Efficient Simulation Methods for Estimating Risk Measures

Yiping Du

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011
ABSTRACT

Efficient Simulation Methods for Estimating Risk Measures

Yiping Du

In this thesis, we analyze the computational problem of estimating financial risk in nested Monte Carlo simulation. An outer simulation is used to generate financial scenarios, and an inner simulation is used to estimate future portfolio values in each scenario. Mean squared error (MSE) for standard nested simulation converges at the rate $k^{-2/3}$, where $k$ is the computational budget.

In the first part of this thesis, we focus on one risk measure, the probability of a large loss, and we propose a new algorithm to estimate this risk. Our algorithm sequentially allocates computational effort in the inner simulation based on marginal changes in the risk estimator in each scenario. Theoretical results are given to show that the risk estimator has an asymptotic MSE of order $k^{-4/5+\epsilon}$, for all positive $\epsilon$, that is faster compared to the conventional uniform inner sampling approach. Numerical results consistent with the theory are presented.

In the second part of this thesis, we introduce a regression-based nested Monte Carlo simulation method for risk estimation. The proposed regression method combines information from different risk factor realizations to provide a better estimate of the portfolio loss function. The MSE of the regression method converges at the rate $k^{-1}$ until reaching an asymptotic bias level which depends on the magnitude of the regression error. Numerical results consistent with our theoretical analysis are provided and numerical comparisons with other methods are also given.

In the third part of this thesis, we propose a method based on weighted regression. Similar to the unweighted regression method, the MSE of the weighted regression method converges at the rate $k^{-1}$ until reaching an asymptotic bias level, which depends on the size of the regression error. However, the weighted approach further reduces MSE by emphasizing scenarios that are more important to the calculation of the risk measure. We find a globally optimal weighting strategy for
general risk measures in an idealized setting. For applications, we propose and test a practically implementable two-pass method, where the first pass uses an unweighted regression and the second pass uses weights based on the first pass.
## Table of Contents

1 **Introduction** .................................................................................................................. 1

  1.1 Introduction and Problem Motivation ............................................................................. 1

  1.2 Problem Formulation: Nested Simulation ..................................................................... 2

  1.3 The Sequential Method ................................................................................................. 6

  1.4 The Unweighted Regression Method ............................................................................ 7

  1.5 The Weighted Regression Method ................................................................................ 9

  1.6 Literature Review ......................................................................................................... 10

2 **The Sequential Method** .................................................................................................. 14

  2.1 Optimal Uniform .......................................................................................................... 14

  2.2 Sequential Sampling ..................................................................................................... 19

  2.3 Analysis ......................................................................................................................... 25

     2.3.1 A Simplified Non-Uniform Estimator ....................................................................... 26

     2.3.2 Asymptotic Analysis ................................................................................................. 29

     2.3.3 Optimal Non-Uniform Threshold Estimator ............................................................. 32

  2.4 Adaptive Allocation Algorithm .................................................................................... 33

  2.5 Numerical Results ......................................................................................................... 36

     2.5.1 Experimental Setting ............................................................................................... 38

     2.5.2 Bias Comparison .................................................................................................... 39

     2.5.3 MSE Comparison .................................................................................................... 41

     2.5.4 Variance Estimation ............................................................................................... 47

  2.6 Conclusion ....................................................................................................................... 48
3 The Unweighted Regression Method

3.1 Delta-Gamma Approximation .................................................. 50
3.2 The Unweighted Regression Algorithm ..................................... 51
3.3 Analysis .................................................................................. 54
   3.3.1 Differentiable Case ............................................................ 55
   3.3.2 Lipschitz Continuous Case ................................................ 57
3.4 Numerical Results ................................................................. 60
   3.4.1 Experimental Setting ......................................................... 61
   3.4.2 Single Asset Example ....................................................... 62
   3.4.3 Multiple Asset Examples .................................................. 65
3.5 Conclusion .............................................................................. 68

4 The Weighted Regression Method ............................................... 73

4.1 The Weighted Regression Algorithm ......................................... 73
4.2 Analysis .................................................................................. 75
4.3 Practical Implementation ......................................................... 77
4.4 Numerical Results ................................................................. 79
   4.4.1 Examples ......................................................................... 79
   4.4.2 Numerical Performance .................................................... 81
4.5 Conclusion .............................................................................. 84

Bibliography ................................................................................. 84

A Proofs Related to the Sequential Method ................................... 89

A.1 Preliminaries .......................................................................... 89
A.2 Asymptotic Bias ..................................................................... 91
A.3 Expected Number of Inner Samples ........................................ 93
A.4 Realized Number of Inner Samples ......................................... 96

B Proofs Related to the Unweighted Regression Method ...................... 100

B.1 Differentiable Case ............................................................... 102
B.2 Lipschitz Continuous Case ................................................... 105
# List of Figures

1.1 The Risk Management Problem ............................................. 4  
1.2 Two-Stage Sampling .......................................................... 5  

2.1 The Benefits of Non-Uniform Sampling .................................. 20  
2.2 The Impact of an Additional Sample ..................................... 22  
2.3 The Threshold Estimator .................................................... 29  
2.4 Bias as a Function of the Total Number of Inner Stage Samples  .................................................... 42  
2.5 Distribution of Inner Stage Samples ..................................... 43  

3.1 The Mean Squared Error in the Single Asset Example ............... 64  
3.2 Approximations in the Single Asset Example .......................... 65  
3.3 The Mean Squared Error in Example $E_{10}$ .......................... 69  
3.4 The Mean Squared Error in Example $E_{10E}$ .......................... 70  
3.5 The Mean Squared Error in Example $E_{100}$ .......................... 71  

4.1 The Mean Squared Error in the One-Dimensional Example .......... 81  
4.2 The Mean Squared Error in the 10-Dimensional Delta-Hedged Example .................................................... 82  
4.3 Approximations in the One-Dimensional Example .................... 83
List of Tables

2.1 Numerical Results of the Gaussian Example . . . . . . . . . . . . . . . . . . . . 45
2.2 Numerical Results of the Put Example . . . . . . . . . . . . . . . . . . . . . . . . 46
3.1 Numerical Results of Four Examples . . . . . . . . . . . . . . . . . . . . . . . . . . 72
Acknowledgments

The work in my thesis represents an outcome of the collaboration with my advisors, Professor Mark Broadie and Professor Ciamac C. Moallemi. I would like to express my deepest gratitude to them for their guidance and instructions. It was memorable working with them.

I am very grateful to Professor Steven Kou, Professor Jose Blanchet, and Professor Martin Haugh for serving on my dissertation committee. Also, I would like to thank everyone who have instructed me or helped me during the course of my study.

Finally, I would like to take this opportunity to give special thanks to my family and friends for their support and encouragement.
Chapter 1

Introduction

1.1 Introduction and Problem Motivation

The measurement and management of risk is an increasingly important function at financial institutions. Of particular concern is financial risk due to changes of the investment portfolio value. Banks and other financial institutions have a significant need for effective methods of risk measurement to recognize, evaluate, and scrutinize the financial risks in their portfolios. A primary goal of risk measurement is to ensure that banks and other financial firms have sufficient capital reserves in relation to their holdings and investment activities in the event of a large loss.

Risk measurement is typically divided into two stages: scenario generation and portfolio revaluation. Scenario generation refers to the sampling of risk factors over a given time horizon. The risk of portfolios is induced by the variability of financial risk factors, and therefore can be quantified by generating possible future scenarios. This step is called outer (or, first) stage sampling. For realistic models with a large number of correlated risk factors, this outer stage is often performed with Monte Carlo simulation. Portfolio re-valuation refers to the computation of the portfolio value at the risk time horizon given a particular scenario of risk factors. New financial securities with complex cash flows have been invented and introduced in the financial markets. A portfolio may contain a large number of securities for most of which analytical pricing formulas do not exist. In these cases, the re-valuation task could require Monte Carlo simulation for this inner (or, second) stage. Thus, in realistic applications, the risk measurement calculation involves a two-stage, nested Monte Carlo simulation. Nested Monte Carlo simulation can represent a prohibitive computational
challenge. To make this nested simulation more effective, various approaches are often employed. In this thesis, we examine different ways of estimating risk measures with a focus on developing improvements to the standard nested Monte Carlo simulation approach.

### 1.2 Problem Formulation: Nested Simulation

Consider the problem of measuring the risk of a portfolio of securities at some future time \( t = \tau \) (the risk horizon), from the perspective of an observer at time \( t = 0 \). Financial institutions are generally interested in risks of their portfolios when \( \tau \) is small, e.g., one day or one week. Denote the current portfolio value by \( X_0 \). The value of the portfolio at time \( \tau \), \( X_\tau \), is, in general, unknown at time 0, and is dependent on a collection of financial risk factors that are realized between times 0 and \( \tau \). Examples of risk factors include stock prices, commodity prices, interest rates, and exchange rates. In particular, suppose that \( \Omega \subset \mathbb{R}^Q \) is a set of possible future risk factors realized by time \( \tau \). We assume that each \( \omega \in \Omega \) incorporates sufficient information so as to determine all asset prices at time \( \tau \). Hence, we will refer to \( \omega \) as a scenario (or, a state). In each scenario \( \omega \in \Omega \), the portfolio has value \( X_\tau (\omega) \). The mark-to-market loss of the portfolio at time \( \tau \), which we will denote by \( L(\omega) \), is thus a function of \( \omega \), and is given by\(^1\)

\[
L(\omega) \triangleq X_0 - X_\tau (\omega).
\]

We assume that there exists a probability space such that \( \omega \) is distributed according to the real-world distribution of risk factors over the state space \( \Omega \), and that \( L(\cdot) \) is a measurable function, so that \( L(\omega) \) has the real-world distribution of portfolio losses. A risk measure is a function that associates the real-world distribution of the portfolio loss \( L \) with a single scalar \( \alpha \). Some common examples of risk measures include value at risk and expected shortfall, which are described, for example, in Artzner et al. (1999), Jorion (2006), and Rockafellar and Uryasev (2002). In this thesis, we specifically consider risk measures of the form

\[
\alpha \triangleq \mathbb{E}[f(L(\omega))],
\]  

\(^{1}\) Without loss of generality, we assume the portfolio has no intermediate cashflows before time \( \tau \), and that the riskless rate is 0 before time \( \tau \).
given a function $f : \mathbb{R} \to \mathbb{R}$, where we assume that $L(\cdot)$ and $f(\cdot)$ are functions such that the expectation in (1.1) exists. Some common examples of risk measures in this form include:

- **Probability of a large loss.** Given a loss threshold $c \in \mathbb{R}$, define

  $$f(L) \triangleq I_{\{L \geq c\}}. \tag{1.2}$$

  In this case, the risk measure $\alpha$ is the probability that the portfolio loss exceeds the level $c$.

- **Expected excess loss.** Given a loss threshold $c \in \mathbb{R}$, define

  $$f(L) \triangleq (L - c)^+. \tag{1.3}$$

  In this case, $\alpha$ is the expected value of losses in excess of the level $c$.

- **Squared tracking error.** Given a target level $c \in \mathbb{R}$, define

  $$f(L) \triangleq (L - c)^2. \tag{1.4}$$

  In this case, $\alpha$ is the expected squared error of the portfolio loss relative to the target $c$.

In order to estimate a risk measure $\alpha$, we face two challenges. First, typically, the space of possible scenarios $\Omega$ is quite large, if not infinite. Thus, one approach is to approximate the distribution of the portfolio loss $L$ with an empirical distribution obtained by Monte Carlo sampling. In particular, consider a set of $n$ scenarios $\omega^{(1)}, \ldots, \omega^{(n)}$ that are independent and identically distributed according to the real-world distribution of the risk factors $\omega$. These samples are referred to as the outer stage of the simulation. We can approximate the risk measure by

$$\frac{1}{n} \sum_{i=1}^{n} f \left( L \left( \omega^{(i)} \right) \right). \tag{1.5}$$

However, another challenge in computing the risk measure $\alpha$ is that, even in a single scenario $\omega$, often the portfolio loss $L(\omega)$ is difficult to compute in closed form. In many financial applications, the portfolio may contain a collection of complex, path-dependent securities whose values depend on random cashflows between times $\tau$ and a long-term horizon $T$. This is illustrated in Figure 1.1.

---

2 An indicator function $I(A)$ is a function that has the value 1 if the expression $A$ is true, and value 0 otherwise.
CHAPTER 1. INTRODUCTION

Figure 1.1: Illustration of the risk management problem. The scenario, or the realization of risk factors between times 0 and $\tau$, is denoted by $\omega$ and is distributed according to the real-world distribution. Conditioned on the scenario $\omega$, the portfolio loss $L(\omega)$ depends on the expected value of further random cashflows between $\tau$ and $T$, under the risk-neutral distribution.

Hence, the loss $L(\omega)$ may also need to be numerically estimated. A common way to estimate the portfolio loss in a given scenario is via an inner stage Monte Carlo simulation of the expected cashflows of the portfolio over the interval $[\tau, T]$. The inner simulation occurs under the risk-neutral distribution, conditioned on the scenario $\omega$. Specifically, suppose that $T$ is the longest maturity of all of the securities in the portfolio. The value of the portfolio at the risk horizon $\tau$ in scenario $\omega^{(i)}$ is equal to the expected discounted cashflows of the portfolio over the interval $[\tau, T]$. If $\hat{Z}_{i,1}, \ldots, \hat{Z}_{i,m}$ are $m$ i.i.d. samples of losses generated according to this inner stage simulation, each with mean $L(\omega^{(i)})$, then we can approximate the loss $L(\omega^{(i)})$ at scenario $\omega^{(i)}$ by

$$\hat{L}(\omega^{(i)}, \zeta^{(i)}) \triangleq \frac{1}{m} \sum_{j=1}^{m} \hat{Z}_{i,j}, \quad (1.4)$$

where $\zeta^{(i)}$ is an independent random variable that captures the randomness of the inner stage simulation, and is identically distributed for each scenario $\omega^{(i)}$. In other words, $L(\omega^{(i)})$ can be approximated by a quantity $\hat{L}(\omega^{(i)}, \zeta^{(i)})$ that is the sample average of $m$ independent and identically distributed samples of the discounted cashflows. Then we have the standard nested simulation estimator defined by

$$\hat{\alpha}_{SN(m,n)} \triangleq \frac{1}{n} \sum_{i=1}^{n} f(\hat{L}(\omega^{(i)}, \zeta^{(i)})). \quad (1.5)$$

This procedure is illustrated in Figure 1.2.

