$ *and assume that assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 are satisfied. Then for any* $0 < h \leq h^*$*, where* h^* *is specified in* (5.3.4)*, and* $t \in (nT, (n+1)T]$*, it follows that,*

$$\mathcal{W}_1(\mathcal{L}(\xi_t^v), \Xi_{t,v}^{n,h})) \leq \sqrt{h\left(C_1 e^{-an/2}\mathbb{E}[\mathcal{V}_2(\theta_0)] + C_2\right)} \tag{5.3.59}$$

where C_1 *and* C_2 *are given in equation* (5.3.64)*.*

Proof. In order to prove this lemma, we use the inequality $\mathcal{W}_1 \leq \mathcal{W}_2$. Now observe that,

$$\mathbb{E}[|\xi_t^v - \Xi_{t,v}^{n,h}|^2] \leq 2h^2\mathbb{E}\left[\left|\left(\int_{nT}^t \nabla F_{\mathbf{x}}^{\varepsilon}(\xi_{\lfloor u \rfloor}^v) - v_{\lfloor u \rfloor}\right) du\right|^2\right] + 4hL^2\int_{nT}^t \mathbb{E}[|\xi_u^v - \xi_{\lfloor u \rfloor}^v|^2]du + 4hL^2\int_{nT}^y \mathbb{E}[|\xi_u^v - \Xi_{u,v}^{n,h}|^2]du. \tag{5.3.60}$$

Now as a consequence of Lemma 5.3.4 and 5.3.5, we have,

$$\mathbb{E}[|\xi_t^v - \Xi_{t,v}^{n,h}|^2] \leq 2h\left(ae^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)] + \mathring{B}\,(a+1) + a\kappa_{0,2}\right) + 4hL^2\left(ae^{-ah\lfloor t \rfloor}\mathbb{E}[\mathcal{V}_2(\theta_0)] + B_1\right) + 4hL^2\int_{mT}^t \mathbb{E}[|\xi_u^v - \Xi_{u,v}^{n,h}|^2]du \tag{5.3.61}$$

Now, by Gronwall's inequality, we have,

$$\mathbb{E}[|\xi_t^v - \Xi_{t,v}^{n,h}|^2] \leq 4he^{4L^2}\left(a(1+L^2)e^{-ahnT}\mathbb{E}[\mathcal{V}_2(\theta_0)] + \mathring{B}\,(a+1) + a\kappa_{0,2} + L^2B_1\right). \tag{5.3.62}$$

Since, $0 < h \leq 1$, we have $1/2 < nT \leq 1$, therefore,

$$\mathcal{W}_2^2(\mathcal{L}(\xi_t^v), \mathcal{L}(\Xi_{t,v}^{n,h})) \leq \mathbb{E}[|\xi_t^v - \Xi_{t,v}^{n,h}|^2] \leq h\left(C_1e^{-an/2}\mathbb{E}[\mathcal{V}_2(\theta_0)] + C_2\right), \tag{5.3.63}$$

where,

$$C_1 = 4e^{4L^2}a(1+L^2) \text{ and } C_2 = 4e^{4L^2}\left(\mathring{B}\,(a+1) + a\kappa_{0,2} + L^2B_1\right), \tag{5.3.64}$$

with $\mathring{B}$ and B_1 given in Lemma 5.3.2 and Lemma 5.3.4 respectively. □

Lemma 5.3.7. *Consider $n \in \mathbb{N}$ and assume that assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 are satisfied. Then for any $0 < h \leq h^*$, where h^* is specified in* (5.3.4)*, and $t \in (nT, (n+1)T]$, it follows that,*

$$\mathcal{W}_1(\mathcal{L}(\Xi_{t,v}^{n,h}), \mathcal{L}(\tilde{z}_t)) \leq \sqrt{h}\left(e^{-\min\{\mathfrak{c}_1, a/2\}n/2}C_3\mathbb{E}[\mathcal{V}_4(\theta_0)] + C_4\right) \tag{5.3.65}$$

where C_3 and C_4 are given in (5.3.69)*.*

Proof (outline). The proof of this lemma follows the same argument as in [76]. The reader may refer to [76, Lemma 5] for the detailed proof. Below, we provide an outline of the purpose of the exposition. We begin the proof by observing that, $\tilde{z}_t = \Xi_{t,v}^{0,h}$. Now

by the contraction argument in [16, Proposition 3.14], we have,

$$\begin{aligned}\mathcal{W}_1(\mathcal{L}(\Xi^{n,h}_{t,v}),\mathcal{L}(\tilde{z}_t)) &\leq \sum_{j=1}^{n}\mathcal{W}_1(\mathcal{L}(\Xi^{j,h}_{t,v}),\mathcal{L}(\Xi^{j-1,h}_{t,v}))\\ &\leq \sum_{j=1}^{n} w_{1,2}(\mathcal{L}(\Xi_t^{jT,\xi^v_{jT},h}),\mathcal{L}(\Xi_t^{jT,\Xi^{j-1,h}_{jT,v},h}))\\ &\leq \sum_{j=1}^{n}\mathfrak{c}_2 e^{-\mathfrak{c}_1(n-j)} w_{1,2}(\mathcal{L}(\xi^v_{jT}),\mathcal{L}(\Xi^{j-1,h}_{jT,v})),\end{aligned} \tag{5.3.66}$$

where the second inequality is by the fact that, $\mathcal{W}_1(\zeta_1,\zeta_2)\leq w_{1,2}(\zeta_1,\zeta_2)$, and the third inequality is because of Proposition 5.3.1. Now by the definition of $w_{1,2}(\cdot,\cdot)$ and the Cauchy-Schwarz inequality, we have,

$$w_{1,2}(\mathcal{L}(\xi^v_{jT}),\mathcal{L}(\Xi^{j-1,h}_{jT,v}))\leq \mathcal{W}_2(\mathcal{L}(\xi^v_{jT}),\mathcal{L}(\Xi^{j-1,h}_{jT,v}))\left[1+\left(\mathbb{E}[\mathcal{V}_4(\xi^v_{jT})]\right)^{1/2}+\left(\mathbb{E}[\mathcal{V}_4(\Xi^{j-1,h}_{jT,v})]\right)^{1/2}\right]. \tag{5.3.67}$$

Now using the inequality $2ab\leq \nu a^2+\nu^{-1}b^2$ with $\nu=1/\sqrt{h}$, Lemma 5.3.3, corollary 5.3.1 and (5.3.63), we can prove that,

$$\begin{aligned}\mathcal{W}_1(\mathcal{L}(\Xi^{n,h}_{t,v}),\mathcal{L}(\tilde{z}_t)) &\leq \mathfrak{c}_2\sqrt{h}\left(1+\frac{2}{\min\{\mathfrak{c}_1,a/2\}}\right)e^{-\min\{\mathfrak{c}_1,a/2\}(n)/2}\mathbb{E}[\mathcal{V}_4(\theta_0)]\left(e^{\min\{\mathfrak{c}_1,a/2\}}C_1+12\right)\\ &\quad+\frac{\mathfrak{c}_2\sqrt{h}}{1-e^{\mathfrak{c}_1}}\left(15+12\mathring{C}\left(1+\frac{1}{a}\right)+9\mathfrak{v}_5(K_{\mathcal{V}}(4))+C_2\right)\\ &\leq \sqrt{h}\left(e^{-\min\{\mathfrak{c}_1,a/2\}n/2}C_3\mathbb{E}[\mathcal{V}_4(\theta_0)]+C_4\right)\end{aligned} \tag{5.3.68}$$

where,

$$\begin{aligned}C_3&=\mathfrak{c}_2\left(1+\frac{2}{\min\{\mathfrak{c}_1,a/2\}}\right)\left(e^{\min\{\mathfrak{c}_1,a/2\}}C_1+12\right)\\ C_4&=\frac{\mathfrak{c}_2}{1-e^{\mathfrak{c}_1}}\left(15+12\mathring{C}\left(1+\frac{1}{a}\right)+9\mathfrak{v}_5(K_{\mathcal{V}}(4))+C_2\right),\end{aligned} \tag{5.3.69}$$

with $\mathfrak{c}_2,\mathfrak{c}_1$ given in Proposition 5.3.1, C_1,C_2 given in Lemma 5.3.6, $\mathring{C}$ given in Lemma 5.3.2, and $K_{\mathcal{V}}(4)$ given in Lemma 5.3.1. □

Lemma 5.3.8. *Consider $n\in\mathbb{N}$ and assume that assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 are satisfied. Then for any $0<h\leq h^*$, where h^* is specified in* (5.3.4)*, and*

$t \in (nT, (n+1)T]$, *it follows that,*

$$\mathcal{W}_2(\mathcal{L}(\Xi^{n,h}_{t,v}), \mathcal{L}(\tilde{z}_t)) \leq h^{1/4}\left(e^{-\min\{\mathfrak{c}_1, a/2\}n/4} C_5 (\mathbb{E}[\mathcal{V}_4(\theta_0)])^{1/2} + C_6\right), \tag{5.3.70}$$

where C_5 *and* C_6 *are given in* (5.3.71).