Recall that when analytical formulas are available for the portfolio re-valuation step, a primary challenge of a single-level Monte Carlo simulation is to reduce the variance of the simulation risk estimator. However, in the nested setting, simulation is also used in the inner stage for the portfolio re-valuation step and additional sources of variability are introduced. Besides, the function $f$ is generally non-linear, and then the inner stage simulation introduces bias into the computation.
CHAPTER 1. INTRODUCTION

Figure 1.2: Illustration of two-stage sampling. The outer stage generates $n$ financial scenarios $\omega^{(1)}, \ldots, \omega^{(n)}$. Conditional on each scenario $\omega^{(i)}$, $m$ inner stage samples are generated, which determine the cashflows of the portfolio from time $\tau$ to time $T$. Notice that the outer scenarios are generated according to the real-world distribution, and the inner stage samples are generated according to the risk-neutral distribution.

Hence, both bias and variance need to be balanced in order to minimize the total error in the simulation risk estimate.

The problem of estimating risk measures via nested simulation has been studied by a number of authors, e.g., Lee (1998), Lee and Glynn (2003), Gordy and Juneja (2008, 2010), and Hong and Juneja (2009). These authors primarily consider and analyze standard (i.e., uniform) nested simulation estimators. Such estimators employ a constant number of inner samples across portfolio re-valuation calculations, thus allocating computational effort uniformly across all scenarios. They demonstrate that, asymptotically, the bias of a uniform estimator is a function of the number of inner samples used in each portfolio re-valuation, while the variance of a uniform estimator is a function of the number of outer scenarios. They characterize the asymptotically optimal uniform estimator. This estimator balances a limited computational budget between using many outer scenarios, so as to lower variance, and using many inner samples in each scenario, so as to lower bias, in a way that minimizes the overall mean squared error (MSE) among the class of uniform estimators.

In this thesis, we consider the problem where portfolio losses are analytically unavailable, and we have to estimate them by nested simulation. We propose different methods to numerically compute the risk measures defined by (1.1), and investigate the trade-off between the bias and the variance introduced by nested simulation. In the first method, instead of sampling uniformly, we develop a strategy that samples non-uniform numbers of inner stage samples at each outer stage
scenario. Specifically, we put more inner stage samples at scenarios that have a greater marginal impact on the overall risk calculation. This strategy of allocating inner stage samples proves to be more efficient since the non-uniform distribution of inner stage samples leads to lower bias. In the second method, we propose a risk estimation procedure using unweighted regression. This tailored regression method estimates portfolio losses by using spatial information of nearby scenarios, and then fits the curve of portfolio losses globally. In the third method, we propose a weighted regression method, which is similar to the unweighted regression method but emphasizes scenarios that are more important to the calculation of the risk measure.

1.3 The Sequential Method

In Chapter 2, we will consider what is perhaps the most basic risk measure, the probability of a large loss; that is, given a threshold \( c \in \mathbb{R} \), we are interested in estimating the probability of the loss \( L \) exceeding \( c \). Denote the resulting probability by

\[
\alpha \triangleq \mathbb{E} [I_{\{L(\omega) \geq c\}}] = P(L(\omega) \geq c).
\]

The problem of estimating the probability of a loss via uniform nested simulation was first analyzed by Lee (1998) and Lee and Glynn (2003), and was subsequently considered by Gordy and Juneja (2008, 2010). Our algorithm seeks to exploit the fact that accurate portfolio re-valuation is not equally important across all scenarios. Nested simulation can be made much more efficient by allocating computational effort non-uniformly across scenarios. Non-uniform estimators have been suggested previously by others in a number of contexts (e.g., Lee and Glynn (2003); Lesnevski et al. (2004, 2007); Lan et al. (2008); Gordy and Juneja (2008)). Here, we propose and analyze a novel class of non-uniform estimators based on the idea of allocating additional effort to scenarios with a greater expected marginal change to the risk measure. Specifically, the main contributions are as follows:

1. We propose a non-uniform nested simulation algorithm for estimating the probability of a loss.

Our algorithm proceeds by allocating the inner stage samples for portfolio re-valuation in a sequential fashion. At each time step, it myopically selects the scenario where one additional
inner stage sample will have the greatest marginal impact to the estimated loss probability. This algorithm is simple to implement and incurs minimal computational overhead.

2. We provide an analysis that demonstrates the lower asymptotic bias of our approach.

Given \( m \) inner stage samples in each scenario, a uniform nested estimator has an asymptotic bias of order \( m^{-1} \). We analyze a simplified variation of our non-uniform estimator, and demonstrate that with an average of \( \bar{m} \) inner stage samples per scenario, the asymptotic bias is of order \( \bar{m}^{-2+\epsilon} \), for all positive \( \epsilon \). Hence, for the same overall number of samples, the non-uniform estimator reduces bias by an order of magnitude. This theoretical analysis builds on ideas from sequential hypothesis testing, and highlights the relationship between our non-uniform estimation algorithm and classical sequential hypothesis testing.

3. We provide an analysis that demonstrates the lower asymptotic MSE of our approach.

Given a computational budget of \( k \), the optimal uniform nested estimator results in an asymptotic MSE of order \( k^{-2/3} \). Since non-uniform sampling provides a lower bias for the same number of inner stage samples, some of this computational savings can be used for the generation of additional outer scenarios to lower variance. We show that our non-uniform method has an asymptotic MSE of order \( k^{-4/5+\epsilon} \), for all positive \( \epsilon \). Further, we demonstrate a practical implementation of our non-uniform estimator that adaptively balances bias (inner sampling) and variance (outer scenario generation).

4. We demonstrate the practical benefits of our method via numerical experiments.

Numerical experiments demonstrate that the performance of our non-uniform nested estimation algorithm is up to two orders of magnitude better than competing methods. Hence, we illustrate that the results achievable in practice are consistent with the gains suggested by the theory.

1.4 The Unweighted Regression Method

In Chapter 3, we present a risk estimation procedure based on unweighted regression. This method is applicable to a large category of risk measures. Since the prices of securities are not analytically available in closed form, instead of taking the sample average, we use regression, which provides
a global approximation to the portfolio loss function. Specifically, the main contributions are as follows:

1. **We propose a method for nested simulation based on regression.**

   The method proceeds as in standard nested simulation, but it uses a linear combination of regression basis functions, instead of sample averages, to more accurately approximate portfolio losses. Since the entire distribution of portfolio losses is estimated by this method, it can be easily applied to compute a large range of risk measures.

2. **We provide a theoretical analysis of the performance of our method.**

   We analyze the asymptotic behavior of the regression-based risk estimator. The analysis shows that the MSE of the regression-based risk estimator converges at the rate $n^{-1+\delta}$, for all positive $\delta$, where $n$ is the number of outer stage scenarios, plus a non-diminishing bias term that is determined by the magnitude of the regression error. The performance of the regression estimator does not depend on the dimension of a problem except through the quality of the regression basis functions. With well-chosen basis functions, the regression method performs significantly better than the standard nested simulation method.

3. **The theoretical analysis shows the optimal tradeoff between inner and outer sampling in the regression method.**

   Given a fixed number of inner stage samples $k = mn$, our analysis shows that the asymptotically minimum mean squared error is achieved when $m = 1$ and $n = k$ in the regression method. That is, the regression method works best with one inner stage sample per outer stage scenario. With this optimal choice, the regression method recovers the $k^{-1}$ convergence rate of non-nested simulation, until the MSE reaches an asymptotic bias level determined by the regression model error.

4. **We provide numerical examples that compare the performance of the regression method with other methods.**

   We compare the mean squared error of the regression estimator with other methods used in the literature and in practice. Numerical results are provided that are consistent with
the theoretical analysis and demonstrate the advantage of the regression method over other approaches.

1.5 The Weighted Regression Method

In Chapter 4, we examine a method via weighted regression to estimate risk measures. This proposed weighted regression method estimates portfolio losses by using spatial information of scenarios and emphasizing scenarios more important to the calculation of the risk measure. The main contributions are as follows:

1. *We propose a method via weighted regression to estimate risk measures.*

   Beginning with the unweighted regression method proposed in Chapter 3, we present a weighted regression method that assigns more weight to scenarios that are more important to the calculation of the risk measure. Due to the nonlinearity of the function $f$ in the category of risk measures defined by (1.1), outer stage scenarios are not equally importantly involved in risk measurement. Therefore, the weighted regression method is reasonable since larger weights are assigned to scenarios that deserve more precise estimation.

2. *We provide a heuristic analysis that leads to a globally optimal weighting strategy.*

   When the computational budget is large enough, the MSE of the regression estimator will be dominated by the asymptotic bias. Based on a heuristic analysis, we give an explicit choice of weight function that leads to the global minimum of an upper bound of the squared bias.

3. *We suggest a practically implementable weighting strategy.*

   Our globally optimal weight function depends on knowledge of the portfolio loss in each scenario, which is unobservable in practice. To remedy this, we propose a two-pass weighting strategy, where the first pass is an unweighted regression, and the second pass is a weighted regression with the weight function based on the first pass. This two-pass procedure does not rely on knowledge of the unobservable portfolio loss, and is close to the one derived from analysis when the computational budget is large.

4. *We provide numerical examples to illustrate the practical benefits of our method.*
In our numerical examples, the two-pass weighted regression method yields an estimator that is superior to standard nested simulation. It performs comparably to unweighted regression with a small computational budget, and has better performance when the computational budget is large.

1.6 Literature Review

Overviews of financial risk measurement and management are given in Crouhy et al. (2000), Jorion (2006), and McNeil et al. (2006). There is a large literature on the properties of alternative risk measures (see, e.g., Artzner et al. (1999); Rockafellar and Uryasev (2002); Föllmer and Schied (2002)). Variance reduction techniques to improve single stage estimation are given in Glasserman et al. (2000, 2002).

Most closely related to our sequential method for nested simulation is that of Lee (1998) and Lee and Glynn (2003), who consider the problem of estimating the probability of a large loss and analyze nested simulation estimators and their convergence properties under uniform inner stage sampling. They consider two settings, where the underlying scenario space is either continuous or discrete. They establish that, given a total computational budget of $k$, the optimal uniform nested estimator results on an asymptotic MSE of order $k^{-2/3}$ in the continuous case and $k^{-1} \log k$ in the discrete case. Independently, in a pair of papers, Gordy and Juneja (2008, 2010) also consider estimating the probability of large loss in the continuous case, under a different set of assumptions. They also consider two additional risk measures (value at risk and expected shortfall). For each of these three risk measures, they derive asymptotic bias and variance results for uniform inner stage sampling. This allows them to derive the optimal allocation of effort between outer and inner stage sampling and derive the optimal asymptotic MSE of order $k^{-2/3}$. They also propose a jackknife procedure for reducing bias.

The idea of non-uniform nested estimation of risk measures dates back to at least the work of Lee and Glynn (2003). In the discrete case, they identify a class of non-uniform nested estimators for the probability of a large loss with asymptotic MSE of order $k^{-1} \log k$. In this setting, the

---

3 In this thesis, we will consider only continuous scenario spaces. Note that the theory is qualitatively different in the discrete case versus the continuous case.
non-uniform estimator achieves the same asymptotic convergence rate as the uniform estimator, but with a better constant. For each outer stage scenario, they use a specific amount of inner stage samples, which is a function of the outer stage scenario. More discussion of Lee and Glynn (2003) is given in Section 2.2. Lesnevski et al. (2004, 2007) propose a non-uniform nested estimator for a related discrete problem: they estimate the worst case expected loss across a finite set of scenarios. They are able to develop confidence intervals for their estimation procedure. Frye (1998) and Shaw (1997) propose an algorithm that generates no inner stage samples at a certain set of outer stage scenarios so as to reduce the computational cost. Lan et al. (2007, 2008), Liu and Staum (2010), and Liu et al. (2010) extend this work to the specific case of estimating expected shortfall, and their method dynamically allocates the computational budget to scenarios that are likely to lead to large losses. Compared with standard nested simulation, this procedure achieves lower MSE in numerical experiments. However, no theoretical proof of better efficiency is developed. Contemporaneous with the present work, Gordy and Juneja (2008) suggest a broad class of non-uniform estimators for estimating the probability of a large loss, as in the present setting. Their method dynamically allocates more inner stage samples to scenarios that are likely to have portfolio losses close to the threshold $c$ in (1.2). Their description is rather general, while we provide a concrete algorithm with theoretical support, which covers the idea in Gordy and Juneja (2008) but also reflects more characters of a good dynamic allocation.

Note that some of the non-uniform estimators in this prior literature have similarities to the non-uniform estimator that we propose; we discuss these in Section 2.2. Critically, however, none of this prior work is able to establish theoretically that a non-uniform estimator converges at a faster asymptotic order than is possible with uniform estimators.

There are some connections between nested simulation to estimate risk and ranking and selection (R&S) procedures, which search for the best among a finite number of systems. For an overview of ranking and selection, see Kim and Nelson (2005) and the book by Chen and Lee (2010). Each R&S system corresponds to an outer sample, and sampling a performance measure from a system corresponds to an inner sample. Many R&S procedures rely on myopic rules to determine an allocation of inner samples (e.g., Frazier et al. (2008)) and the spirit of our sequential procedure is similar: our method sequentially allocates inner stage samples to scenarios with a greater expected marginal change to the risk measure. R&S typically considers a finite and small number of systems,
whereas our outer sampling draws from an infinite and often multi-dimensional domain. The R&S objective of finding the best performing system is also different than our objective of minimizing the asymptotic MSE of the risk estimator across the range of outer stage scenarios. Moreover, the algorithm proposed by our method in Chapter 2 which is easily implementable, has a fairly clear motivation that reflects fundamental characters of the risk estimation problem.

Sun et al. (2009) consider nested simulation in the context of estimating conditional variance. Specifically they consider the problem of estimating the variance of conditional expectation of portfolio losses in scenarios, and show that as the computational budget increases, the optimal number of inner stage samples approaches a finite limit.

Another important approach is the delta-gamma approximation, which is illustrated in, e.g., Rouvinez (1997), Britten-Jones and Schaefer (1999), and Duffie and Pan (2001). This method uses quadratic approximation, i.e., the second order Taylor approximation, to model the portfolio value. Exact calculation (see, e.g., Rouvinez (1997)) or Monte Carlo sampling of risk factors (see, e.g., Glasserman et al. (2000)) is usually combined with the delta-gamma approximation. Neither of these two techniques needs to generate any inner stage samples, and with these techniques the delta-gamma approximation works faster than nested simulation. However, due to the simple form of quadratic functions, the delta-gamma approximation generally will not converge to the true curve.

A standard reference for regression analysis is White (2001). Hong and Juneja (2009) discuss the benefit of using kernel smoothing to estimate the risk measures in the form of (1.1). Compared with standard nested simulation, the combination of the nested simulation and kernel smoothing works better when the dimension of the problem is less than or equal to three while it performs worse otherwise. To achieve the minimum MSE, they suggest \( m = 1 \) as the optimal setting, i.e., there are \( k \) outer stage scenarios and 1 inner stage sample at each scenario. Another recent method is given in Liu and Staum (2010), where they estimate expected shortfall, and explore an approach based on stochastic kriging described in, for instance, Ankenman et al. (2010). Stochastic kriging builds a metamodel that interpolates portfolio values based on nearby outer stage scenarios. The methods in Hong and Juneja (2009) and Liu and Staum (2010) use local information from nearby scenarios.

If a method relies on information from nearby scenarios, its performance on higher-dimensional
problems is suspect. However, the convergence rate of our regression method does not depend on the dimension of a problem, i.e., the number of risk factors considered. Meanwhile, features of portfolios can be captured by regression if the quality of the basis functions are good. Therefore, it is fairly straightforward to apply our analysis and procedure to realistic higher-dimensional problems.
Chapter 2

The Sequential Method

This chapter focuses on developing a non-uniform nested simulation method that achieves a higher order of convergence. The algorithm and analysis are developed for one specific risk measure, the probability of a large loss.

2.1 Optimal Uniform

The Uniform estimator (see Algorithm 1) is a nested simulation procedure that combines the estimates from the outer and inner levels of simulation in the obvious way to produce an overall estimate of the loss probability. We say that this estimator samples uniformly in the sense that a constant number of inner stage samples are used for each outer stage scenario.

The Uniform estimator is a function of two parameters: \( n \), the number of outer stage scenarios, and \( m \), the number of inner stage samples per outer stage scenario. This raises an obvious question: what are the best choices for the parameters \( m \) and \( n \)? This question has been addressed in the work of Lee (1998) and Gordy and Juneja (2008, 2010). We follow the latter approach.

Denote the Uniform estimate of the probability of a large loss by

\[
\hat{\alpha}_{SN(m,n)} \equiv \text{Uniform}(m,n).
\]

The obvious objective is to choose parameters \((m,n)\) so as to minimize the mean squared error (MSE) of the estimator \(\hat{\alpha}_{SN(m,n)}\), subject to the constraint of a limited budget of computational resources. The Uniform estimator involves outer scenario generation and inner sampling. We will
CHAPTER 2. THE SEQUENTIAL METHOD

Algorithm 1: Estimate the probability of a large loss using a uniform nested simulation. The parameter $m$ is the number of inner samples per scenario. The parameter $n$ is the number of outer scenarios.
CHAPTER 2. THE SEQUENTIAL METHOD

make the assumption that the computational effort of this estimator is dominated by the latter\footnote{This will typically be true since the risk horizon \( \tau \) is often short relative to the time horizon \( T \) of realized cashflows.}

Given parameters \((m, n)\), a total of \( mn \) inner samples are generated in order to compute the estimate \( \hat{\alpha}_{SN(m,n)} \). Thus, given a computational work budget \( k \) on the total number of inner samples, we have the optimization problem:

\[
\begin{align*}
\text{minimize} & \quad E \left[ (\hat{\alpha}_{SN(m,n)} - \alpha)^2 \right] \\
\text{subject to} & \quad mn \leq k, \\
& \quad m, n \geq 0. \quad (2.1)
\end{align*}
\]

The mean squared error objective can be decomposed into variance and bias terms according to

\[
E \left[ (\hat{\alpha}_{SN(m,n)} - \alpha)^2 \right] = E \left[ (\hat{\alpha}_{SN(m,n)} - E[\hat{\alpha}_{SN(m,n)}])^2 \right] + \left( E[\hat{\alpha}_{SN(m,n)}] - \alpha \right)^2. \quad (2.2)
\]

In order to analyze the asymptotic behavior of the MSE, we need the following technical assumption\footnote{We follow the assumptions in Gordy and Juneja (2008, 2010). For an alternative set of assumptions, see Lee (1998).}.

**Assumption 2.1.** Denote by \( L(\omega) \) the portfolio loss in scenario \( \omega \) at time \( \tau \), and denote by \( \hat{L} \) an estimator of the form (1.4) for \( L(\omega) \), based on the average of \( m \) i.i.d. inner stage samples. Assume that

1. The joint probability density function \( p_m(\ell, \hat{\ell}) \) of \((L, \hat{L})\) and its partial derivatives \( \frac{\partial}{\partial \ell} p_m(\ell, \hat{\ell}) \) and \( \frac{\partial^2}{\partial \ell^2} p_m(\ell, \hat{\ell}) \) exist for each \( m \) and \((\ell, \hat{\ell})\).

2. For each \( m \geq 1 \), there exist functions \( f_{0,m}(\cdot) \), \( f_{1,m}(\cdot) \), and \( f_{2,m}(\cdot) \) so that

\[
p_m(\ell, \hat{\ell}) \leq f_{0,m}(\hat{\ell}), \quad \left| \frac{\partial}{\partial \ell} p_m(\ell, \hat{\ell}) \right| \leq f_{1,m}(\hat{\ell}), \quad \left| \frac{\partial^2}{\partial \ell^2} p_m(\ell, \hat{\ell}) \right| \leq f_{2,m}(\hat{\ell}),
\]

for all \((\ell, \hat{\ell})\). Further,

\[
\sup_m \int_{-\infty}^{\infty} |\hat{\ell}|^r f_{i,m}(\hat{\ell}) d\hat{\ell} < \infty, \quad \text{for all } i = 0, 1, 2, \text{ and } 0 \leq r \leq 4.
\]
Gordy and Juneja (2008, 2010) establish the following:

**Theorem 2.1.** Suppose that Assumption 2.1 holds, and denote by \( f(\cdot) \) the density of the loss variable \( L \). As \( m \to \infty \), the bias of the Uniform estimator asymptotically satisfies

\[
E \left[ \hat{\alpha}_{SN(m,n)} - \alpha \right] = \frac{\theta_c}{m} + O \left( m^{-3/2} \right),
\]

where

\[
\theta_c \triangleq -\Upsilon'(c), \quad \Upsilon(c) \triangleq \frac{1}{2} f(c) E \left[ \sigma^2(\omega) \mid L(\omega) = c \right],
\]  
(2.3)

and \( \sigma^2(\omega) \) is the variance of the inner stage samples in scenario \( \omega \).

Theorem 2.1 directly provides an asymptotic analysis of the bias term in the MSE (2.2). Theorem 2.1 can immediately be employed to analyze the variance term, as in the following corollary:

**Corollary 2.1.** Under the conditions of Theorem 2.1, as \( m \to \infty \), the variance of the Uniform estimator satisfies

\[
\text{Var} \left( \hat{\alpha}_{SN(m,n)} \right) = \frac{\alpha(1-\alpha)}{n} + O \left( m^{-1}n^{-1} \right).
\]

**Proof.** Note that

\[
\text{Var} \left( \hat{\alpha}_{SN(m,n)} \right) = \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} I \left\{ L(\omega^{(i)},\zeta^{(i)}) \geq c \right\} \right)
\]

\[
= \frac{1}{n} \text{Var} \left( I \left\{ L(\omega^{(i)},\zeta^{(i)}) \geq c \right\} \right)
\]

\[
= \frac{E \left[ \hat{\alpha}_{SN(m,n)} \right] \left( 1 - E \left[ \hat{\alpha}_{SN(m,n)} \right] \right)}{n},
\]

where we have used the fact that the loss estimates \( \{ \hat{L}(\omega^{(i)},\zeta^{(i)}) \} \) are independent and identically distributed. Applying Theorem 2.1,

\[
\text{Var} \left( \hat{\alpha}_{SN(m,n)} \right) = \frac{\alpha(1-\alpha)}{n} + \frac{E \left[ \hat{\alpha}_{SN(m,n)} - \alpha \right] \left( 1 - E \left[ \hat{\alpha}_{SN(m,n)} \right] \right)}{n} + \frac{\alpha E \left[ \alpha - \hat{\alpha}_{SN(m,n)} \right]}{n}
\]

\[
= \frac{\alpha(1-\alpha)}{n} + O \left( m^{-1}n^{-1} \right).
\]

---

3 In what follows, given arbitrary sequences \( \{ f_N \} \) and \( \{ g_N \} \), and a positive sequence \( \{ q_N \} \), as \( N \to \infty \), we will say that \( f_N = g_N + O(q_N) \) if \( \limsup_{N \to \infty} | f_N - g_N | / q_N < \infty \), i.e., if the difference between \( f \) and \( g \) is asymptotically bounded above by some constant multiple of \( q \). Similarly, we will say that \( f_N = g_N + o(q_N) \) if \( \limsup_{N \to \infty} | f_N - g_N | / q_N = 0 \), i.e., if the difference between \( f \) and \( g \) is asymptotically dominated by every constant multiple of \( q \). Finally, we will say that \( f_N = g_N + \Theta(q_N) \) if \( 0 < \liminf_{N \to \infty} | f_N - g_N | / q_N \leq \limsup_{N \to \infty} | f_N - g_N | / q_N < \infty \), i.e., if the difference between \( f \) and \( g \) is asymptotically bounded above and below by constant multiples of \( q \).
Theorem 2.1 and Corollary 2.1 provide a complete asymptotic characterization of the MSE of the Uniform estimator. The asymptotic variance of the estimator is determined by the number of scenarios $n$ and decays as $n^{-1}$, while the asymptotic bias of the estimator is determined by the number of inner stage samples per scenario $m$ and decays as $m^{-1}$.

Given a computational budget of a total of $k$ inner stage samples, a naive choice of parameters $(m, n)$ might be to sample equally in the outer and inner stages, i.e., set $m = n = k^{1/2}$. This would result in an asymptotic squared bias of order $k^{-1}$ and an asymptotic variance of order $k^{-1/2}$, and an overall asymptotic MSE of order $k^{-1/2}$. Since the variance is asymptotically dominating the squared bias and determining the MSE, the naive Uniform estimator is clearly not optimal.

One could do better by using fewer inner stage samples per scenario and increasing the number of scenarios.

In order to find the optimal Uniform estimator, using Theorem 2.1 and Corollary 2.1, we can approximate the minimum MSE problem (2.1) by the optimization problem

$$
\begin{align*}
\text{minimize} & \quad \frac{\alpha (1 - \alpha)}{n} + \frac{\theta_c^2}{m^2} \\
\text{subject to} & \quad mn \leq k, \\
& \quad m, n \geq 0.
\end{align*}
$$

This suggests optimal allocations

$$
\begin{align*}
m^* &= k^{1/3} / \beta^*, \\
n^* &= \beta^* k^{2/3}, \\
\beta^* &\triangleq \left( \frac{\alpha (1 - \alpha)}{2 \theta_c^2} \right)^{1/3},
\end{align*}
$$

and the optimal asymptotic mean squared error

$$
\mathbb{E} \left[ (\hat{\alpha}_{SN}(m,n) - \alpha)^2 \right] = 3 (\beta^*)^2 k^{-2/3} + o(k^{-2/3}).
$$

The optimal allocations suggested by (2.4) involve, asymptotically, order $k^{2/3}$ outer stage scenarios and order $k^{1/3}$ inner stage samples per scenario.

**Remark 2.1.** Although we know the asymptotic orders of magnitudes of $m^*$ and $n^*$, the optimal constant factors depend on the constant $\theta_c$ and it is not clear how to effectively estimate $\theta_c$ a priori. As we will see in Section 2.5, the choice of this constant factor is critical to the practical performance of a uniform estimator. Therefore, in general cases, it is not clear how to achieve the optimal performance of $\hat{\alpha}_{SN}(m,n)$ given a finite $k$. 


Finally, it is instructive to compare the rate of convergence of the optimal Uniform estimator in a two-level nested Monte Carlo simulation to that of an estimator of the probability of a large loss in a single-level Monte Carlo simulation. In the latter case, scenarios \( \omega^{(1)}, \ldots, \omega^{(n)} \) are generated. It is assumed that in each scenario \( \omega^{(i)} \), the loss \( L(\omega^{(i)}) \) can be exactly computed, and the probability is estimated via (1.3). Note that the estimator (1.3) is unbiased, and has a variance proportional to \( n^{-1} \). In a single-level simulation, then, the amount of work is proportional to \( n \), while the MSE of the estimator decays proportional to \( n^{-1} \). In a two-level simulation, however, as shown above, the amount of work is proportional to \( k \), while the MSE decays at best at a rate of \( k^{-2/3} \). This slower rate of decay is due to the bias introduced by the inner level of simulation.

### 2.2 Sequential Sampling

The Uniform estimator described in Sections 2.1 employs a constant number of inner stage samples for each outer stage scenario. It is intuitively clear to see that this may be an inefficient strategy. As an illustrative example, consider the situation depicted in Figure 2.1. Here, we wish to estimate the loss probability associated with the shaded region. There are two outer stage scenarios \( \omega^{(1)} \) and \( \omega^{(2)} \), associated with the portfolio losses \( L(\omega^{(1)}) \) and \( L(\omega^{(2)}) \) respectively. These true losses are approximated, in each scenario, by the estimated losses \( \hat{L}(\omega^{(1)}, \zeta^{(1)}) \) and \( \hat{L}(\omega^{(2)}, \zeta^{(2)}) \).