Proof (outline). The proof of this lemma follows the same line of argument as given above while applying $\mathcal{W}_2(\zeta_1, \zeta_2) \leq \sqrt{2w_{1,2}(\zeta_1, \zeta_2)}$. The reader may refer to [76, corollary 6] for the complete argument. The constants C_5 and C_6 are given as,

$$\begin{aligned} C_5 &= \sqrt{2\mathfrak{c}_2}\left(2\sqrt{2} + e^{\min\{\mathfrak{c}_1/2.a/4\}/2}\sqrt{C_1}\right)\left(1 + \frac{2}{\min\{\mathfrak{c}_1/2, a/4\}}\right), \\ C_6 &= \frac{\sqrt{2\mathfrak{c}_2}}{1 - e^{-\mathfrak{c}_1/2}}\left(1 + 2\sqrt{2} + 2\sqrt{2\mathring{C}(1 + a^{-1})} + \sqrt{3\mathfrak{v}_5(K_{\mathcal{V}}(4))} + \sqrt{C_2}\right), \end{aligned} \tag{5.3.71}$$

where $\mathfrak{c}_2, \mathfrak{c}_1$ are given in Proposition 5.3.1, C_1, C_2 in Lemma 5.3.6, $\mathring{C}$ in Lemma 5.3.2 and $K_{\mathcal{V}}(4)$ in Lemma 5.3.1. □

Lemma 5.3.9. *Consider* $n \in \mathbb{N}$ *and assume that assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 are satisfied. Then for any* $0 < h \leq h^*$, *where* h^* *is specified in* (5.3.4), *and* $t \in (nT, (n+1)T]$, *it follows that,*

$$\mathcal{W}_1(\mathcal{L}(\xi^v_t), \pi^{\mathbf{x},\varepsilon}_\delta) \leq \bar{C}_1\sqrt{h} + \bar{C}_2 e^{-\bar{C}_3 n}, \tag{5.3.72}$$

where $\bar{C}_2$ *and* $\bar{C}_3$ *are given in* (5.3.76) *and* $\bar{C}_1$ *is given is* (5.3.64).

Proof (outline). To begin with, by triangle inequality, Lemma 5.3.6 and Lemma 5.3.7, we have,

$$\begin{aligned} \mathcal{W}_1(\mathcal{L}(\xi^v_t), \pi^{\mathbf{x},\varepsilon}_\delta) &\leq \mathcal{W}_1(\mathcal{L}(\xi^v_t), \mathcal{L}(\Xi^{n,h}_{t,v})) + \mathcal{W}_1(\mathcal{L}(\Xi^{n,h}_{t,v}), \mathcal{L}(\xi_t)) + \mathcal{W}_1(\mathcal{L}(\tilde{z}_t), \pi^{\mathbf{x},\varepsilon}_\delta) \\ &\leq \bar{C}_1\sqrt{h}\left(e^{-\min\{a/2, \mathfrak{c}_1\}n/2}\mathbb{E}[\mathcal{V}_4(\theta_0)] + 1\right) + \mathcal{W}_1(\mathcal{L}(\tilde{z}_t), \pi^{\mathbf{x},\varepsilon}_\delta) \end{aligned} \tag{5.3.73}$$

where $\bar{C}_1 = \max\{\sqrt{C_1} + C_3, \sqrt{C_2} + C_4\}$. Further, we know,

$$\mathcal{W}_1(\mathcal{L}(\tilde{z}_t), \pi^{\mathbf{x},\varepsilon}_\delta) \leq w_{1,2}(\mathcal{L}(\tilde{z}_t), \pi^{\mathbf{x},\varepsilon}_\delta) \leq \mathfrak{c}_2 e^{-\mathfrak{c}_1 ht} w_{1,2}(\theta_0, \pi^{\mathbf{x},\varepsilon}_\delta).$$

Therefore,

$$\begin{aligned}\mathcal{W}_1(\mathcal{L}(\xi_t^v), \pi_\delta^{\mathbf{x},\varepsilon}) &\leq \bar{C}_1\sqrt{h}\left(e^{-\min\{a/2,\mathfrak{c}_1\}n/2}\mathbb{E}[\mathcal{V}_4(\theta_0)]+1\right) + \mathfrak{c}_2 e^{-\mathfrak{c}_1 ht} w_{1,2}(\theta_0, \pi_\delta^{\mathbf{x},\varepsilon}) \\ &\leq \bar{C}_1\sqrt{h}\left(e^{-\min\{a/2,\mathfrak{c}_1\}n/2}\mathbb{E}[\mathcal{V}_4(\theta_0)]+1\right) \qquad (5.3.74)\\ &\quad + \mathfrak{c}_2 e^{-\mathfrak{c}_1 ht}\left(1+\mathbb{E}[\mathcal{V}_2(\theta_0)] + \int_{\mathbb{R}^d}\mathcal{V}_2(\theta)\pi_\delta^{\mathbf{x},\varepsilon}d(\theta)\right) \\ &\leq \bar{C}_1\sqrt{h}\left(e^{-\min\{a/2,\mathfrak{c}_1\}n/2}\mathbb{E}[\mathcal{V}_4(\theta_0)]+1\right) \\ &\quad + \mathfrak{c}_2 e^{-\min\{a/2,\mathfrak{c}_1\}n/2}\left(1+\mathbb{E}[\mathcal{V}_2(\theta_0)] + \int_{\mathbb{R}^d}\mathcal{V}_2(\theta)\pi_\delta^{\mathbf{x},\varepsilon}d(\theta)\right) \\ &\leq \bar{C}_1\sqrt{h} + \bar{C}_2 e^{-\bar{C}_3 n} \qquad (5.3.75)\end{aligned}$$

where,

$$\begin{aligned}\bar{C}_2 &= \left(\mathbb{E}[\mathcal{V}_4(\theta_0)]+1+\mathbb{E}[\mathcal{V}_2(\theta_0)] + \int_{\mathbb{R}^d}\mathcal{V}_2(\theta)\pi_\delta^{\mathbf{x},\varepsilon}d\theta\right), \\ \bar{C}_3 &= \mathfrak{c}_2\min\{a/2,\mathfrak{c}_1\}/2,\end{aligned} \qquad (5.3.76)$$

with $\mathfrak{c}_2, \mathfrak{c}_1$ being given in Proposition 5.3.1 and $\bar{C}_1$ is given in (5.3.64). □

Lemma 5.3.10. *Consider $n \in \mathbb{N}$ and assume that assumptions 5.3.1, 5.3.2, 5.3.3, and 5.3.4 are satisfied. Then for any $0 < h \leq h^*$, where h^* is specified in (5.3.4), and $t \in (nT, (n+1)T]$, it follows that,*

$$\mathcal{W}_2(\mathcal{L}(\xi_t^v), \pi_\delta^{\mathbf{x},\varepsilon}) \leq \bar{C}_4 h^{1/4} + \bar{C}_5 e^{-\bar{C}_6 n}, \qquad (5.3.77)$$

where $\bar{C}_4, \bar{C}_5$ and $\bar{C}_6$ are given in (5.3.78).

Proof (outline). As before, by triangle inequality, Lemma 5.3.8 and (5.3.63) we have,

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(\xi_t^v), \pi_\delta^{\mathbf{x},\varepsilon}) &\leq \mathcal{W}_2(\mathcal{L}(\xi_t^v), \mathcal{L}(\Xi_{t,v}^{n,h})) + \mathcal{W}_2(\mathcal{L}(\Xi_{t,v}^{n,h}), \mathcal{L}(\xi_t)) + \mathcal{W}_2(\mathcal{L}(\xi_t), \pi_\delta^{\mathbf{x},\varepsilon}) \\ &\leq \sqrt{h(C_1 e^{-an/2}\mathbb{E}[\mathcal{V}_2(\theta_0)]+C_2)} + h^{1/4}\left(e^{-\min\{\mathfrak{c}_1,a/2\}n/4}C_5(\mathbb{E}[\mathcal{V}_4(\theta_0)])^{1/2}+C_6\right) \\ &\quad + \sqrt{2w_{1,2}(\mathcal{L}(\xi_t), \pi_\delta^{\mathbf{x},\varepsilon})} \\ &\leq \sqrt{h(C_1 e^{-an/2}\mathbb{E}[\mathcal{V}_2(\theta_0)]+C_2)} + h^{1/4}\left(e^{-\min\{\mathfrak{c}_1,a/2\}n/4}C_5(\mathbb{E}[\mathcal{V}_4(\theta_0)])^{1/2}+C_6\right) \\ &\quad + \sqrt{2\mathfrak{c}_2 e^{-\mathfrak{c}_1 n}\left(1+\mathbb{E}[\mathcal{V}_2(\theta_0)] + \int_{\mathbb{R}^d}\mathcal{V}_2(\theta)\pi_\delta^{\mathbf{x},\varepsilon}d(\theta)\right)} \\ &\leq \bar{C}_4 h^{1/4} + \bar{C}_5 e^{-\bar{C}_6 n}\end{aligned}$$

where,

$$\bar{C}_5 = \left(\sqrt{C_1}(\mathbb{E}[\mathcal{V}_2(\theta_0)])^{1/2} + C_5(\mathbb{E}[\mathcal{V}_2(\theta_0)])^{1/4} + \sqrt{2\mathfrak{c}_2 \left(1 + \mathbb{E}[\mathcal{V}_2(\theta_0)] + \int_{\mathbb{R}^d} \mathcal{V}_2(\theta)\pi_\delta^{\mathbf{x},\varepsilon} d(\theta)\right)} \right),$$
$$\bar{C}_4 = \sqrt{C_2} + C_6,$$
$$\bar{C}_6 = \frac{\min\{\mathfrak{c}_1, a/2\}}{4}, \tag{5.3.78}$$

with $\mathfrak{c}_2, \mathfrak{c}_1$ given in Proposition 5.3.1 , C_1, C_2 given in (5.3.64) and C_5, C_6 given in (5.3.71). □

We now present the proof of the main theorem.