Suppose that, under a uniform nested simulation, the portfolio losses estimated in each scenario are distributed according to the dashed probability distributions. Then, it is clear that it would be advantageous to employ fewer inner stage samples at scenario \( \omega^{(1)} \), and more inner stage samples at scenario \( \omega^{(2)} \). This is because the loss probability estimate \( \hat{\alpha} \) is calculated according to

\[
\hat{\alpha} \triangleq \frac{1}{n} \sum_{i=1}^{n} I\{\hat{L}(\omega^{(i)}, \zeta^{(i)}) \geq c\}. \tag{2.6}
\]

Thus, only the ordinal position of the estimates \( \hat{L}(\omega^{(1)}, \zeta^{(1)}) \) and \( \hat{L}(\omega^{(2)}, \zeta^{(2)}) \) relative to the loss threshold \( c \) is relevant. Given the uncertainty in the estimate \( \hat{L}(\omega^{(1)}, \zeta^{(1)}) \), it is fairly certain that \( L(\omega^{(1)}) < c \), and, indeed, this could likely be inferred using fewer inner samples in scenario \( \omega^{(1)} \). Given the uncertainty in the estimate \( \hat{L}(\omega^{(2)}, \zeta^{(2)}) \), the fact that \( L(\omega^{(2)}) \geq c \), on the other hand, is much less certain. Without more inner samples in this scenario, there may be significant risk of misclassifying \( \omega^{(2)} \). These observations suggest that a non-uniform sampling strategy may be
CHAPTER 2. THE SEQUENTIAL METHOD

Figure 2.1: An illustration of the benefits of non-uniform sampling. The uncertainty in the loss \( \hat{L}(\omega^{(1)}, \zeta^{(1)}) \) estimated in scenario \( \omega^{(1)} \) is unlikely to impact the overall probability of large loss estimate. Hence, the number of inner samples \( m_1 \) in this scenario can be chosen to be small. In scenario \( \omega^{(2)} \), however, a large number of inner samples \( m_2 \) should be used.

superior: the number of inner samples \( m_1 \) employed at scenario \( \omega^{(1)} \) should be less than the number of inner samples \( m_2 \) employed at scenario \( \omega^{(2)} \).

The discussion above suggests that in a scenario \( \omega \) with a loss \( L(\omega) \) that is much greater than \( c \) or much less than \( c \), few inner samples are necessary. If the loss \( L(\omega) \) is close to \( c \), however, many inner samples are necessary. Unfortunately, a priori, it is not clear how to do this. It is impossible to know the value of \( L(\omega) \) — this is exactly what we seek to estimate via the inner Monte Carlo simulation.

We propose a procedure that simultaneously maintains estimates of the loss in each scenario, while sequentially attempting to allocate additional inner samples across the outer scenarios. We will first motivate our algorithm with an informal justification, and then give a precise description. In particular, suppose that there are \( n \) scenarios \( \omega^{(1)}, \ldots, \omega^{(n)} \). For each scenario \( \omega^{(i)} \), suppose that \( m_i \) inner samples \( \hat{Z}_{i,1}, \ldots, \hat{Z}_{i,m_i} \) have been made, resulting in the loss estimate

\[
\hat{L}(\omega^{(i)}, \zeta^{(i)}) \equiv \frac{1}{m_i} \sum_{j=1}^{m_i} \hat{Z}_{i,j}.
\]

This results in an overall probability of a large loss estimate \( \hat{\alpha} \) given by (2.6).

Without loss of generality, assume that \( \hat{L}(\omega^{(i)}, \zeta^{(i)}) \geq c \). Suppose we wish to perform one
additional inner stage sample. If we were to perform the additional sample in scenario $\omega^{(i)}$, this would result in a new loss estimate given by

$$
\hat{L}'(\omega^{(i)}, \zeta^{(i)}) \triangleq \frac{1}{m_i + 1} \sum_{j=1}^{m_i + 1} Z_{i,j} = \frac{1}{m_i + 1} \hat{Z}_{i,m_i+1} + \frac{m_i}{m_i + 1} \hat{L}(\omega^{(i)}, \zeta^{(i)}).
$$

The additional sample will only impact the estimate $\hat{\alpha}$ if the $\hat{L}(\omega^{(i)}, \zeta^{(i)})$ is on the opposite side of the threshold level $c$ than $\hat{L}'(\omega^{(i)}, \zeta^{(i)})$, i.e., if $\hat{L}'(\omega^{(i)}, \zeta^{(i)}) < c$. This is illustrated in Figure 2.2.

In order to myopically maximize the impact of the single additional sample, we will seek to choose the scenario $\omega^{(i)}$ that maximizes the probability of such a sign change. Suppose that the additional sample $\hat{Z}_{i,m_i+1}$ has variance $\sigma_i^2 \triangleq \sigma^2(\omega^{(i)})$. Observe that

$$
P \left( \hat{L}'(\omega^{(i)}, \zeta^{(i)}) < c \right) = P \left( \hat{Z}_{i,m_i+1} - L(\omega^{(i)}) < -m_i \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right) - (L(\omega^{(i)}) - c)
\approx P \left( \hat{Z}_{i,m_i+1} - L(\omega^{(i)}) < -m_i \left| \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right| \right)
\leq \left( 1 + \frac{m_i^2}{\sigma_i^2} \left| \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right|^2 \right)^{-1}. \tag{2.7}
$$

Here, the approximation follows from the assumption that $m_i \gg 1$, so that $-m_i \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \approx -m_i \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c$. The inequality follows from the one-sided Chebyshev inequality. By analogous consideration of the symmetric case (where $\hat{L}(\omega^{(i)}, \zeta^{(i)}) < c$), a myopic allocation rule that seeks to maximize the probability of a sign change estimated via the Chebyshev bound (2.7) will choose to add the additional inner sample in scenario $\omega^{(i*)}$ where

$$
i^* \in \arg\min_i \frac{m_i}{\sigma_i} \left| \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right|. \tag{2.8}
$$

An alternative justification for the myopic rule (2.8) arises if the additional sample $\hat{Z}_{i,m_i+1}$ is drawn from a location-scale family of distributions, e.g., if $\hat{Z}_{i,m_i+1}$ is normally distributed. Such a distribution is specified by a mean $L(\omega^{(i)})$ and a variance $\sigma_i^2$, so that

$$
P \left( \hat{Z}_{i,m_i+1} < z \right) = G \left( \frac{z - L(\omega^{(i)})}{\sigma_i} \right),
$$

where $G$ is an increasing function. In this case,

$$
P \left( \hat{L}'(\omega^{(i)}, \zeta^{(i)}) < c \right) \approx P \left( \hat{Z}_{i,m_i+1} - L(\omega^{(i)}) < -m_i \left| \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right| \right)
= G \left( -\frac{m_i}{\sigma_i} \left| \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right| \right). \tag{2.9}
$$
CHAPTER 2. THE SEQUENTIAL METHOD

Maximizing the probability of a sign change according to (2.9) also results in the myopic rule (2.8).

We call the quantity minimized in (2.8), \( (m_i/\sigma_i)|\hat{L}(\omega^{(i)}, \zeta^{(i)}) - c| \), the error margin associated with the scenario \( \omega^{(i)} \). The allocation rule (2.8), which picks a scenario by greedily minimizing the error margin, makes intuitive sense qualitatively. It encourages additional inner samples at scenarios that are close to the loss boundary (i.e., \( |\hat{L}(\omega^{(i)}, \zeta^{(i)}) - c| \) is small), scenarios with few inner samples (i.e., \( m_i \) is small), or scenarios with significant variability in the portfolio losses (i.e., \( \sigma_i \) is large). The Sequential estimator of Algorithm 2 employs the allocation rule (2.8). This estimator takes a triple \((m^0, \bar{m}, n)\) of input parameters. Here, \( n \) is the desired number of outer stage scenarios, \( m^0 \) is the initial number of inner stage samples per scenario, and \( \bar{m} \) is the desired average number of inner stages samples per scenario at the conclusion of the algorithm. The algorithm proceeds as follows: first, \( n \) scenarios are generated, and, for each scenario, \( m^0 \) inner stage samples are performed. The remaining \( \bar{m}n - m^0n \) inner stage samples are allocated one at a time in a sequential fashion myopically, as in (2.8).

Note that the Sequential estimator requires access to the conditional standard deviation \( \sigma_i^2 \) of losses in each scenario \( \omega^{(i)} \), in order to compute the error margin. These are not required for the Uniform estimator and, moreover, are typically not known in practice. However, these conditional standard deviations can be estimated in an online fashion over the course of the estimation.

Figure 2.2: An additional inner sample in scenario \( \omega^{(i)} \) will only change the overall probability of loss estimate if \( \hat{L}(\omega^{(i)}, \zeta^{(i)}) \) changes sign.
CHAPTER 2. THE SEQUENTIAL METHOD

1: procedure Sequential($m^0$, $\tilde{m}$, $n$)  
2:     for $i$ ← 1 to $n$ do  
3:         generate scenario $\omega^{(i)}$  
4:         conditioned on scenario $\omega^{(i)}$, generate i.i.d. samples $\hat{Z}_{i,1}, \ldots, \hat{Z}_{i,m^0}$ of portfolio losses  
5:         $m_i$ ← $m^0$  
6:     end for  
7:     while $\sum_{i=1}^{n} m_i < \tilde{m}n$ do  
8:         set  
9:         
10:         
11:     end while  
12:     compute an estimate of the probability of a large loss  
13:         $\hat{\alpha} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\hat{L}(\omega^{(i)}, \zeta^{(i)}) \geq c\}$  
14:     return $\hat{\alpha}$  

Algorithm 2: Estimate the probability of a large loss using a sequential non-uniform nested simulation.  
The parameter $m^0$ is the initial number of inner samples per scenario. The parameter $\tilde{m}$ is the average  
number of inner samples per scenario at the conclusion of the simulation. The parameter $n$ is the  
number of scenarios.
Further, the Sequential estimator requires additional computational overhead beyond that of the Uniform estimator. However, this is minimal: the only additional requirement is to track scenarios in order of error margin. This can be accomplished efficiently via a priority queue data structure (see, e.g., Cormen et al. (2002)). With a priority queue, determining the scenario with minimum error margin (line 8 in Algorithm 2) can be accomplished in constant time (i.e., in an amount of time independent of $m$ and $n$). Once a new inner sample is generated for a scenario (lines 9-10 in Algorithm 2), order $\log n$ time would be required to update the priority queue data structure. In practice, this is not significant.

The Sequential estimator also requires more memory than the Uniform estimator. In particular, the Uniform estimator can be implemented in a way where scenarios are processed one-at-a-time and never need to be simultaneously stored in memory. Such an implementation would have a constant memory requirement (i.e., independent of $m$ and $n$). For the Sequential estimator, each of the $n$ outer scenarios must be stored in memory over the course of the algorithm, hence the memory requirement is of order $n$. In practice, even given a very large number of scenarios (e.g., millions), each of very high dimension (e.g., thousands), this memory requirement is well within the reach of commodity hardware. Each inner sample may require simulating multiple steps over a long time horizon, but the memory requirement is minimal since all intermediate computations are discarded and only the inner sample loss is recorded.

The Sequential estimator has some similarities to non-uniform estimators that have been proposed in the literature. Lee and Glynn (2003) suggest a non-uniform nested estimator in the case where the scenario space is discrete. They choose the number of inner samples $m_i$ in each scenario $\omega(i)$ so as to optimize certain large deviation asymptotics. Using a Gaussian approximation as a heuristic, this results in the allocation

$$m_i \propto \frac{\sigma_i^2}{(L(\omega(i)) - c)^2}.$$  \hspace{1cm} (2.10)

Since the loss $L(\omega(i))$ in scenario $\omega(i)$ is unknown, Lee and Glynn (2003) propose a two-pass algorithm: in the first pass, a small number of inner samples are generated in each scenario and are

---

4The non-uniform Threshold estimator that will be discussed in Section 2.3.1 does not require any additional computational or memory overhead beyond that of the standard Uniform estimator.
used to compute inner sample allocations in a second “production run.”

Our \textit{Sequential} estimator differs from (2.10) in several fundamental ways: first, the allocation (2.10) is loosely analogous to minimizing the square of the error margin, as opposed to the error margin itself. Second, the allocation (2.10) is accomplished with multiple passes, while our estimator is fully sequential. Indeed, in Section 2.3 tools from sequential analysis will prove fundamental in the theoretical analysis of our estimators. Finally, and most importantly, in the setting of Lee and Glynn (2003), non-uniform sampling does not provide a qualitatively different rate of convergence than uniform sampling. Given a total computational budget of order \( k \), both the uniform and non-uniform methods achieve an asymptotic MSE of order \( k^{-1} \log k \), albeit with different constants. As we shall see in Section 2.3, we will be able to establish theoretically that a non-uniform estimator converges at a faster asymptotic order than is possible with uniform estimators.

Gordy and Juneja (2008) suggest a general class of multi-pass “dynamic allocation” schemes for non-uniform nested estimation. Such schemes would, for example, divide the simulation into a sequence of \( J \) phases, where in the \( j \)th phase inner samples would only be allocated to scenarios \( \omega^{(i)} \) if \( \hat{L}(\omega^{(i)}, \zeta^{(i)}) \geq c - \epsilon_j \). Here, \( \epsilon_1 > \epsilon_2 > \cdots > \epsilon_J \) is a sequence of thresholds. Gordy and Juneja (2008) provide some numerical evidence that such schemes may provide a significant improvement over uniform estimators, but the choice of specific parameters of the algorithm (e.g., the number of phases \( J \) or the thresholds \( \{\epsilon_j\} \)) is left as a direction for future research.

### 2.3 Analysis

In Section 2.2 we introduced the non-uniform \textit{Sequential} estimator and motivated this algorithm via an informal discussion. In this section, we will provide an analysis of non-uniform estimation. We begin in Section 2.3.1 by introducing a simplified variation of the \textit{Sequential} estimator. This simplified estimator preserves the myopic and non-uniform behavior of the \textit{Sequential} estimator, but is more amenable to analysis. Moreover, the simplified estimator is reminiscent of a compound sequential hypothesis test, and highlights connections to the classical field of sequential analysis. In Section 2.3.2 we provide an asymptotic analysis of the bias and variance of simplified non-uniform estimator. Finally, in Section 2.3.3 we discuss optimal parameter choices for the simplified non-uniform estimator. We demonstrate that this estimator has an asymptotic MSE of order \( k^{-4/5+\epsilon} \),
for all positive $\epsilon$, as a function of the computational budget $k$. This can be compared to the asymptotic MSE of order $k^{-2/3}$ of the optimal uniform estimator.

### 2.3.1 A Simplified Non-Uniform Estimator

Analysis of the Sequential estimator described in Section 2.2 presents a number of challenges. Foremost among these is the fact that, over the course of the nested simulation of the Sequential estimator, the loss estimates $\hat{L}(\omega^{(1)}, \zeta^{(1)}), \ldots, \hat{L}(\omega^{(n)}, \zeta^{(n)})$ are dependent random variables. This dependence is induced by the myopic selection rule (2.8), which, at each point in time, simultaneously depends upon all of the loss estimates. In order to make the analysis tractable, we will consider a modification of the Sequential estimator, which results in independent loss estimates while maintaining the spirit of myopic non-uniform sampling.