Proof Theorem 5.3.1. To begin with, we observe that,

$$\begin{aligned}
\mathbb{E}[F(\theta_t^v)] - \inf_{\theta\in\mathbb{R}^d} F(\theta) &= \mathbb{E}[F(\theta_t^v)] - \mathbb{E}[F^\varepsilon(\theta_t^v)] + \mathbb{E}[F^\varepsilon(\theta_t^v)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) + \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) - \inf_{\theta\in\mathbb{R}} F(\theta) \\
&\leq \left|\mathbb{E}[F(\theta_t^v)] - \mathbb{E}[F^\varepsilon(\theta_t^v)]\right| + \mathbb{E}[F^\varepsilon(\theta_t^v)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) + \left|\inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) - \inf_{\theta\in\mathbb{R}} F(\theta)\right| \\
&\leq \mathbb{E}[|F(\theta_t^v) - F^\varepsilon(\theta_t^v)|] + \mathbb{E}[F^\varepsilon(\theta_t^v)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) + \sup_{\theta\in\mathbb{R}^d}|F(\theta) - F^\varepsilon(\theta)| \\
&\leq 2\varepsilon + \mathbb{E}[F^\varepsilon(\theta_t^v)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta).
\end{aligned} \tag{5.3.79}$$

We now further decompose the second term in the above inequality,

$$\mathbb{E}[F^\varepsilon(\theta_t^v)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) = \mathbb{E}[F^\varepsilon(\theta_t^v)] - \mathbb{E}[F^\varepsilon(z_\infty)] + \mathbb{E}[F^\varepsilon(z_\infty)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta), \tag{5.3.80}$$

where $z_\infty \sim \pi_\delta^{\mathbf{x},\varepsilon}$. By Lemma 6 in [71], we have that,

$$\mathbb{E}[F^\varepsilon(\theta_t^v)] - \mathbb{E}[F^\varepsilon(z_\infty)] \leq \int_{(\mathbb{R}^d)^N} \mathbb{P}^{\otimes n} d\mathbf{x} \left((\bar{\sigma}L + B)\mathcal{W}_2(\mathcal{L}(\theta_t^v), \pi_\delta^{\mathbf{x},\varepsilon})\right), \tag{5.3.81}$$

where $\bar{\sigma} = \max\{\mathbb{E}[|\theta_t^v|^2], \mathbb{E}[|z_\infty|^2]\} \leq (2b + 2d\delta^{-1} + \mathring{B}(1 + a^{-1}))$. Moreover, by Lemma 5.3.10, we have,

$$\mathcal{W}_2(\mathcal{L}(\theta_t^v), \pi_\delta^{\mathbf{x},\varepsilon}) \leq \bar{C}_4 h^{1/4} + \bar{C}_5 e^{-\bar{C}_6 hn}, \tag{5.3.82}$$

therefore,

$$\mathbb{E}[F^\varepsilon(\theta_t^v)] - \mathbb{E}[F^\varepsilon(z_\infty)] \leq \bar{C}_{1,7} h^{1/4} + \bar{C}_{2,7} e^{-\bar{C}_6 hn}, \tag{5.3.83}$$

where,

$$\bar{C}_{1,7} = \bar{C}_4(2b + 2d\delta^{-1} + \mathring{B}(1 + a^{-1})) \text{ and } \bar{C}_{2,7} = \bar{C}_5(2b + 2d\delta^{-1} + \mathring{B}(1 + a^{-1})). \quad (5.3.84)$$

Finally, by Proposition 11 and 12 in [71], we have,

$$\begin{aligned}\mathbb{E}[F^\varepsilon(z_\infty)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) &\leq \mathbb{E}[F^\varepsilon(z_\infty)] - \mathbb{E}[F^\varepsilon_{\mathbf{x}}(z_\infty)] + \mathbb{E}[F^\varepsilon_{\mathbf{x}}(z_\infty)] - \inf_{\theta\in\mathbb{R}} F^\varepsilon(\theta) \\ &\leq \frac{\bar{C}_8}{N} + \frac{d}{2\delta}\log\left(\frac{eL}{a}\left(\frac{b\delta}{d} + 1\right)\right), \end{aligned} \quad (5.3.85)$$

for some constant $\bar{C}_8 > 0$. Collating everything together, we get the final result. □

5.4 Numerical Illustration

In this section, we complement our theoretical results with numerical examples where we conduct numerical experiments to evaluate the effectiveness of the SVRG-LD method discussed. We address the minimum-CVaR portfolio optimization problem (5.1.14), employing both the conventional SGLD approach, where $v_n = \nabla f^\varepsilon(\theta^v_n, Y^{\mathbf{x}}_{n+1})$, and the SVRG-LD methods developed within this chapter, and subsequently compare their outcomes.

As discussed in Section 5.1, we assume the availability of N samples from the $\mathcal{L}(X)$ *i.e.*, $\mathbf{x} = \{X_1, \ldots, X_N\}$ and consequently attempt to solve the optimization problem given as,

$$\widetilde{\mathrm{CVaR}}^{\mathbf{x}}_\nu(\Phi) \coloneqq \inf_{\theta\in\mathbb{R}}\left\{\vartheta + \frac{1}{1-\nu}\frac{1}{N}\sum_{k=1}^{N}(\Phi(w, X_k) - \vartheta)_+ + \frac{\eta}{2}|\theta|^2\right\}. \quad (5.4.1)$$

where $(x)_+$ represents the $\max\{x, 0\}$. From the above equation,

$$f(\theta, X) = \vartheta + \frac{1}{1-\nu}(\Phi(w, X) - \vartheta)_+ + \frac{\eta}{2}|\theta|^2 \quad (5.4.2)$$

Further, we define $\Phi(w, X)$ as follows,

$$\Phi(w, X_k) = \sum_{j=1}^{d}\frac{e^{w_j}X^j_k}{\sum_{k=1}^{d} e^{w_k}}, \quad (5.4.3)$$

and therefore, $g_i(w) = \dfrac{e^{w_i}}{\sum_{j=0}^{d} e^{w_j}}$. Moreover, in order to deal with the discontinuous gradient of $f(\theta, X)$, we consider a smooth approximation given as,

$$f^{\varepsilon}(\theta, X) = \vartheta + \frac{1}{1-\nu} S^{\varepsilon}(\phi(w, X) - \vartheta) + \frac{\eta}{2}|\theta|^2, \tag{5.4.4}$$

where,

$$S^{\varepsilon}(x) := \varepsilon \ln\left(1 + e^{\frac{1}{\varepsilon}x}\right), \quad \text{for some } \varepsilon > 0. \tag{5.4.5}$$

Finally, in order to perform our numerical experimentation, we consider a portfolio of two assets X_1 and X_2, wherein in the first case, we consider $X_1 \sim \mathcal{N}(1, 4)$ and $X_2 \sim \mathcal{N}(0, 1)$. On the other hand the second case, we consider $X_1 \sim t_{d.f.}$ $X_2 \sim \mathcal{N}(0, 1)$ where *d.f.* denotes the degree of freedom and is equal to 1000 for our simulation. We consider, for our implementation, the sample size $N = 1000$, $\delta = 10^8$, $\eta = 10^{-8}$ and the confidence interval $\nu = 0.95$. We run our algorithm for $h = \{0.02, 0.005, 0.0025\}$ and for $50,000$ iterations where we consider the initial epoch length $m_0 = 10$. In order to estimate the ER (Excess Risk), we run 1000 simulations and approximate ER as,

$$\text{ER} \approx \frac{1}{1000} \sum_{j=1}^{1000} F(\theta_{50000}^{v,j}) - \text{CVaR}_{\nu}(\Phi), \tag{5.4.6}$$

where,

$$F(\theta) = \vartheta + \frac{1}{1-\nu} \frac{1}{N} \sum_{k=1}^{N} (\Phi(w, X_k) - \vartheta)_+ + \frac{\eta}{2}|\theta|^2.$$

Figures 5.1, 5.2, and 5.3 illustrate the trajectories of the weights w_1, w_2 and parameter ϑ for $X_1 = \mathcal{N}(1, 4)$ and $X_2 = \mathcal{N}(0, 1)$ concerning $h \in \{0.02, 0.005, 0.0025\}$. Meanwhile, Figures 5.4, 5.5, and 5.6 demonstrate the convergence of $\widetilde{\text{CVaR}}_{\nu}^{\mathbf{x}}(\Phi)$ under varying values of h. Likewise, the trajectories of weights w_1, w_2 and parameter ϑ for $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0, 1)$ in relation to $h \in \{0.02, 0.005, 0.0025\}$ are depicted in Figures 5.7, 5.8, and 5.9, with Figures 5.10, 5.11, and 5.12 showing the convergence of $\widetilde{\text{CVaR}}_{\nu}^{\mathbf{x}}(\Phi)$ for different values of h. Finally, Figure 5.13 details the rate of descent of ER as the step-size h diminishes.

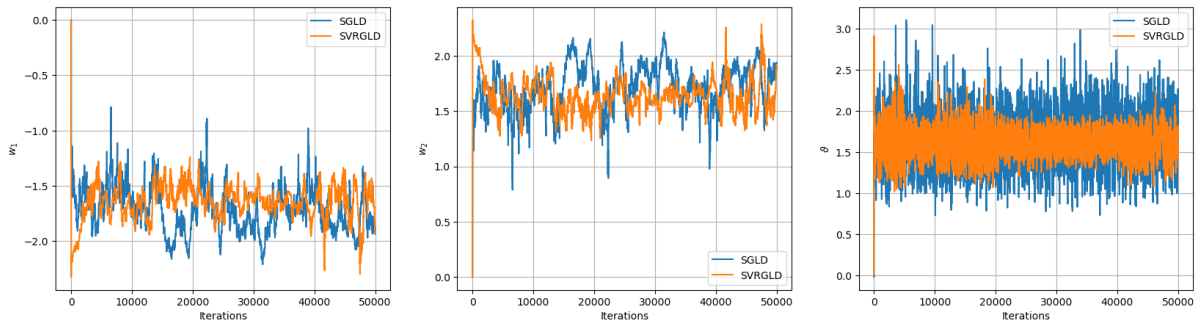


Figure 5.1: Trajectories of w_1, w_2 and ϑ for SGLD and SVRG-LD with step size $h = 0.02$, where $X_1 = \mathcal{N}(1, 4)$ and $X_2 = \mathcal{N}(0, 1)$. At this larger step size, both methods exhibit higher variance, with SGLD showing more pronounced fluctuations whereas, SVRG-LD maintains relatively smoother paths.

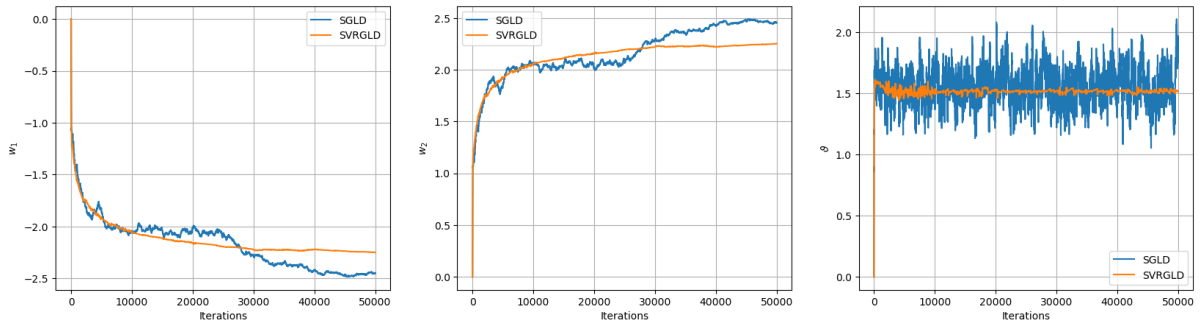


Figure 5.2: Trajectories of w_1, w_2 and ϑ for SGLD and SVRG-LD with step size $h = 0.005$, where $X_1 = \mathcal{N}(1, 4)$ and $X_2 = \mathcal{N}(0, 1)$. Both algorithms converge toward similar parameter values, but SVRG-LD demonstrates significantly reduced variance and smoother convergence.

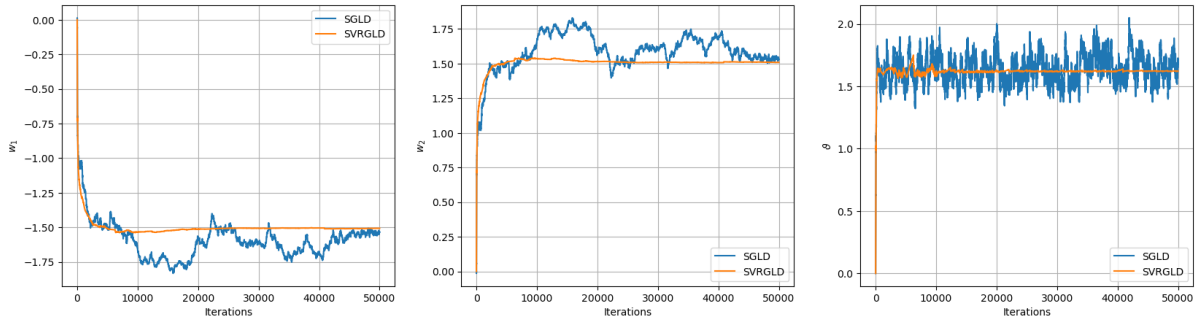


Figure 5.3: Trajectories of w_1, w_2 and ϑ for SGLD and SVRG-LD with step size $h = 0.0025$, where $X_1 = \mathcal{N}(1, 4)$ and $X_2 = \mathcal{N}(0, 1)$. Both algorithms converge toward similar parameter values, but SVRG-LD demonstrates significantly reduced variance and smoother convergence.

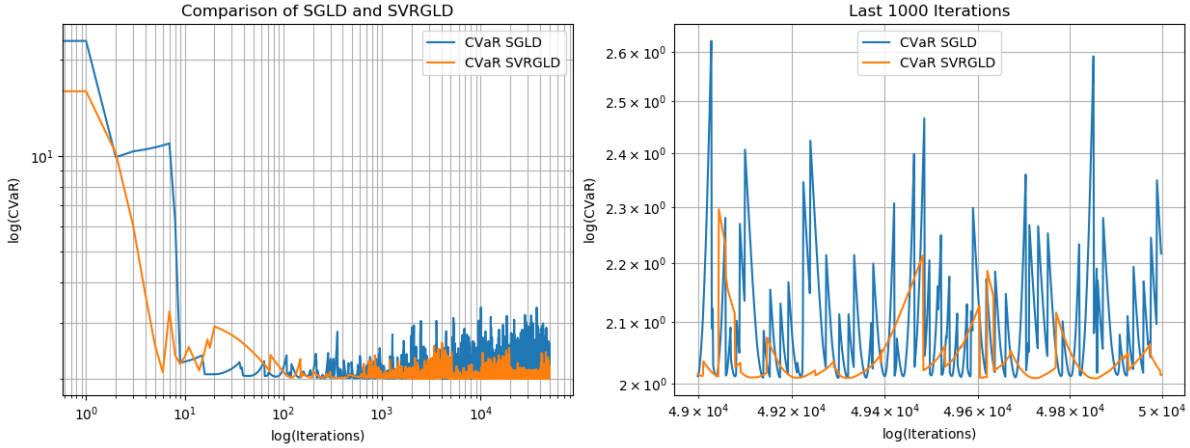


Figure 5.4: Log-scale evolution of the estimated $\widetilde{\mathrm{CVaR}}^{\mathbf{x}}_{\nu}(\Phi)$ under SGLD and SVRG-LD with step size $h = 0.02$, for $X_1 = \mathcal{N}(1,4)$ and $X_2 = \mathcal{N}(0,1)$. SVRG-LD achieves faster intial convergence to optimal CVaR and maintains the values throughout the iterations. The right panel (last 1000 iterations) highlights the lower variance and more stable behavior of SVRG-LD compared to the noisier and more erratic path of SGLD.

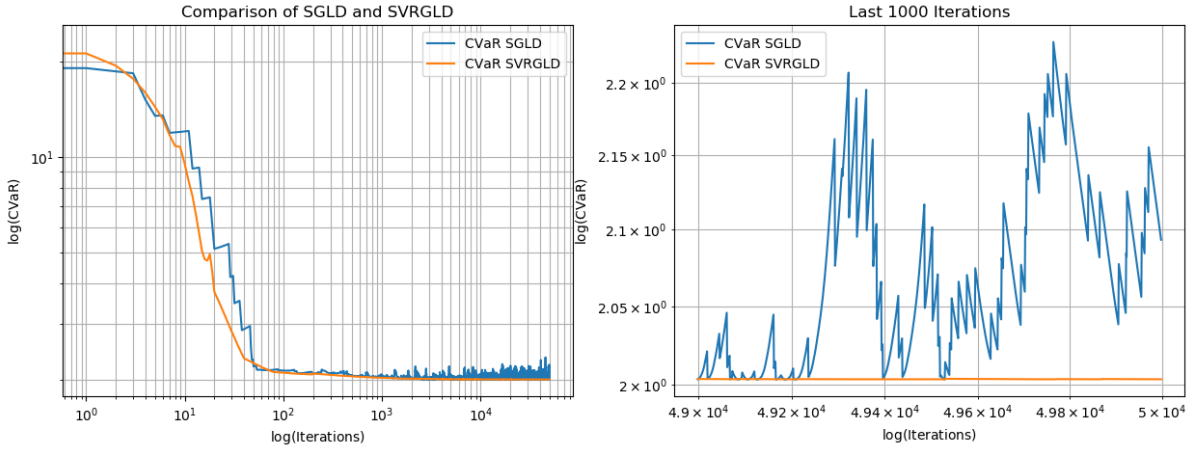


Figure 5.5: Log-scale evolution of the estimated $\widetilde{\mathrm{CVaR}}^{\mathbf{x}}_{\nu}(\Phi)$ under SGLD and SVRG-LD with step size $h = 0.005$, for $X_1 = \mathcal{N}(1,4)$ and $X_2 = \mathcal{N}(0,1)$. SVRG-LD achieves faster and smoother convergence to optimal CVaR values, while SGLD exhibits greater fluctuations, especially during the final iterations. The right panel highlights SVRG-LD's significantly lower variance and more stable behavior compared to SGLD.