In particular, recall that the Sequential estimator takes as input a parameter $\bar{m}$, specifying the desired average number of inner samples in each scenario, and a parameter $n$, specifying the desired number of scenarios. Over the course of the algorithm, $\bar{m}n$ total inner stage samples will be generated. These samples are allocated in a sequential fashion so as to myopically minimize the error margin $(m_i/\sigma_i)|\hat{L}(\omega^{(i)}, \zeta^{(i)}) - c|$, uniformly over $1 \leq i \leq n$.

If we imagine the algorithm to be in a state where a significant number of inner samples have been generated, i.e., $m_i \gg 1$ for each $i$, then one would expect the error margins to be roughly constant — if not, more inner samples would have been generated for the scenarios with lower error margins. One could achieve a similar effect by fixing a threshold $\gamma > 0$, and continuing to add inner stage samples to each scenario $\omega^{(i)}$ until the error margin exceeds $\gamma$, i.e.,

$$\frac{m_i}{\sigma_i} \left| \hat{L}(\omega^{(i)}, \zeta^{(i)}) - c \right| \geq \gamma.$$  \hspace{1cm} (2.11)

This is precisely what is done by the Threshold estimator of Algorithm 3.

At a high level, the Sequential and Threshold estimators are quite similar. Both seek to non-uniformly allocate inner stage samples based on minimization of the error margin. However, they are parameterized differently. The Sequential estimator takes as an input the parameter $\bar{m}$, which is the mean number of inner stage samples. On the other hand, the Threshold estimator takes as input the parameter $\gamma$, which is the threshold for the error margin. As argued earlier, for large values of $\bar{m}$ and $\gamma$, these two algorithms yield similar results. Further, we will see numerical evidence for this in Section 2.5.
CHAPTER 2. THE SEQUENTIAL METHOD

Algorithm 3: Estimate the probability of a large loss using a threshold-based non-uniform nested simulation. The parameter $\gamma$ is the error margin threshold. The parameter $n$ is the number of scenarios.
From a practical perspective, the **Sequential** estimator is more natural. In particular, if all other parameters are fixed, it is easy to choose a value for $\bar{m}$. This parameter explicitly specifies the total number of inner stage samples to be generated by $\bar{m}n$, and therefore determines the running time of the algorithm. Thus, we can choose $\bar{m}$ based on the available running time. In the **Threshold** estimator, the parameter $\gamma$ implicitly specifies the total number of inner stage samples to be generated, and hence indirectly determines the running time. It is not clear, however, how to make choice of $\gamma$ a priori that ensure a certain running time, for example.

From a theoretical perspective, however, the **Threshold** estimator proves much more amenable to analysis. The main reason is that, at any point during the execution of the algorithm, the loss estimates $\hat{L}(\omega^{(1)},\zeta^{(1)}), \ldots, \hat{L}(\omega^{(n)},\zeta^{(n)})$ are independent and identically distributed random variables. This i.i.d. structure will prove crucial in the analysis of Section 2.3.2 as it allows the analysis of the overall algorithm via the analysis of a single outer stage scenario.

Moreover, the **Threshold** estimator has another interesting interpretation. Given a threshold $\gamma$, consider a scenario $\omega^{(i)}$ with inner loss samples $\hat{Z}_{i,1}, \hat{Z}_{i,2}, \ldots$. Examining (2.11), the algorithm will generate $m_i$ inner stage samples in this scenario, with

$$m_i = \inf \left\{ m > 0 : \left| S_m^{(i)} \right| \geq \gamma \right\},$$

(2.12)

where, for $m \geq 0$, the partial sum is defined by

$$S_m^{(i)} = \sum_{j=1}^{m} \frac{1}{\sigma_i} \left( \hat{Z}_{i,j} - c \right).$$

(2.13)

Note that $\{S_m^{(i)}, m \geq 0\}$ is a random walk with unit variance increments. Then, the number of samples $m_i$ is determined by the first exit time of the random walk from the interval $(-\gamma, \gamma)$. This is illustrated in Figure 2.3. If the exit occurs through the upper barrier at $\gamma$, then $\hat{L}(\omega^{(i)},\zeta^{(i)}) > c$ and the scenario is declared to be a loss exceeding $c$. If the exist occurs through the lower barrier at $-\gamma$, then $\hat{L}(\omega^{(i)},\zeta^{(i)}) < c$ and the scenario is declared not to be a loss exceeding $c$.

The interpretation of the threshold policy in terms of the first exit of a random walk is reminiscent of sequential hypothesis testing (see, e.g., [Siegmund (1985)]). Indeed, for each scenario $\omega^{(i)}$, the threshold estimator is defining a sequential compound hypothesis test of whether the i.i.d. unit variance random variables $\{(Z_{i,j} - c)/\sigma_i\}$ have a positive or negative mean. As we show next, techniques from sequential analysis will prove helpful in theoretical analysis of our algorithm.
2.3.2 Asymptotic Analysis

Define \( \hat{\alpha}_{\text{SEQ}(m,n)} \) to be the threshold estimate, i.e.,

\[
\hat{\alpha}_{\text{SEQ}(m,n)} \triangleq \text{Threshold}(\gamma,n).
\]

As in Section 2.1, we will analyze the accuracy of this estimator by decomposing the mean squared error into bias and variance terms. We begin with an assumption:

**Assumption 2.2.** Assume that:

1. Conditional on an outer stage scenario \( \omega^{(i)} \in \Omega \), the inner stage samples \( \hat{Z}_{i,1}, \hat{Z}_{i,2}, \ldots \) are i.i.d. normal random variables. Denote the standard deviation of these samples by \( \sigma(\omega^{(i)}) \).
2. Given a scenario \( \omega \in \Omega \), define the normalized excess loss

\[
\mu(\omega) \triangleq \frac{L(\omega) - c}{\sigma(\omega)}.
\]

Then, the probability density function \( p \) of \( \mu \),

\[
p(u) \triangleq \frac{d}{du} P(\mu \leq u),
\]

exists and is continuously differentiable in a neighborhood of 0.
The second condition of Assumption 2.2 is a technical condition that is reminiscent of the first condition of Assumption 2.1. The first condition of Assumption 2.2 is motivated by the random walk interpretation of Section 2.3.1. In particular, consider the random walk formed by the partial sums \( \{ S_m^{(i)}, m \geq 0 \} \) from (2.13). By the functional central limit theorem, under a proper scaling, this process converges to a Brownian motion, i.e., a random walk with normal increments. The first condition makes the assumption that the unscaled random walk also has normal increments.

We are interested in the accuracy of the Threshold estimator in the asymptotic regime where the resulting estimate converges to the true value, i.e., as \( n \to \infty \) (many outer stage scenarios) and \( \gamma \to \infty \) (many inner stage samples). Our first result is the following theorem, which characterizes the asymptotic bias of this estimator:

**Theorem 2.2.** Under Assumption 2.2 as \( \gamma \to \infty \), the asymptotic bias of the Threshold estimator satisfies

\[
E \left[ \hat{\alpha}_{\text{SEQ}}(m,n) - \alpha \right] = O\left(\gamma^{-2}\right).
\]

The proof of Theorem 2.2 is provided in Appendix A. It relies on the random walk interpretation of Section 2.3.1 as well as techniques from sequential analysis. Specifically, exponential martingales are used in combination with the optional stopping theorem.

The following is an immediate corollary of Theorem 2.2 and provides an asymptotic expression for the variance of the simplified sequential estimator:

**Corollary 2.2.** Under the conditions of Theorem 2.2 as \( \gamma \to \infty \), the variance of the Threshold estimator satisfies

\[
\text{Var}(\hat{\alpha}_{\text{SEQ}}(m,n)) = \frac{\alpha(1-\alpha)}{n} + O\left(\gamma^{-2}n^{-1}\right).
\]

**Proof.** Note that

\[
\text{Var}(\hat{\alpha}_{\text{SEQ}}(m,n)) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{L}(\omega^{(i)},\zeta^{(i)}) \geq c\}\right)
\]

\[
= \frac{1}{n} \text{Var}\left(\mathbb{1}\{\hat{L}(\omega^{(1)},\zeta^{(1)}) \geq c\}\right)
\]

\[
= \frac{E[\hat{\alpha}_{\text{SEQ}}(m,n)]}{n} \left(1 - E[\hat{\alpha}_{\text{SEQ}}(m,n)]\right)
\]

where we have used the fact that the loss estimates \( \{\hat{L}(\omega^{(i)},\zeta^{(i)})\} \) are independent and identically
CHAPTER 2. THE SEQUENTIAL METHOD

31

distributed. Applying Theorem 2.2,

\[
\text{Var} \left( \hat{\alpha}_{\text{SEQ}}(m,n) \right) = \frac{\alpha(1-\alpha)}{n} + \frac{\mathbb{E} \left[ \hat{\alpha}_{\text{SEQ}}(m,n) - \alpha \right] (1 - \mathbb{E} \left[ \hat{\alpha}_{\text{SEQ}}(m,n) \right])}{n} + \frac{\alpha \mathbb{E} [\alpha - \hat{\alpha}_{\text{SEQ}}(m,n)]}{n}
\]

\[= \frac{\alpha(1-\alpha)}{n} + O \left( \gamma^{-2}n^{-1} \right). \]

The total run-time of the Threshold estimator is proportional to the total number of inner stage samples generated. Note, however, by the nature of the algorithm, the number of inner samples is stochastic. Hence, define \( \bar{m}(\gamma) \) to be the expected number of inner stage samples at a single outer stage scenario, given parameter \( \gamma \). That is,

\[\bar{m}(\gamma) \triangleq \mathbb{E} \left\{ \inf \left\{ m > 0 : \frac{m}{\sigma(\omega)} \left| \hat{L}(\omega, \zeta) - c \right| \geq \gamma \right\} \right\} \tag{2.14}\]

Here, the expectation is over the scenario \( \omega \) and the corresponding loss estimate \( \hat{L}(\omega, \zeta) \). Then, given parameters \((\gamma, n)\), the Threshold estimator has expected run-time proportional to \( \bar{m}(\gamma)n \). The following theorems, whose proof is given in Appendix A, characterizes the rate of growth of this run-time, as a function of \( \gamma \):

**Theorem 2.3.** Under Assumption 2.2 as \( \gamma \to \infty \), the expected number of inner stages samples in each scenario under the Threshold estimator satisfies

\[\bar{m}(\gamma) = O \left( \gamma \ln \gamma \right).\]

Note that Theorem 2.3 is intuitive given the first exit time interpretation of Figure 2.3. In particular, for large values of \( \gamma \), the amount of time required for a random walk starting at the origin with drift \( \mu \neq 0 \) to exit the interval \((-\gamma, \gamma)\) is approximately \( \gamma/|\mu| \). If the random walk has zero drift, the exit time is approximately \( \gamma^2 \). In our case, the expected number of samples \( \bar{m}(\gamma) \) is averaged over various possibilities of drift given by \( \mu(\omega) \triangleq (L(\omega) - c)/\sigma(\omega) \). The probability of this drift being exactly zero is zero, by the second condition of Assumption 2.2. However, arbitrarily small drifts are possible and thus \( \bar{m}(\gamma) \) is slightly larger than \( O(\gamma) \).

While Theorem 2.3 provides an \( O(\gamma \log \gamma) \) bound on the expected number of inner stage samples per scenario, it might be the case that the realized number of inner stage samples per scenario is larger. The following theorem guarantees that, so long as the number of scenarios \( n \) is sufficiently
large, an $O(\gamma \log \gamma)$ bound continues to hold on the number of realized samples per scenario with high probability. The proof can be found in Appendix A.

**Theorem 2.4.** Under Assumption 2.2, suppose that $C_0, \gamma_0 > 0$ are constants so that, for all $\gamma \geq \gamma_0$,

$$
\bar{m}(\gamma) \leq C_0 \gamma \log \gamma.
$$

(Such constants are guaranteed to exist by Theorem 2.3.) Further, suppose the number of scenarios $n \triangleq n(\gamma)$ is chosen as a function of $\gamma$, and that there exist constants $C_1, \gamma_1 > 0$, so that for all $\gamma \geq \gamma_1$,

$$
n(\gamma) \geq C_1 \gamma.
$$

That is, $n$ asymptotically grows at least linearly in $\gamma$. Then, for any $\epsilon, \delta > 0$, there exists $\gamma_2 > 0$ so that for all $\gamma \geq \gamma_2$,

$$
P \left( \frac{1}{n} \sum_{i=1}^{n} m_i \geq (C_0 + \epsilon) \gamma \log \gamma \right) < \delta.
$$

### 2.3.3 Optimal Non-Uniform Threshold Estimator

Theorems 2.2 and 2.3 and Corollary 2.2 allow a comparison between the **UNIFORM** estimator and the non-uniform **THRESHOLD** estimator. In particular, suppose $\hat{\alpha}_{SN}(m,n)$ is the **UNIFORM** estimate with $n$ scenarios and $m$ inner stage samples. As discussed in Section 2.1, when $m, n \to \infty$, this has asymptotic bias and variance

$$
E \left[ \hat{\alpha}_{SN}(m,n) - \alpha \right] = \frac{\theta_c}{m} + O \left( m^{-3/2} \right), \quad \text{Var} \left( \hat{\alpha}_{SN}(m,n) \right) = \frac{\alpha (1 - \alpha)}{n} + O \left( m^{-1} n^{-1} \right). \tag{2.15}
$$

On the other hand, suppose that $\hat{\alpha}_{SEQ}(m,n)$ is the non-uniform **THRESHOLD** estimator with $n$ scenarios and a threshold of $\gamma$. By Theorem 2.3, this estimator will employ, on average, $\bar{m} \triangleq \bar{m}(\gamma) = O(\gamma^{1+\epsilon})$ inner stage samples per scenario, for any positive $\epsilon$. We can express the asymptotic bias and variance results of Theorem 2.2 and Corollary 2.2 as a function of $n$ and $\bar{m}$ by

$$
E \left[ \hat{\alpha}_{SEQ}(m,n) - \alpha \right] = O \left( \bar{m}^{-2+\epsilon} \right), \quad \text{Var} \left( \hat{\alpha}_{SEQ}(m,n) \right) = \frac{\alpha (1 - \alpha)}{n} + O \left( \bar{m}^{-2+\epsilon} n^{-1} \right), \tag{2.16}
$$

for all positive $\epsilon$.