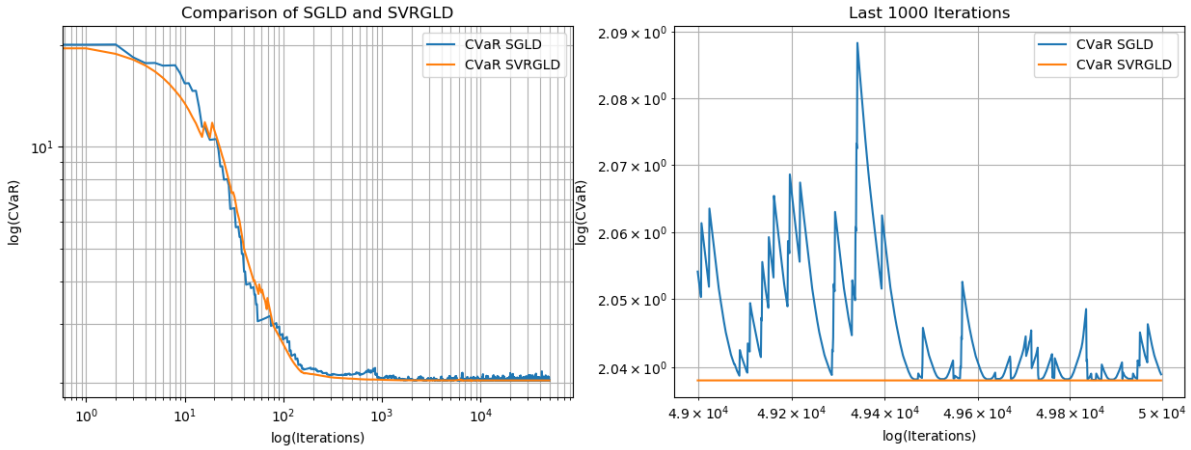


Figure 5.6: Log-scale evolution of the estimated $\widetilde{\text{CVaR}}^{\mathbf{x}}_{\nu}(\Phi)$ under SGLD and SVRG-LD with step size $h = 0.0025$, for $X_1 = \mathcal{N}(1, 4)$ and $X_2 = \mathcal{N}(0, 1)$. SVRG-LD achieves faster and smoother convergence to optimal CVaR values, while SGLD exhibits greater fluctuations, especially during the final iterations. The right panel highlights SVRG-LD's significantly lower variance and more stable behavior compared to SGLD.

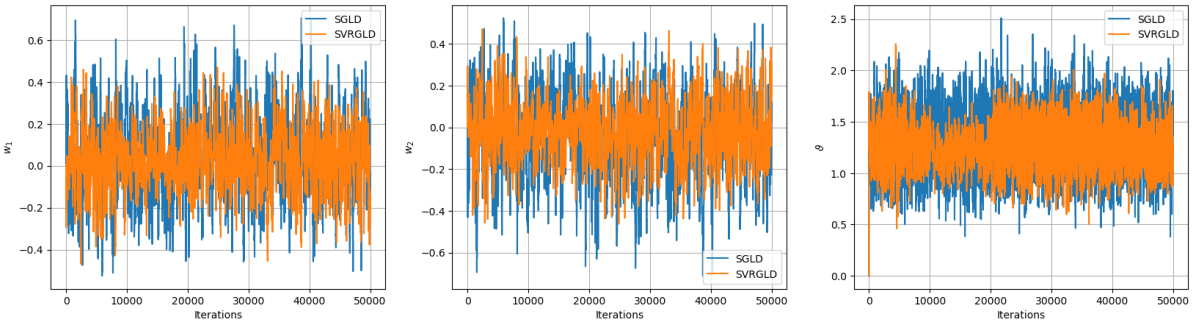


Figure 5.7: Trajectories of w_1, w_2, and and ϑ for SGLD and SVRG-LD with step size $h = 0.02$, where $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0, 1)$. At this larger step size, both methods exhibit higher variance.

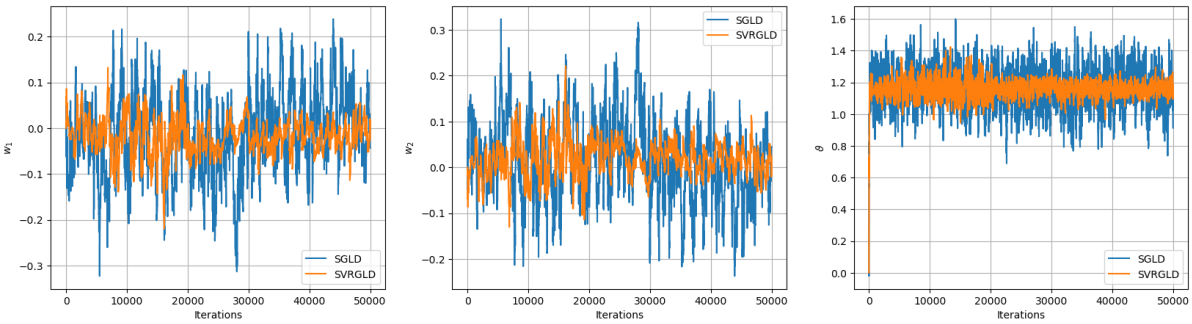


Figure 5.8: Trajectories of w_1, w_2, and and ϑ for SGLD and SVRG-LD with step size $h = 0.005$, where $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0, 1)$. At this step size, both methods exhibit higher variance, with SGLD showing more pronounced fluctuations whereas, SVRG-LD maintains relatively smoother paths.

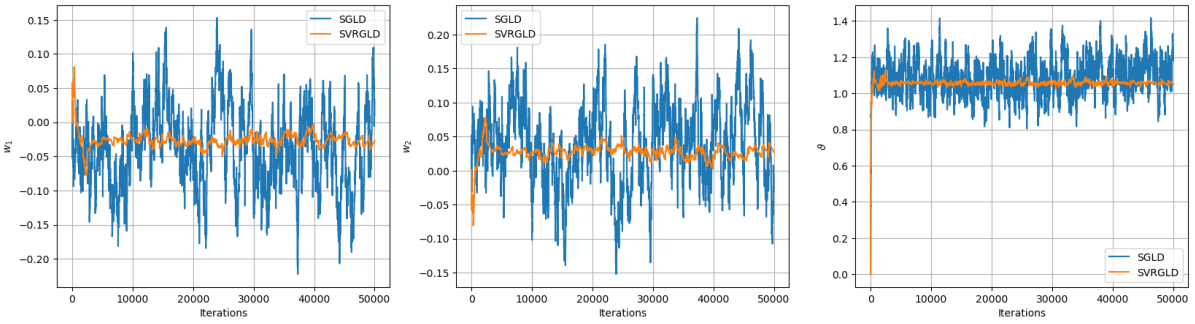


Figure 5.9: Trajectories of w_1, w_2, and and ϑ for SGLD and SVRG-LD with step size $h = 0.0025$, where $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0,1)$. Both algorithms converge toward similar parameter values, but SVRG-LD demonstrates significantly reduced variance and smoother convergence.

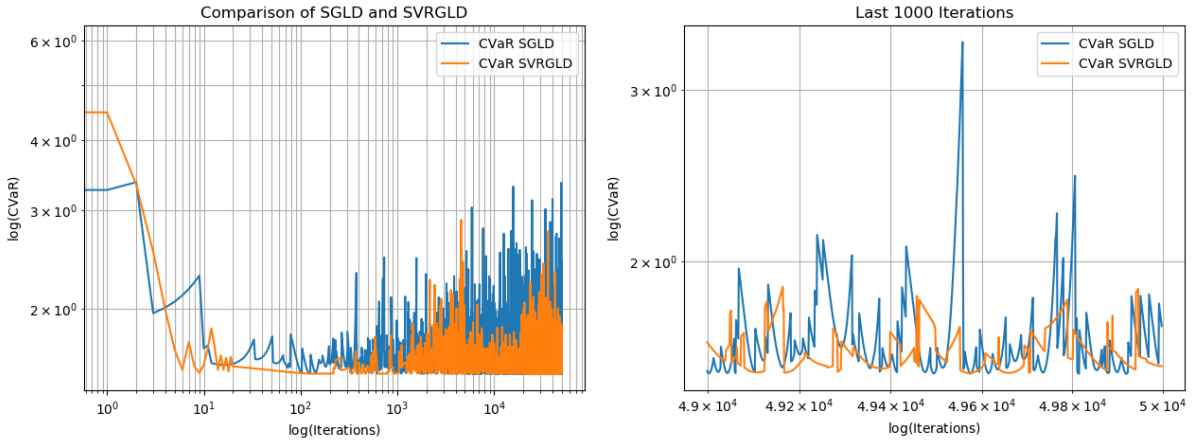


Figure 5.10: Log-scale evolution of the estimated $\widetilde{\mathrm{CVaR}}^{\mathbf{x}}_{\nu}(\Phi)$ under SGLD and SVRG-LD with step size $h = 0.02$, for $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0,1)$. SVRG-LD achieves faster and smoother convergence to the optimal CVaR values, while SGLD exhibits greater fluctuations, particularly during the later stages of sampling. The right panel highlights SVRG-LD's significantly lower variance and more stable behavior compared to SGLD in the final 1000 iterations.

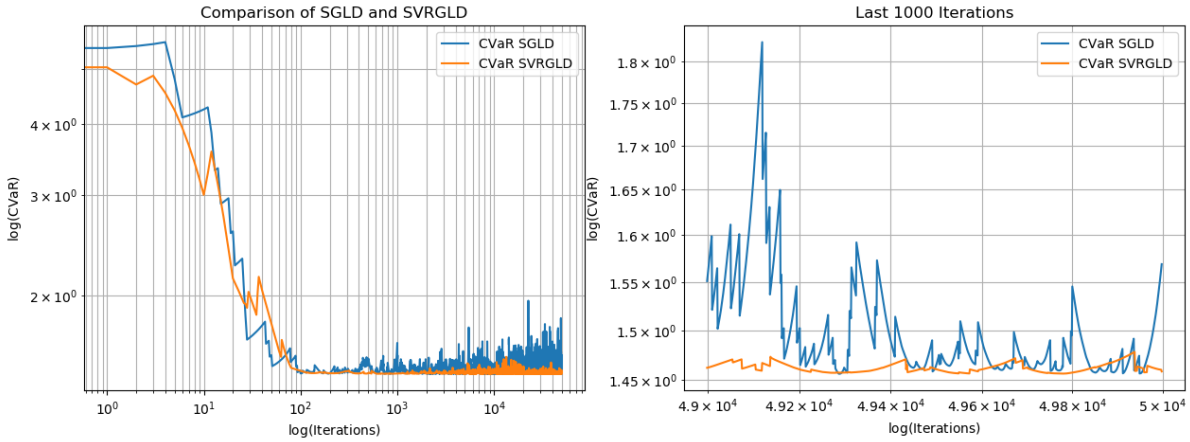


Figure 5.11: Log-scale evolution of the estimated $\widetilde{\mathrm{CVaR}}^{\mathbf{x}}_{\nu}(\Phi)$ under SGLD and SVRG-LD with step size $h = 0.005$, for $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0,1)$. Both methods converge to similar CVaR levels, but SVRG-LD achieves this with markedly lower variance. The right panel shows that in the final 1000 iterations, SVRG-LD maintains stable and concentrated estimates, while SGLD displays persistent fluctuations, underscoring SVRG-LD's robustness even at smaller step sizes.

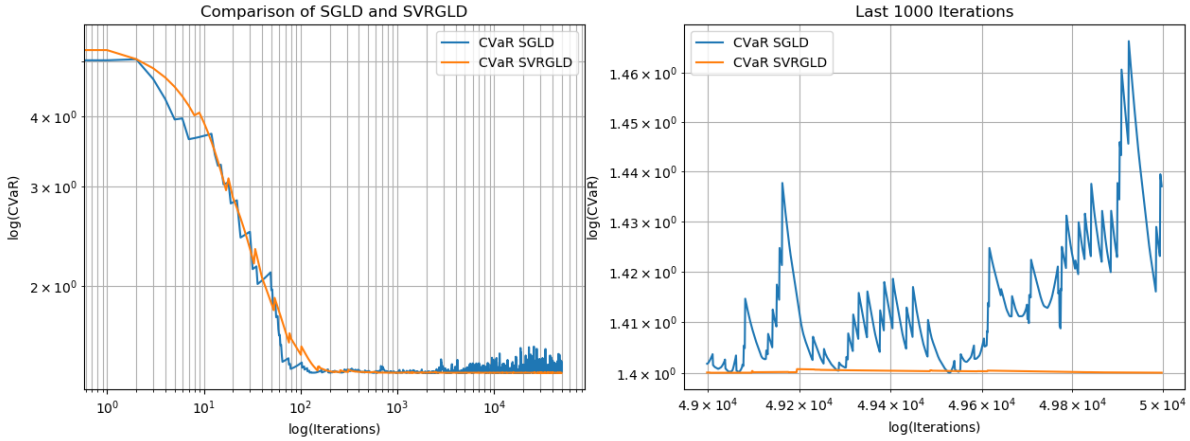
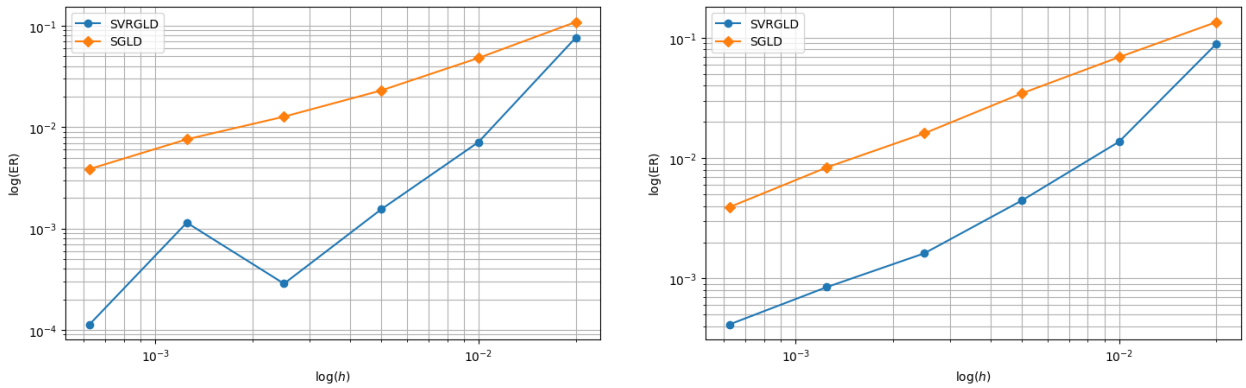


Figure 5.12: Log-scale evolution of the estimated $\widetilde{\mathrm{CVaR}}^{\mathbf{x}}_{\nu}(\Phi)$ under SGLD and SVRG-LD with step size $h = 0.005$, for $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0,1)$. Both methods exhibit similar convergence behavior in the early phase, but SVRG-LD achieves superior stability in later iterations. As shown in the right panel, SVRG-LD maintains an almost flat trajectory with minimal variance, while SGLD continues to fluctuate, highlighting the variance-reduction advantage of SVRG-LD at smaller step sizes.



(a) $X_1 = \mathcal{N}(1, 4)$ and $X_2 = \mathcal{N}(0, 1)$. (b) $X_1 = t_{\{d.f.=1000\}}$ and $X_2 = \mathcal{N}(0, 1)$.

Figure 5.13: Excess Risk as a function of step size h. log-log scale.

5.5 Summary

In this study, we focused on the efficient calculation of the minimum-CVaR portfolio. In this regard, we investigated a variance-reduced version of stochastic gradient Langevin dynamics. We conducted a comprehensive theoretical analysis and estimated the non-asymptotic error bound for expected ER. It is apparent from the findings presented in Section 5.4 that the SVRG-LD method facilitates high-precision sampling from the target distribution, yielding a significantly improved estimate of CVaR for the identical step size.

Chapter 6

Conclusion and Future Research

In this thesis, we explored Monte Carlo simulation and its variants as powerful tools for addressing computationally intensive challenges in quantitative finance. Our analysis focused on three key areas: derivative pricing, risk management, and portfolio optimization.

We developed a hybrid algorithm for derivative pricing that combines Multilevel Richardson-Romberg extrapolation (ML2R) with adaptive importance sampling to enhance computational efficiency. Numerical experiments further confirmed its effectiveness within the quantitative finance framework. From the numerical results it is evident that Adaptive Importance Sampling ML2R perform better than ML2R while achieving the desired level of accuracy. We also observed that employing the importance sampling procedure on every level of resolution might not be necessary as it would lead to an increase in computational cost without any substantial variance reduction. For instance, we observed that under the Milstein scheme, employing importance sampling only on the coarsest level would lead to substantial variance reduction without any significant computational overhead. Therefore, employing importance sampling at a low variance level would not provide a significant variance reduction but may lead to an increased computational cost. Hence, we need to be vigilant while incorporating the discussed importance sampling procedure in the ML2R framework.

We also examined stochastic optimization under a biased sampling framework, leveraging Sample Average Approximation (SAA) to solve stochastic optimization problems. Our investigation into uniform convergence properties provided valuable insights into the computational cost required to estimate optimal values. We observed that integrating MLMC in the SAA framework could lead to an improved sample complexity both in the context of getting an $\epsilon-$optimal solution and RMSE of $\mathcal{O}(\epsilon)$. The theoretical findings were complemented by the numerical estimation of CVaR, where we estimated the

optimality gap of a candidate solution $\hat{x}$.

Finally, we concentrated on effectively calculating the minimum-CVaR portfolio. To address this issue, we looked at a variance-reduced version of Stochastic Gradient Langevin Dynamics (SVRG-LD). The performance of our strategy was confirmed by numerical tests and backed by a non-asymptotic error constraint for the Expected Excess Risk.

While many of the examples developed in this thesis focus on scalar SDEs, the methods—particularly those based on the MLMC framework—can be naturally extended to systems of SDEs. For example, in the context of importance sampling, such extensions to higher dimensions are straightforward when employing the Euler–Maruyama discretization. However, we acknowledge that high-dimensional settings introduce additional challenges, particularly when using the Milstein scheme, where the computation of Lévy areas becomes non-trivial. Addressing such issues may require specialized numerical techniques or approximations. A comprehensive treatment of these aspects remains an important direction for future research.