Comparing (2.15) and (2.16), we see that, up to the dominant term, the two algorithms achieve the same asymptotic variance of order $n^{-1}$. This is consistent with the discussion in Section 2.1.
which suggests that the asymptotic variance is determined by the randomness in scenario generation. This is exactly the same in the two algorithms. The inner stage sampling is different, however, and this results in a difference in bias for the estimators. Specifically, as a function of the average number of inner stage samples per scenario, the bias of the non-uniform \textit{Threshold} estimator decays approximately as the square of the bias of the \textit{Uniform} estimator.

Given a total work budget of $k$ (i.e., $mn \leq k$), we saw in Section 2.1 that the optimal \textit{Uniform} estimator (in the sense of minimizing MSE) would utilize a number of scenarios $n$ of order $k^{1/3}$, a number of inner stage samples per scenario $m$ of order $k^{2/3}$, and result in an MSE of order $k^{-2/3}$. For the non-uniform \textit{Threshold} estimator, from the results of Section 2.3.2, we can bound the MSE by
\[
E \left[ (\hat{\alpha}_{\text{SEQ}(m,n)} - \alpha)^2 \right] \leq \frac{\alpha(1-\alpha)}{n} + \frac{C}{\gamma^4},
\]
for sufficiently large $n$ and $\gamma$ and an appropriate choice of the constant $C$. We can find a non-uniform \textit{Threshold} estimator with low MSE by minimizing this upper bound over choices of $(\gamma, n)$, subject to an expected total work constraint. That is, we consider optimization problem
\[
\begin{align*}
\text{minimize} & \quad \frac{\alpha(1-\alpha)}{n} + \frac{C}{\gamma^4} \\
\text{subject to} & \quad \bar{m}(\gamma)n \leq k, \\
& \quad \gamma, n \geq 0.
\end{align*}
\]

For any positive $\epsilon$ and given a work budget $k$, suppose we choose $\gamma^* \propto k^{1/5}$ and $n^* \propto k^{4/5-\epsilon}$. Then, we have that $\bar{m}(\gamma^*)n^* = O(k^{1-\epsilon} \log k) = o(k)$. Thus, for sufficiently large $k$, the expected total work will be less that $k$. Indeed, since $(\gamma^*, n^*)$ satisfy the conditions of Theorem 2.4 for sufficiently large $k$ the realized total work will also be less than $k$ with high probability. This choice will result in an MSE of $O(k^{-4/5+\epsilon})$. Hence, the optimal non-uniform \textit{Threshold} estimator converges at a faster rate than any uniform estimator. This is accomplished by generating more outer scenarios ($k^{4/5-\epsilon}$ vs. $k^{2/3}$) and performing less inner stage sampling on average in each scenario ($k^{1/5}$ vs. $k^{1/3}$) than is optimal in the uniform case.

### 2.4 Adaptive Allocation Algorithm

The non-uniform \textit{Sequential} estimator provides a way to determine the placement of inner stage samples across scenarios. The decision of how to allocate computational effort between generating
more scenarios (i.e., the choice of \( n \)) and generating more inner samples across scenarios (i.e.,
the choice of \( \bar{m} \)) is unaddressed, however. The discussion in Section 2.3.3 suggests that, given
a total work budget of \( k \), one should asymptotically approximately choose \( n \propto k^{4/5} \) and \( \bar{m} \propto k^{1/5} \). However, the constants in these asymptotic expressions are unspecified. The choice of these
constants may have an enormous impact on the practical performance of these algorithms. Note
that the **Uniform** estimator faces the same problem — as described in Remark 2.1, the optimal
allocation \((2.4)\) suggested by the analysis of Section 2.1 requires knowledge of the constant \( \theta_c \). It
is not clear, in general, how to determine this constant.

In this section we will consider an adaptive allocation approach. This algorithm is a heuristic
that estimates the optimal choice of \( \bar{m} \) and \( n \) at each point in time. It refines these estimates over
the course of the simulation. The idea of this approach is that, based on the results of Section 2.3,
the variance is determined by the number of scenarios (\( n \)) and the squared bias is determined by
the amount of inner sampling (\( \bar{m} \)). The adaptive algorithm estimates these quantities and then
either increases the number of scenarios or increases the number of inner samples depending on
whether the MSE is dominated by the variance or the squared bias.

Specifically, the **Adaptive** estimator of Algorithm 4 proceeds as follows:

1. The simulation is initialized (lines 2–7) by generating \( n^0 \) scenarios with \( m^0 \) inner samples for
each scenario.

2. The work budget of the simulation \( k \) is divided into \( K \triangleq k/\tau_e \) intervals (or, epochs) of length
\( \tau_e \) (note that we assume for simplicity of exposition that \( K \) is an integer, and that the first
epoch is only of length \( \tau_e - n^0m^0 \) due to the initialization).

3. At the beginning of the \( \ell \)th epoch (line 9), estimates are made for the squared bias and
variance of the loss probability estimate, given the scenarios and samples that have been
generated thus far. Specifically, given the loss probability estimate

\[
\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{L(\omega^{(i)}, \zeta^{(i)}) \geq c\},
\]

the bias is approximated according to

\[
E[\hat{\alpha} - \alpha] \approx \hat{B} \triangleq \hat{\alpha} - \bar{\alpha}, \quad (2.18)
\]
where
\[ \bar{\alpha} \triangleq \frac{1}{n} \sum_{i=1}^{n} N \left( \frac{m_i (\hat{L}(\omega^{(i)}, \zeta^{(i)}) - c)}{\sigma_i} \right), \]
and \( N \) is the cumulative distribution function for the normal distribution. This approximation is based on a central limit theorem heuristic: in each scenario \( \omega^{(i)} \), when the number of samples \( m_i \) is large, each loss estimate \( \hat{L}(\omega^{(i)}, \zeta^{(i)}) \) can be approximated by a normal distribution with mean equal to \( L(\omega^{(i)}) \) and with variance \( \sigma_i^2 / m_i \). Hence, given a fixed set of scenarios \( \omega^{(1)}, \ldots, \omega^{(n)} \), one might estimate the bias via
\[
E[\hat{\alpha} - \alpha] = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{P} \left( \hat{L}(\omega^{(i)}, \zeta^{(i)}) \geq c \right) - \mathbb{I}\{L(\omega^{(i)}) \geq c\} \right\}
\approx \frac{1}{n} \sum_{i=1}^{n} \left\{ N \left( \frac{m_i (L(\omega^{(i)}) - c)}{\sigma_i} \right) - \mathbb{I}\{L(\omega^{(i)}) \geq c\} \right\}.
\]
Since each true loss \( L(\omega^{(i)}) \) is unknown in practice, we can approximate this with its realized estimate \( \hat{L}(\omega^{(i)}, \zeta^{(i)}) \). This results in (2.18). By making a similar heuristic approximation for the variance, we arrive at the expression
\[
\text{Var} (\hat{\alpha}) \approx \hat{V} \triangleq \frac{\bar{\alpha} (1 - \bar{\alpha})}{n}. \tag{2.19}
\]
Note that the estimators (2.18) and (2.19) are meant only as heuristics. Better estimators may be possible and bias, in particular, is notoriously difficult to estimate. For our purposes, however, they only need to be accurate within orders of magnitude so as to allocate computational effort between inner samples and outer scenarios. We will see in the numerical results of Section 2.5 that, empirically, they suffice for this purpose.

4. Suppose there are \( n \) outer scenarios and an average of \( \bar{m} \triangleq \frac{1}{n} \sum_{i=1}^{n} m_i \) inner samples per scenario at the beginning of the \( \ell \)th epoch. From the results in Section 2.3, we expect the squared bias to decrease according to \( \bar{m}^{-4+\epsilon} \) and the variance to decrease in proportion to \( n^{-1} \). Then, assume that the number of scenarios and samples at the end of the \( \ell \)th epoch are given by \( n' \) and \( m' \). We can estimate the squared bias at the end of the \( \ell \)th epoch, as a function of the bias estimate \( \hat{B} \) at the beginning, by
\[
\hat{B}^2 \left( \frac{\bar{m}}{m'} \right)^4.
\]
Similarly, the variance at the end of the \( \ell \)th epoch can be estimated by 

\[ \hat{V} \left( \frac{n}{n'} \right). \]

Thus, at the beginning of the \( \ell \)th epoch, we consider the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \tilde{B}^2 \left( \frac{\bar{m}}{\bar{m}'} \right)^4 + \hat{V} \left( \frac{n}{n'} \right) \\
\text{subject to} & \quad \bar{m}'n' = \bar{mn} + \tau_e, \\
& \quad n \leq n' \leq n + \tau_e, \\
& \quad \bar{m}' \geq 0.
\end{align*}
\]

(2.20)

This problem seeks to make a choice of \((\bar{m}', n')\) that results in a minimal mean squared error at the end of the \( \ell \)th epoch. The first constraint ensures that the total number of inner samples in the \( \ell \)th epoch will equal the epoch length \( \tau_e \). The second constraint ensures that the number of scenarios at the end of the \( \ell \)th epoch is at least the number of scenarios at the beginning, and increases by at most the length of the epoch.

The solution to (2.20) is given by

\[
n' = \min \left\{ \max \left\{ \left( \frac{\hat{V}n}{4\tilde{B}^2\bar{m}^4} (\bar{mn} + \tau_e)^4 \right)^{1/5}, n \right\} , n + \tau_e \right\}, \quad \bar{m}' = \frac{\bar{mn} + \tau_e}{n'}.
\]

(2.21)

After obtaining the target number of scenarios \( n' \) (line 10), \( n' - n \) additional scenarios are generated.

5. Over the course of the \( \ell \)th epoch (lines 13–21) \( \tau_e \) inner samples are generated. These are distributed to ensure that every scenario has at least \( m^0 \) inner samples in total (not per epoch). Once that is the case, inner samples are allocated myopically according to minimum error margin as in the sequential estimator.

### 2.5 Numerical Results

In this section we present numerical results that illustrate the benefits of non-uniform nested estimation. We begin in Section 2.5.1 by describing two settings for our numerical experiments. In Section 2.5.2, we compare the bias of the uniform estimator and the non-uniform threshold estimator.
Algorithm 4: Estimate the probability of a large loss using an adaptive non-uniform nested simulation. This estimator employs a sequential algorithm to determine the placement of inner stage samples across scenarios, and adaptively decides the number of scenarios and inner samples to add by estimating the bias and variance. The parameters $n^0$ and $m^0$ are the initial number of scenarios and inner samples per scenario, respectively. The parameter $\tau_e$ is the epoch length. The parameter $k$ is the total number of inner samples. Note that each standard deviation $\sigma_i$ can be estimated in an online fashion over the course of the simulation, as is discussed in Section 2.5.4.
and sequential estimators. In Section 2.5.3, we compare the MSE of a number of both implementable and idealized uniform and non-uniform estimators. Finally, in Section 2.5.4, we consider issues arising from the estimation of the variance of inner stage samples.

2.5.1 Experimental Setting

Our numerical experiments are set in the context of the following two examples: a portfolio with Gaussian cashflows, where both the outer stage scenarios and inner stage samples are generated from normal distributions, and a put option example, where the portfolio consists of a single put option on an underlying asset whose price follows a geometric Brownian motion process. For both examples, we are interested in computing the probability of a loss. We consider loss thresholds corresponding to 10%, 1%, and 0.1% loss probabilities.

In the Gaussian example, we consider a portfolio with normally distributed risk factors and cashflows. This is the simplest setting in which to test any nested simulation procedure. Specifically, we consider a portfolio with value $X_0 = 0$ at time $t = 0$ and value $X_\tau(\omega) = \omega$ at the risk horizon $\tau$. We assume that the real-valued risk factor $\omega \in \mathbb{R}$ is normally distributed with mean zero and standard deviation $\sigma^2_{\text{outer}} = 1$. Then, the loss $L(\omega) = X_0 - X_\tau(\omega) = -\omega$ is a standard normal random variable. Given a scenario $\omega^{(i)}$, each inner loss sample takes the form $\hat{Z}_{i,j} = -\omega^{(i)} + \sigma_{\text{inner}}W^{(i,j)}$, where $W^{(i,j)}$ is a standard normal random variable and $\sigma_{\text{inner}} = 5$ is the standard deviation of the inner stage samples.

In this case, given a loss threshold $c$, the probability of a loss exceeding $c$ is given by $\alpha = N(-c)$. We choose the values 1.282, 2.326, and 3.090 for the loss threshold $c$, corresponding to loss probabilities $\alpha$ of 10%, 1%, and 0.1%, respectively.

In the put option example, we assume that the portfolio consists of a long position in a single put option. This example is more complex since the portfolio cashflows are non-linear and follow highly skewed distributions, which vary substantially across outer stage scenarios. Here, the underlying asset follows a geometric Brownian motion with an initial price of $S_0 = 100$. The drift of this process under the real-world distribution used in the outer stage of simulation is $\mu = 8\%$. The annualized volatility is $\sigma = 20\%$. The riskless rate is $r_f = 3\%$. The strike of the put option is $K = 95$ and the maturity is $T = 0.25$ years (i.e., three months). The risk horizon is $\tau = 1/52$ years (i.e., one week). With these parameters, the initial value of the put is $X_0 = 1.669$ given by the
Black-Scholes formula.

Denote by $S_\tau(\omega)$ the underlying asset price at the risk horizon $\tau$. This random variable is generated according to

$$S_\tau(\omega) \triangleq S_0 e^{(\mu - \sigma^2/2)\tau + \sigma \sqrt{\tau} \omega},$$

where the real-valued risk-factor $\omega$ is a standard normal random variable. The portfolio loss at the risk horizon $\tau$ is given by

$$L(\omega) = X_0 - E\left[e^{-r_f(T-\tau)} \max (K - S_T(\omega, W), 0) \mid \omega \right],$$

where the expectation is taken over the random variable $W$, which is an independently distributed standard normal, and $S_T(\omega, W)$ is given by

$$S_T(\omega, W) \triangleq S_\tau(\omega)e^{(r_f - \sigma^2/2)(T-\tau) + \sigma \sqrt{T-\tau} W}.$$

Note that, given a fixed value of $\omega$ and a standard normal $W$, the random variable $S_T(\omega, W)$ is distributed according to the risk-neutral distribution of underlying asset price at the option maturity $T$, conditional on asset price $S_\tau(\omega)$ at the risk horizon $\tau$. Given an outer scenario $\omega^{(i)}$, each inner loss sample takes the form

$$\hat{Z}_{i,j} = X_0 - e^{-r_f(T-\tau)} \max \left(K - S_T(\omega^{(i)}, W^{(i,j)}), 0\right),$$

where $W^{(i,j)}$ is an independent standard normal random variable. Notice that outer stage scenarios are generated using the real-world distribution governed by the drift $\mu$, while inner stage scenarios used to generate future put option prices are generated using the risk-neutral distribution governed by the drift $r_f$.