As observed in the conduct of the numerical experiment the effect of the hybrid method is prominent when the rate of convergence of variance $\beta > 1$. However, as discussed $\beta > 1$ with the Milstein scheme for a high dimensional SDE is difficult due to Lévy area simulation, which is computationally expensive. Therefore, in future research, we would like to study the importance of sampling under the antithetic MLMC paradigm for problems dealing with higher dimensional SDEs. Similar to the discussion in Chapter 2 and Chapter 3, one could leverage the Central Limit Theorem developed and discussed in [11] to design the importance sampling estimator for high dimensional SDEs. Another direction of research we would like to explore is to integrate the importance sampling technique with the adaptive sampling MLMC technique to improve the computational cost in the nested expectation paradigm.

The most natural extension of the study undertaken in Chapter 3 would be to examine the effect of biased sampling in a multistage stochastic program. From the computational perspective, we can introduce retrospective approximation [69] in the biased sampling framework and conduct an extensive analysis with respect to convergence and sample complexity with respect to both Monte Carlo and Multilevel Monte Carlo setups. Moreover, in this study, we only look into deterministic constraints. It would be an interesting problem to study the SAA under the biased sampling framework with chance constraints, such as risk-averse portfolio optimization problems [74] and stochastic control problems [61]. Additionally, it would be interesting to incorporate various variance reduction techniques to further improve the performance of the SAA.

In future, we would also like to focus on examining the impact of bias induced in the sample X_i's and performing a comprehensive error analysis to provide an upper bound for the Excess Risk. The impact of biased sampling has been recently studied in the stochastic gradient descent paradigms with extension to the multilevel framework (see, e.g. [22]). However, no such study has been conducted using the SGLD framework. We would like to explore this area and undertake a theoretical analysis with the aim of providing a bound for the Excess Risk. We can further incorporate the SVRG-LD in the biased framework to improve our computational results. Additionally, for quicker convergence and less computing load, we will study the concept of a multilevel gradient estimator under biased sampling.

Overall, our findings highlight the potential of advanced Monte Carlo techniques in financial applications, offering improved accuracy and computational efficiency across derivative pricing, risk assessment, and portfolio optimization.

Bibliography

[1] S. Ahn, A. Korattikara, and M. Welling. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, pages 1–9, 2012.

[2] M. B. Alaya, K. Hajji, and A. Kebaier. Importance Sampling and Statistical Romberg Method. *Bernoulli*, 21(4):1947–1983, 2015.

[3] M. B. Alaya, K. Hajji, and A. Kebaier. Improved Adaptive Multilevel Monte Carlo and Applications to Finance. *arXiv Preprint arXiv:1603.02959*, 2016.

[4] M. B. Alaya and A. Kebaier. Central Limit Theorem for the Multilevel Monte Carlo Euler Method. *The Annals of Applied Probability*, 25(1):211–234, 2015.

[5] Y. M. Y. and. A General Stochastic Programming Problem. *Journal of Cybernetics*, 1(4):106–112, 1971.

[6] C. Andrieu, É. Moulines, and P. Priouret. Stability of Stochastic Approximation Under Verifiable Conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312, 2005.

[7] B. Arouna. Adaptive Monte Carlo Method: A Variance Reduction Technique. *Monte Carlo Methods and Applications*, 10(1):1–24, 2004.

[8] D. Banholzer, J. Fliege, and R. Werner. On Rates of Convergence for Sample Average Approximations in the Almost Sure Sense and in Mean. *Mathematical Programming*, pages 1–39, 2022.

[9] O. Bardou, N. Frikha, and G. Pagès. Computing VaR and CVaR Using Stochastic Approximation and Adaptive Unconstrained Importance Sampling. *Monte Carlo Methods and Applications*, 15(3):173–210, 2009.

[10] M. Ben Alaya, K. Hajji, and A. Kebaier. Adaptive Importance Sampling for Multilevel Monte Carlo Euler Method. *Stochastics*, 95(2):303–327, 2023.

[11] M. Ben Alaya, A. Kebaier, and T. B. T. Ngo. Central Limit Theorem for the Antithetic Multilevel Monte Carlo Method. *The Annals of Applied Probability*, 32(3):1970–2027, 2022.

[12] F. Bolley, I. Gentil, and A. Guillin. Convergence to Equilibrium in Wasserstein Distance for Fokker–Planck Equations. *Journal of Functional Analysis*, 263(8):2430–2457, 2012.

[13] F. Bourgey. *Stochastic Approximations for Financial Risk Computations*. PhD thesis, Institut Polytechnique de Paris, 2020.

[14] M. Broadie, Y. Du, and C. C. Moallemi. Efficient Risk Estimation via Nested Sequential Simulation. *Management Science*, 57(6):1172–1194, 2011.

[15] N. Chatterji, N. Flammarion, Y. Ma, P. Bartlett, and M. Jordan. On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR, 2018.

[16] N. H. Chau, É. Moulines, M. Rásonyi, S. Sabanis, and Y. Zhang. On Stochastic Gradient Langevin Dynamics With Dependent Data Streams: The Fully Nonconvex Case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

[17] C. Chen, W. Wang, Y. Zhang, Q. Su, and L. Carin. A Convergence Analysis for a Class of Practical Variance-Reduction Stochastic Gradient MCMC. *Science China Information Sciences*, 62:1–13, 2019.

[18] H.-F. Chen. Robbins-Monro Algorithm. *Stochastic Approximation and Its Applications*, pages 1–24, 2002.

[19] H.-F. Chen, L. Guo, and A.-J. Gao. Convergence and Robustness of the Robbins-Monro Algorithm Truncated at Randomly Varying Bounds. *Stochastic Processes and Their Applications*, 27:217–231, 1987.

[20] S. X. Chen. Nonparametric Estimation of Expected Shortfall. *Journal of Financial Econometrics*, 6(1):87–107, 2008.

[21] J. Chu and L. Tangpi. Non-Asymptotic Estimation of Risk Measures Using Stochastic Gradient Langevin Dynamics. *SIAM Journal on Financial Mathematics*, 15(2):503–536, 2024.

[22] S. Crépey, N. Frikha, and A. Louzi. A Multilevel Stochastic Approximation Algorithm for Value-at-Risk and Expected Shortfall Estimation. *arXiv Preprint arXiv:2304.01207*, 2023.

[23] A. S. Dalalyan. Theoretical Guarantees for Approximate Sampling From Smooth and Log-Concave Densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[24] D. Dentcheva, S. Penev, and A. Ruszczyński. Statistical Estimation of Composite Risk Functionals and Risk Optimization Problems. *Annals of the Institute of Statistical Mathematics*, 69:737–760, 2017.

[25] K. A. Dubey, S. J Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing. Variance Reduction in Stochastic Gradient Langevin Dynamics. *Advances in Neural Information Processing Systems*, 29, 2016.

[26] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via Convex Optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[27] A. Durmus and É. Moulines. Non-Asymptotic Convergence Analysis for the Unadjusted Langevin Algorithm. *The Annals of Applied Probability*, pages 1551–1587, 2017.

[28] A. Durmus and É. Moulines. High-Dimensional Bayesian Inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[29] Y. M. Ermoliev and V. I. Norkin. Sample Average Approximation Method for Compound Stochastic Optimization Problems. *SIAM Journal on Optimization*, 23(4):2231–2263, 2013.

[30] N. Frikha. Multi-Level Stochastic Approximation Algorithms. *The Annals of Applied Probability*, 26(2):933–985, 2016.

[31] M. Giles. Improved Multilevel Monte Carlo Convergence Using the Milstein Scheme. In *Monte Carlo and Quasi-Monte Carlo Methods 2006*, pages 343–358. Springer, 2008.

[32] M. B. Giles. Multilevel Monte Carlo Path Simulation. *Operations Research*, 56(3):607–617, 2008.

[33] M. B. Giles. MLMC for Nested Expectations. *Contemporary Computational Mathematics: A Celebration of the 80th Birthday of Ian Sloan*, pages 425–442, 2018.

[34] M. B. Giles, K. Debrabant, and A. Rössler. Analysis of Multilevel Monte Carlo Path Simulation Using the Milstein Discretisation. *Discrete & Continuous Dynamical Systems - B*, 24(8):3881, 2019.

[35] M. B. Giles and A.-L. Haji-Ali. Multilevel Nested Simulation for Efficient Risk Estimation. *SIAM/ASA Journal on Uncertainty Quantification*, 7(2):497–525, 2019.

[36] M. B. Giles and L. Szpruch. Antithetic Multilevel Monte Carlo Estimation for Multidimensional SDEs. In *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 367–384. Springer, 2013.

[37] M. B. Giles and L. Szpruch. Antithetic Multilevel Monte Carlo Estimation for Multi-Dimensional SDEs Without Lévy Area Simulation. *The Annals of Applied Probability*, 24(4):1585–1620, 2014.

[38] M. B. Giles and B. J. Waterhouse. Multilevel Quasi-Monte Carlo Path Simulation. In *Advanced Financial Modelling*, pages 165–182. De Gruyter, 2009.

[39] D. Giorgi, V. Lemaire, and G. Pagès. Limit Theorems for Weighted and Regular Multilevel Estimators. *Monte Carlo Methods and Applications*, 23(1):43–70, 2017.

[40] P. Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer, 2004.

[41] M. B. Gordy and S. Juneja. Nested Simulation in Portfolio Risk Measurement. *Management Science*, 56(10):1833–1848, 2010.

[42] A.-L. Haji-Ali, J. Spence, and A. L. Teckentrup. Adaptive Multilevel Monte Carlo for Probabilities. *SIAM Journal on Numerical Analysis*, 60(4):2125–2149, 2022.

[43] P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 2014.

[44] S. Heinrich. Multilevel Monte Carlo Methods. pages 58–67, 2001.

[45] H.F.Chen and Y. Zhu. Stochastic Approximation Procedures With Randomly Varying Truncations. *Science in China, Ser. A*, 1986.

[46] T. Homem-de Mello and G. Bayraksan. Monte Carlo Sampling-Based Methods for Stochastic Optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

[47] L. J. Hong, S. Juneja, and G. Liu. Kernel Smoothing for Nested Estimation With Application to Portfolio Risk Measurement. *Operations Research*, 65(3):657–673, 2017.

[48] L. J. Hong and G. Liu. Monte Carlo Estimation of Value-at-Risk, Conditional Value-at-Risk and Their Sensitivities. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pages 11–14. IEEE.

[49] Y. Hu, X. Chen, and N. He. Sample Complexity of Sample Average Approximation for Conditional Stochastic Optimization. *SIAM Journal on Optimization*, 30(3):2103–2133, 2020.

[50] C.-R. Hwang. Laplace's Method Revisited: Weak Convergence of Probability Measures. *The Annals of Probability*, pages 1177–1182, 1980.

[51] G. Iyengar and A. K. C. Ma. Fast Gradient Descent Method for Mean-CVaR Optimization. *Annals of Operations Research*, 205:203–212, 2013.

[52] A. Kebaier and J. Lelong. Coupling Importance Sampling and Multilevel Monte Carlo Using Sample Average Approximation. *Methodology and Computing in Applied Probability*, 20(2):611–641, 2018.

[53] Y. Kinoshita and T. Suzuki. Improved Convergence Rate of Stochastic Gradient Langevin Dynamics With Variance Reduction and Its Application to Optimization. In *Advances in Neural Information Processing Systems*, volume 35, pages 19022–19034, 2022.

[54] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The Sample Average Approximation Method for Stochastic Discrete Optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[55] P. E. Kloeden and E. Platen. Stochastic Differential Equations. In *Numerical Solution of Stochastic Differential Equations*, pages 103–160. Springer, 1992.

[56] H. Lam and H. Qian. Bounding Optimality Gap in Stochastic Optimization via Bagging: Statistical Efficiency and Stability. *arXiv Preprint arXiv:1810.02905*, 2018.

[57] S. Laruelle, C.-A. Lehalle, et al. Optimal Posting Price of Limit Orders: Learning by Trading. *Mathematics and Financial Economics*, 7(3):359–403, 2013.

[58] J. Lelong. Almost Sure Convergence of Randomly Truncated Stochastic Algorithms Under Verifiable Conditions. *Statistics & Probability Letters*, 78(16):2632–2636, 2008.

[59] V. Lemaire and G. Pagès. Unconstrained Recursive Importance Sampling. *The Annals of Applied Probability*, 20(3):1029–1067, 2010.

[60] V. Lemaire and G. Pagès. Multilevel Richardson–Romberg Extrapolation. *Bernoulli*, 23(4A):2643–2692, 2017.

[61] T. Lew, R. Bonalli, and M. Pavone. Sample Average Approximation for Stochastic Programming With Equality Constraints. *SIAM Journal on Optimization*, 34(4):3506–3533, 2024.

[62] L. Liang, A. Neufeld, and Y. Zhang. Non-Asymptotic Convergence Analysis of the Stochastic Gradient Hamiltonian Monte Carlo Algorithm With Discontinuous Stochastic Gradient With Applications to Training of ReLU Neural Networks. *arXiv Preprint arXiv:2409.17107*, 2024.

[63] D.-Y. Lim, A. Neufeld, S. Sabanis, and Y. Zhang. Langevin Dynamics Based Algorithm e-TH ε O POULA for Stochastic Optimization Problems With Discontinuous Stochastic Gradient. *Mathematics of Operations Research*, 2024.

[64] D.-Y. Lim, A. Neufeld, S. Sabanis, and Y. Zhang. Non-Asymptotic Estimates for TUSLA Algorithm for Non-Convex Learning With Applications to Neural Networks With ReLU Activation Function. *IMA Journal of Numerical Analysis*, 44(3):1464–1559, 2024.

[65] D.-Y. Lim and S. Sabanis. Polygonal Unadjusted Langevin Algorithms: Creating Stable and Efficient Adaptive Algorithms for Neural Networks. *Journal of Machine Learning Research*, 25(53):1–52, 2024.

[66] W. K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo Bounding Techniques for Determining Solution Quality in Stochastic Programs. *Operations Research Letters*, 24(1):47–56, Feb. 1999.

[67] A. Neufeld, M. N. C. En, and Y. Zhang. Robust SGLD Algorithm for Solving Non-Convex Distributionally Robust Optimisation Problems. *arXiv Preprint arXiv:2403.09532*, 2024.

[68] G. Pagès. Numerical Probability. *Universitext, Springer*, 2018.

[69] R. Pasupathy. On Choosing Parameters in Retrospective-Approximation Algorithms for Stochastic Root Finding and Simulation Optimization. *Operations Research*, 58(4-part-1):889–901, 2010.

[70] M. Pelletier. Asymptotic Almost Sure Efficiency of Averaged Stochastic Algorithms. *SIAM Journal on Control and Optimization*, 39(1):49–72, 2000.

[71] M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Non-Asymptotic Analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 2017.

[72] C.-h. Rhee and P. W. Glynn. Unbiased Estimation With Square Root Convergence for SDE Models. *Operations Research*, 63(5):1026–1043, 2015.

[73] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.

[74] R. T. Rockafellar, S. Uryasev, et al. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–42, 2000.

[75] A. Ruszczyński and A. Shapiro. Stochastic Programming Models. *Handbooks in Operations Research and Management Science*, 10:1–64, 2003.

[76] S. Sabanis and Y. Zhang. A Fully Data-Driven Approach to Minimizing CVaR for Portfolio of Assets via SGLD With Discontinuous Updating. *arXiv Preprint arXiv:2007.01672*, 2020.

[77] I. Sato and H. Nakagawa. Variance Reduction in Stochastic Gradient Langevin Dynamics for Bayesian Sampling Tasks. *Journal of Machine Learning Research*, 15(3):982–1010, 2014.

[78] K. Sekimoto. *Stochastic Energetics*. Springer, 2010.

[79] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

[80] D. Talay and L. Tubaro. Expansion of the Global Error for Numerical Schemes Solving Stochastic Differential Equations. *Stochastic Analysis and Applications*, 8(4):483–509, 1990.

[81] A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via Sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[82] Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17:1–33, 2016.

[83] M. Wang, E. X. Fang, and H. Liu. Stochastic Compositional Gradient Descent: Algorithms for Minimizing Compositions of Expected-Value Functions. *Mathematical Programming*, 161:419–449, 2017.

[84] M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

[85] J. Wellner et al. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 2013.

[86] P. Xu, J. Chen, D. Zou, and Q. Gu. Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[87] J. Zhang et al. Low-Precision Stochastic Gradient Langevin Dynamics for Efficient Sampling in Bayesian Inference. *Proceedings of the International Conference on Machine Learning (ICML)*, pages 26624–26644, 2022.

[88] H. Zhu and E. Zhou. Estimation of Conditional Value-at-Risk for Input Uncertainty With Budget Allocation. In *2015 Winter Simulation Conference (WSC)*, pages 655–666. IEEE, 2015.

[89] D. Zou, P. Xu, and Q. Gu. Subsampled Stochastic Variance-Reduced Gradient Langevin Dynamics. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.

[90] D. Zou, P. Xu, and Q. Gu. Faster Convergence of Stochastic Gradient Langevin Dynamics for Non-Log-Concave Sampling. In *Proceedings of the 37th International*

Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 3337–3347, 2020.

[91] D. Zou, P. Xu, and Q. Gu. Sampling From Non-Log-Concave Distributions via Stochastic Variance-Reduced Gradient Langevin Dynamics. In *AISTATS 2019-22nd International Conference on Artificial Intelligence and Statistics*, 2020.

List of Accepted or Communicated Papers

Based on the work in this thesis, the following research articles have been accepted or communicated.

1. Sinha, Devang, and Siddhartha P. Chakrabarty. "Multilevel Richardson-Romberg and Importance Sampling for Efficient Simulation". (*In Review*)
2. Sinha, Devang, and Siddhartha P. Chakrabarty. "A Review of Efficient Multilevel Monte Carlo Algorithms for Derivative Pricing and Risk Management." MethodsX 10 (2023): 102078.
3. Sinha, Devang, and Siddhartha P. Chakrabarty. "Multilevel Monte Carlo in Sample Average Approximation: Convergence, Complexity and Application. (*In Revision*)
4. Sinha, Devang. "Non-Asymptotic Estimation of Conditional Value-at-Risk via Stochastic Variance Reduced Gradient Langevin Dynamics". (*In Review*)