It is not difficult to see that the loss $L(\omega)$ is strictly increasing in the risk factor $\omega$. Hence, the probability of a loss exceeding a threshold $c$ can be computed according to $\alpha = P(L \geq c) = P(\omega \geq \omega^*) = N(-\omega^*)$, where $\omega^*$ is the unique solution to $L(\omega^*) = c$. We choose the values 0.859, 1.221, and 1.390 for the loss threshold $c$, corresponding to loss probabilities $\alpha$ of 10%, 1%, and 0.1%, respectively.

### 2.5.2 Bias Comparison

As established in Section 2.3, the advantage of non-uniform inner stage sampling relative to uniform sampling is that, for the same total quantity of inner samples, a lower bias is attained. In this
section, we numerically compare the \texttt{Uniform} estimator and the non-uniform \texttt{Threshold} and \texttt{Sequential} estimators on the basis of bias.

For this purpose, we generate a fixed sequence $\omega^{(1)}, \ldots, \omega^{(n)}$ of $n = 10,000$ outer stage scenarios. In order to eliminate any noise in our comparison due to randomness in scenario generation, we choose the scenarios in a deterministic and stratified manner, so that

$$P(\omega \leq \omega^{(i)}) = i/(n + 1), \quad \text{for all } 1 \leq i \leq n.$$ 

Given the stratified scenarios, we numerically compute the bias of each estimator, measured over 1,000 independent trials, as the total number of inner stage samples (i.e., the work budget) is varied from $k = 20,000$ to $k = 4,000,000$. In the case of the \texttt{Uniform} estimator, this is accomplished by varying the number of inner stage samples per scenario from $m = 2$ to $m = 400$. For the non-uniform \texttt{Sequential} estimator, this is accomplished by using $m^0 = 2$ initial inner samples per scenario, and then varying the average number of inner stage samples per scenario from $\bar{m} = 2$ to $\bar{m} = 400$. In the case of the non-uniform \texttt{Threshold} estimator, the threshold parameter $\gamma$ was varied over the interval $(5 \times 10^{-5}, 2 \times 10^{-1})$ and the expected total number of inner stage samples was plotted (averaged over the independent trials). This range of $\gamma$ was experimentally chosen so that the range of expected total inner samples for the \texttt{Threshold} algorithm coincided with the range of total inner samples for the other algorithms.

The results for both the Gaussian example and the put option example with $\alpha = 1\%$ are plotted in Figure 2.4. In both cases, the non-uniform \texttt{Threshold} and \texttt{Sequential} estimators exhibit a lower bias than the \texttt{Uniform} estimator, given the same work budget. Further, for the \texttt{Uniform} estimator, the results are consistent with the bias decreasing with order $k^{-1}$, as suggested by Theorem 2.1. For the non-uniform \texttt{Threshold} estimator, the results are consistent with the bias decreasing according to $k^{-2+\epsilon}$ for any positive $\epsilon$, as suggested by the theory presented in Section 2.3. Note that the performance of the \texttt{Threshold} and \texttt{Sequential} estimators is largely indistinguishable. This strongly suggests that our theoretical analysis of the rate of convergence

---

5The exact rates of decay (i.e., the asymptotic slopes in Figure 2.4) are challenging to accurately estimate numerically. This is because it is not computationally feasible to compute the estimators over many orders of magnitude of $k$. The results in Figure 2.4 are not intended as numerical “proof” of a particular rate of convergence, but are rather intended to illustrate that the numerical convergence is consistent with our earlier theoretical analyses.
of Threshold estimator in Section 2.3 provides a good proxy for the rate of convergence of the Sequential estimator.

Figure 2.5 gives some qualitative insight into the inner sampling behavior of the non-uniform Sequential estimator. Here, we have plotted the number of inner samples (averaged across the 1,000 independent trials) at a scenario against the loss in the scenario. The amount of inner sampling employed by the Sequential varies over two orders of magnitude across scenarios, with much more sampling taking place close to the loss threshold \(c\) than far away from it.

### 2.5.3 MSE Comparison

In this section, we will provide an overall comparison of the MSE achieved by various uniform and non-uniform estimators, given a fixed computational budget of \(k\) inner stage samples. We consider each of the following estimators:

- **Optimal uniform.** This is the Uniform estimator with parameters chosen optimally, as in Section 2.3. The simulation budget is allocated according to \(m = k^{1/3}/\beta^*\) and \(n = \beta^*k^{2/3}\), where the constant \(\beta^*\), given in (2.4), is chosen to minimize MSE. Note that, in general, it is not clear how to determine the value \(\beta^*\) given the problem parameters. For both the Gaussian and put option examples here, we are able to use closed-form expressions for the probability distribution of losses to exactly compute this constant.

- **\(1/3:2/3\) uniform.** This is the Uniform estimator with \(m = k^{1/3}\) and \(n = k^{2/3}\). Based on the analysis in Section 2.1 this estimator has MSE that decays with same order \((k^{-2/3})\) as the optimal uniform estimator, but with a suboptimal constant. This is meant to illustrate the case where the constant \(\beta^*\) of the optimal uniform estimator is unknown, and an arbitrary choice of constant \((\beta^* = 1)\) is made.

- **Optimal sequential.** This is the Sequential non-uniform sampling estimator where \(n\) is chosen optimally to minimize MSE. Here, \(m^0 = 2\) initial samples were used. The parameters \((\bar{m}, n)\) controlling the average number of inner samples and the number of scenarios were varied over choices satisfying the simulation work budget, i.e., \(\bar{m}n = k\). The choice that resulted in the minimum MSE was selected. The optimal sequential estimator is an idealized
Figure 2.4: The vertical axis shows the bias in absolute terms, i.e., the absolute value of difference between the estimated loss probability and the true loss probability $\alpha$, as a function of the total number of inner stage samples. In the case of the Threshold algorithm, the expected total number of samples is shown. A set of $n = 10,000$ stratified outer scenarios was used. The bias of the non-uniform Threshold and Sequential estimators is consistent with the predicted theoretical decay of $k^{-2+\epsilon}$, for any positive $\epsilon$. Similarly, the bias of the Uniform estimator is consistent with the predicted theoretical decay of $k^{-1}$. 
Figure 2.5: The number of inner stage samples as a function of the loss in each scenario, averaged over 1,000 trials. Here, \( k = 4,000,000 \) inner stage samples are distributed across \( n = 10,000 \) stratified scenarios. The \text{UNIFORM} estimator employs \( m = 400 \) inner samples for each scenario. The non-uniform \text{SEQUENTIAL} estimator varies the number of samples over two orders of magnitude, and employs many more samples close to the loss threshold \( c \).
algorithm meant to capture the best possible performance than can be achieved using the **sequential** estimator.

**Adaptive.** This is the **Adaptive** estimator of Section 2.4 which utilizes sequential non-uniform sampling and adaptively allocates computational effort between outer stage scenarios and inner stage samples. Here, \( n^0 = 500 \) initial scenarios were used, with \( m^0 = 2 \) initial inner samples per scenario. An epoch length of \( \tau_e = 100,000 \) was used.

**Adaptive (\( \hat{\sigma}_i \)).** This is a variation of the **Adaptive** estimator in which the variance of inner samples is estimated. This will be discussed shortly in Section 2.5.4.

The numerical results for the two examples (the Gaussian and put option examples, each with thresholds corresponding to three different loss probabilities) using the five estimators are summarized in Tables 2.1 and 2.2. In all cases, a computational budget of \( k = 4,000,000 \) inner stage samples was used. The results in each case are computed over 1,000 independent trials.

The numerical results in Tables 2.1 and 2.2 can be interpreted naturally through a series of pairwise comparisons, as follows:

- **Optimal uniform vs. 1/3:2/3 uniform.** These are both asymptotically optimal **Uniform** estimators; they differ only by the choice of constant \( \beta^* \). The practical performance of these two estimators, however, is dramatically different. This highlights the sensitivity of the **Uniform** estimator in practice to the choice of constant. Note that computing the constant \( \beta^* \), as given in (2.4), requires knowledge of the constant \( \theta_c \), defined by (2.3). It is not clear how to estimate this constant in practice, and this constant may vary dramatically across different problem instances.

- **Optimal uniform vs. optimal sequential.** These represent the best possible performance that can be achieved by the **Uniform** and **sequential** estimators. Neither of these estimators is implementable in practice — the former because it depends on a parameter that cannot be readily determined from the problem data, the latter because it requires exploration over the choice of parameters. However, by contrasting them we can see a comparison of uniform and non-uniform sampling on an equal footing. This comparison clearly illustrates benefits of non-uniform sampling. In every test case, the optimal sequential estimator has the lowest
with thresholds corresponding to three different loss probabilities. The results are computed over 1

Table 2.1: Numerical results for five estimation algorithms over three test cases of the Gaussian example

<table>
<thead>
<tr>
<th>Gaussian</th>
<th>$n$</th>
<th>$\bar{m}$</th>
<th>variance</th>
<th>bias$^2$</th>
<th>MSE</th>
<th>std. err.</th>
<th>MSE norm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 10%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\frac{1}{3}:\frac{2}{3}$ uniform</td>
<td>25,199</td>
<td>159</td>
<td>4.0·10^{-6}</td>
<td>2.8·10^{-4}</td>
<td>2.9·10^{-4}</td>
<td>2.1·10^{-6}</td>
<td>35.4</td>
</tr>
<tr>
<td>optimal uniform</td>
<td>4,499</td>
<td>889</td>
<td>2.1·10^{-5}</td>
<td>8.6·10^{-6}</td>
<td>3.0·10^{-5}</td>
<td>1.2·10^{-6}</td>
<td>3.7</td>
</tr>
<tr>
<td>adaptive ($\hat{\sigma}_i$)</td>
<td>14,968</td>
<td>281</td>
<td>7.0·10^{-6}</td>
<td>2.7·10^{-6}</td>
<td>9.7·10^{-6}</td>
<td>4.7·10^{-7}</td>
<td>1.2</td>
</tr>
<tr>
<td>adaptive</td>
<td>12,802</td>
<td>321</td>
<td>7.2·10^{-6}</td>
<td>1.5·10^{-6}</td>
<td>8.6·10^{-6}</td>
<td>3.9·10^{-7}</td>
<td>1.0</td>
</tr>
<tr>
<td>optimal sequential</td>
<td>12,395</td>
<td>323</td>
<td>6.8·10^{-6}</td>
<td>1.4·10^{-6}</td>
<td>8.2·10^{-6}</td>
<td>3.7·10^{-7}</td>
<td>1.0</td>
</tr>
</tbody>
</table>

| $\alpha = 1\%$              |      |           |          |          |      |           |           |
| $\frac{1}{3}:\frac{2}{3}$ uniform | 25,199 | 159   | 6.1·10^{-7} | 2.8·10^{-5} | 2.8·10^{-5} | 2.6·10^{-7} | 60.9     |
| optimal uniform              | 5,089 | 786     | 2.3·10^{-6} | 1.0·10^{-6} | 3.3·10^{-6} | 1.5·10^{-7} | 7.2      |
| adaptive ($\hat{\sigma}_i$) | 16,177 | 250   | 7.0·10^{-7} | 3.7·10^{-9} | 7.0·10^{-7} | 3.1·10^{-8} | 1.5      |
| adaptive                     | 16,118 | 251   | 7.1·10^{-7} | 4.1·10^{-9} | 7.2·10^{-7} | 3.1·10^{-8} | 1.6      |
| optimal sequential           | 30,860 | 130   | 3.5·10^{-7} | 1.1·10^{-7} | 4.6·10^{-7} | 1.8·10^{-8} | 1.0      |

| $\alpha = 0.1\%$            |      |           |          |          |      |           |           |
| $\frac{1}{3}:\frac{2}{3}$ uniform | 25,199 | 159   | 8.2·10^{-8} | 1.1·10^{-6} | 1.2·10^{-6} | 1.9·10^{-8} | 48.0     |
| optimal uniform              | 7,788 | 514     | 1.7·10^{-7} | 7.9·10^{-8} | 2.5·10^{-7} | 1.3·10^{-8} | 10.0     |
| adaptive ($\hat{\sigma}_i$) | 30,798 | 132   | 3.5·10^{-8} | 4.7·10^{-10} | 3.5·10^{-8} | 1.6·10^{-9} | 1.4      |
| adaptive                     | 30,628 | 132   | 3.8·10^{-8} | 5.0·10^{-10} | 3.8·10^{-8} | 3.2·10^{-9} | 1.5      |
| optimal sequential           | 56,686 | 71    | 1.8·10^{-8} | 6.5·10^{-9} | 2.5·10^{-8} | 1.1·10^{-9} | 1.0      |

Table 2.1: Numerical results for five estimation algorithms over three test cases of the Gaussian example with thresholds corresponding to three different loss probabilities. The results are computed over 1,000 independent trials, each with a total simulation budget of $k = 4,000,000$. The results reported include the number of outer stage scenarios ($n$) and the average number of inner stage samples per scenario ($\bar{m}$) employed by each estimator, as well as the variance, the squared bias, the mean squared error (MSE), and the standard error of the MSE for each estimator. The last column contains MSE results normalized relative to the optimal sequential estimator.
### Table 2.2: Numerical results for five estimation algorithms over three test cases of the put option example with thresholds corresponding to three different loss probabilities. The results are computed over 1,000 independent trials, each with a total simulation budget of $k = 4,000,000$. The results reported include the number of outer stage scenarios ($n$) and the average number of inner stage samples per scenario ($\bar{m}$) employed by each estimator, as well as the variance, the squared bias, the mean squared error (MSE), and the standard error of the MSE for each estimator. The last column contains MSE results normalized relative to the optimal sequential estimator.

<table>
<thead>
<tr>
<th></th>
<th>$n$</th>
<th>$\bar{m}$</th>
<th>variance</th>
<th>bias$^2$</th>
<th>MSE</th>
<th>std. err.</th>
<th>norm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>put option</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSE</td>
<td>std. err.</td>
<td></td>
</tr>
<tr>
<td>$\alpha = 10%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSE</td>
<td>std. err.</td>
<td></td>
</tr>
<tr>
<td>$1/3:2/3$ uniform</td>
<td>25,199</td>
<td>159</td>
<td>4.1$\cdot$10$^{-6}$</td>
<td>5.1$\cdot$10$^{-4}$</td>
<td>5.1$\cdot$10$^{-4}$</td>
<td>2.9$\cdot$10$^{-6}$</td>
<td>58.6</td>
</tr>
<tr>
<td>optimal uniform</td>
<td>5,095</td>
<td>785</td>
<td>1.9$\cdot$10$^{-5}$</td>
<td>2.4$\cdot$10$^{-5}$</td>
<td>4.2$\cdot$10$^{-5}$</td>
<td>1.6$\cdot$10$^{-6}$</td>
<td>4.8</td>
</tr>
<tr>
<td>adaptive ($\hat{\sigma}_i$)</td>
<td>6,671</td>
<td>601</td>
<td>1.5$\cdot$10$^{-5}$</td>
<td>4.8$\cdot$10$^{-6}$</td>
<td>2.0$\cdot$10$^{-5}$</td>
<td>9.2$\cdot$10$^{-7}$</td>
<td>2.3</td>
</tr>
<tr>
<td>adaptive</td>
<td>7,325</td>
<td>547</td>
<td>1.3$\cdot$10$^{-5}$</td>
<td>2.1$\cdot$10$^{-7}$</td>
<td>1.4$\cdot$10$^{-5}$</td>
<td>6.2$\cdot$10$^{-7}$</td>
<td>1.6</td>
</tr>
<tr>
<td>optimal sequential</td>
<td>12,395</td>
<td>323</td>
<td>7.3$\cdot$10$^{-6}$</td>
<td>1.5$\cdot$10$^{-6}$</td>
<td>8.7$\cdot$10$^{-6}$</td>
<td>3.8$\cdot$10$^{-7}$</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha = 1%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSE</td>
<td>std. err.</td>
<td></td>
</tr>
<tr>
<td>$1/3:2/3$ uniform</td>
<td>25,199</td>
<td>159</td>
<td>7.8$\cdot$10$^{-7}$</td>
<td>9.4$\cdot$10$^{-5}$</td>
<td>9.5$\cdot$10$^{-5}$</td>
<td>5.4$\cdot$10$^{-7}$</td>
<td>141.8</td>
</tr>
<tr>
<td>optimal uniform</td>
<td>3,143</td>
<td>1,273</td>
<td>3.8$\cdot$10$^{-6}$</td>
<td>1.2$\cdot$10$^{-6}$</td>
<td>5.0$\cdot$10$^{-6}$</td>
<td>2.1$\cdot$10$^{-7}$</td>
<td>7.5</td>
</tr>
<tr>
<td>adaptive ($\hat{\sigma}_i$)</td>
<td>10,085</td>
<td>401</td>
<td>1.2$\cdot$10$^{-6}$</td>
<td>2.0$\cdot$10$^{-7}$</td>
<td>1.4$\cdot$10$^{-6}$</td>
<td>6.2$\cdot$10$^{-8}$</td>
<td>2.1</td>
</tr>
<tr>
<td>adaptive</td>
<td>9,992</td>
<td>405</td>
<td>1.1$\cdot$10$^{-6}$</td>
<td>1.7$\cdot$10$^{-8}$</td>
<td>1.1$\cdot$10$^{-6}$</td>
<td>4.8$\cdot$10$^{-8}$</td>
<td>1.6</td>
</tr>
<tr>
<td>optimal sequential</td>
<td>19,558</td>
<td>205</td>
<td>5.4$\cdot$10$^{-7}$</td>
<td>1.5$\cdot$10$^{-7}$</td>
<td>6.9$\cdot$10$^{-7}$</td>
<td>3.0$\cdot$10$^{-8}$</td>
<td>1.0</td>
</tr>
<tr>
<td>$\alpha = 0.1%$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSE</td>
<td>std. err.</td>
<td></td>
</tr>
<tr>
<td>$1/3:2/3$ uniform</td>
<td>25,199</td>
<td>159</td>
<td>1.5$\cdot$10$^{-7}$</td>
<td>8.1$\cdot$10$^{-6}$</td>
<td>8.2$\cdot$10$^{-6}$</td>
<td>7.2$\cdot$10$^{-8}$</td>
<td>174.5</td>
</tr>
<tr>
<td>optimal uniform</td>
<td>2,570</td>
<td>1,556</td>
<td>4.4$\cdot$10$^{-7}$</td>
<td>3.9$\cdot$10$^{-8}$</td>
<td>4.8$\cdot$10$^{-7}$</td>
<td>2.7$\cdot$10$^{-8}$</td>
<td>10.2</td>
</tr>
<tr>
<td>adaptive ($\hat{\sigma}_i$)</td>
<td>14,884</td>
<td>274</td>
<td>1.1$\cdot$10$^{-7}$</td>
<td>1.8$\cdot$10$^{-8}$</td>
<td>1.3$\cdot$10$^{-7}$</td>
<td>9.0$\cdot$10$^{-9}$</td>
<td>2.8</td>
</tr>
<tr>
<td>adaptive</td>
<td>14,384</td>
<td>284</td>
<td>9.2$\cdot$10$^{-8}$</td>
<td>5.6$\cdot$10$^{-10}$</td>
<td>9.2$\cdot$10$^{-8}$</td>
<td>1.4$\cdot$10$^{-8}$</td>
<td>2.0</td>
</tr>
<tr>
<td>optimal sequential</td>
<td>26,508</td>
<td>151</td>
<td>3.9$\cdot$10$^{-8}$</td>
<td>8.0$\cdot$10$^{-9}$</td>
<td>4.7$\cdot$10$^{-8}$</td>
<td>2.3$\cdot$10$^{-9}$</td>
<td>1.0</td>
</tr>
</tbody>
</table>
MSE. The MSE improvement relative to the optimal uniform estimator is between a factor of 4 and 10. This improvement is greatest when estimating loss probabilities that are rare (e.g., the $\alpha = 0.1\%$ case).

Further, note that the optimal sequential estimator employs many fewer inner stage samples and many more outer stage scenarios. This is consistent with the theory developed in Section 2.3 and the experiments in Section 2.5.2. The optimal sequential estimator is able to achieve a low bias with fewer inner stage samples, hence it can employ more scenarios with the same computational budget.

- **Optimal sequential vs. adaptive sequential.** The optimal sequential estimator relies on a brute force optimization over the parameters choosing the number of inner samples and outer scenarios; this is not feasible in practice. On the other hand, the adaptive sequential estimator makes this choice dynamically over the course of the simulation and thus is implementable in practice. Comparing these two methods illustrates how much of the benefit of the optimal sequential method can be achieved in practice.

  Across our experiments, the adaptive sequential estimator achieves an MSE between one and two times that of the optimal sequential estimator. In some cases, the adaptive estimator overestimates the true bias and uses too many inner stage samples compared to the optimal allocation. This suggests that there is modest room for improvement in the \texttt{Adaptive} procedure for allocating computational effort between inner and outer stages.

### 2.5.4 Variance Estimation

The \texttt{Adaptive} algorithm requires the value of $\sigma_i$, the standard deviation of the inner stage loss samples $\hat{Z}_{i,1}, \hat{Z}_{i,2}, \ldots$ in scenario $\omega^{(i)}$. In practice, $\sigma_i$ will not be known. However, one can imagine many variations of the \texttt{Adaptive} algorithm where each $\sigma_i$ is estimated over the course of the estimation algorithm.

One such variation replaces each $\sigma_i$ in the \texttt{Adaptive} algorithm with the estimate

\[
\hat{\sigma}_i \triangleq \frac{m_i}{m_i + b} \bar{\sigma}_i + \frac{b}{m_i + b} \bar{\sigma}.
\] (2.22)
Here, we define
\[
\tilde{\sigma}_i \triangleq \left[ \frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left( \hat{Z}_{i,j} - \hat{L}(\omega^{(i)}, \zeta^{(i)}) \right)^2 \right]^{1/2}
\]
to be the sample standard deviation of the inner stage loss samples generated in scenario \(\omega^{(i)}\), and
\[
\bar{\sigma} \triangleq \frac{1}{n} \sum_{i=1}^{n} \tilde{\sigma}_i
\]
to be the overall average of all such sample standard deviations. This procedure balances an ensemble estimate with a local estimate so that the estimated standard deviations can be generated more reliably, especially when there are a small number of inner stage samples at a given scenario. For \(b = 0\), the procedure corresponds to the usual sample standard deviation estimator. For large values of \(b\), the ensemble estimate is given a larger weight.

Numerical results for an adaptive estimator using this procedure for estimating \(\sigma_i\), with \(b = 5\), are given in Tables 2.1 and 2.1. To avoid a prohibitive computational burden, we only update the average \(\bar{\sigma}\) at the end of each specific epoch. The results show that there is a modest to no loss in performance when the estimated \(\hat{\sigma}_i\) is used in place of the true \(\sigma_i\).

2.6 Conclusion

Two-level nested simulation can provide a more realistic assessment of financial risk, but with a considerable computational cost. In this chapter we propose a nested sequential simulation procedure that significantly reduces the computational burden. The savings are achieved by using a non-uniform inner sampling procedure that allocates more resources where the effect on the risk estimation is the greatest, which in turn allows relatively more effort to be devoted to the generation of outer scenarios. The combined effect produces a risk estimator, which converges at a faster rate to the true value. In numerical experiments, mean squared error was reduced by factors ranging from 4 to over 100.

\[\text{6 Numerically stable and efficient algorithms are available for updating sample variance calculations (see, e.g., Chan et al. (1983)). These would allow for rapid calculation of each } \tilde{\sigma}_i \text{ in an online fashion. However, once } \tilde{\sigma} \text{ is updated, every } \tilde{\sigma}_i \text{ will change. This will necessitate rebuilding the priority queue data structure for the scenarios, and may require order } n \text{ time.}\]
The sequential estimation procedure can be combined with previous research on variance reduction for the outer stage scenario generation to achieve further computational savings. The algorithms and results were presented in the context of estimating the probability of a large loss, but it may be possible to apply similar ideas to develop non-uniform algorithms for other risk measures. This remains an open area for future research.
Chapter 3

The Unweighted Regression Method

This chapter focuses on developing a risk estimation procedure via unweighted regression. The regression method uses spatial information of nearby outer stage scenarios to estimate portfolio losses. The algorithm and analysis are developed for the risk measures in the form of (1.1).

3.1 Delta-Gamma Approximation

Before discussing the regression method, we first introduce the delta-gamma approximation, which will be tested and compared with the regression method as a competing method. The delta-gamma approximation estimates \( L(\omega) \) with a quadratic function of \( \omega \). Consider a single representative scenario \( \omega^* \). For example, the scenario \( \omega^* \) can be selected to be the expected value of \( \omega \) under the real-world distribution at time \( \tau \). If \( L(\cdot) \) is a twice differentiable function in a neighborhood of \( \omega^* \), then a quadratic approximation can be made as

\[
\tilde{L}_{DG}(\omega) \triangleq L(\omega^*) + \nabla L(\omega^*)^\top (\omega - \omega^*) + \frac{1}{2} (\omega - \omega^*)^\top \nabla^2 L(\omega^*) (\omega - \omega^*). \tag{3.1}
\]

Given \( \tilde{L}_{DG} \), the risk measure \( \alpha \) can be estimated by

\[
\hat{\alpha}_{DG} \triangleq E \left[ f(\tilde{L}_{DG}(\omega)) \right]. \tag{3.2}
\]

This is referred to as the delta-gamma estimator. This method is discussed, for example, by Rouvinez (1997) and Glasserman et al. (2000).

\[1\] In this chapter, “regression” always refers to unweighted regression.
CHAPTER 3. THE UNWEIGHTED REGRESSION METHOD

Note that this quadratic approximation requires the knowledge of the portfolio loss \( L(\omega^*) \) as well as the delta (i.e., the gradient) \( \nabla L(\omega^*) \) and the gamma (i.e., the Hessian matrix) \( \nabla^2 L(\omega^*) \) at the representative scenario \( \omega^* \). In general, these quantities may be analytically unavailable in closed form. However, since they are only needed in a single scenario, they can be computed to arbitrary accuracy without excessive computational effort. For the purposes of our discussion, we assume they are known exactly.

Further, given the quadratic approximation \( \tilde{L}_{DG} \), we need to evaluate the expectation in (3.2). Observe that, for any scenario \( \omega \), evaluating \( \tilde{L}_{DG}(\omega) \) is easy since it involves only basic vector operations, and, in particular, requires no simulation. Thus, with a modest computational effort, we can perform a single stage Monte Carlo simulation to approximate the estimator \( \hat{\alpha}_{DG} \) to a high degree of accuracy. For the purpose of discussion, we assume the expectation in (3.2) can be exactly evaluated.

3.2 The Unweighted Regression Algorithm

We now introduce a method that is based on regression. The idea is to approximate the portfolio loss \( L(\cdot) \) by an approximation that is easy to evaluate. This is reminiscent of delta-gamma approximation. However, delta-gamma approximation has two major restrictions. First, it uses only a quadratic approximation. We allow higher order approximations, and allow approximations that can be tailored based on knowledge of the portfolio. Second, the delta-gamma approximation computes a local approximation around some representative scenario. There is no reason to expect that such an approximation will accurately describe the portfolio loss across a broad set of scenarios. On the other hand, we will attempt to find an approximation that is globally good.

We make the following standard assumption:

**Assumption 3.A1.** The second moment of the portfolio loss \( L(\omega) \) is finite, i.e., \( E[L(\omega)^2] < \infty \). The estimated loss \( \hat{L}(\omega, \zeta) \) satisfies

\[
E \left[ \hat{L}(\omega, \zeta) \bigg| \omega \right] = L(\omega),
\]

and

\[
\text{Var} \left( \hat{L}(\omega, \zeta) \bigg| \omega \right) = \frac{v(\omega)}{m} < \infty.
\]
Also, the conditional variance \( v(\omega) \) satisfies
\[
E[v(\omega)] < \infty.
\] (3.5)

Notice that [3.3] states that the portfolio loss estimate is unbiased. Equation [3.4] implies that the conditional variance of the portfolio loss estimate decays at the rate \( m^{-1} \) as a function of the number of inner stage samples \( m \), with a scenario-dependent constant \( v(\omega) \). Equation [3.5] implies that the portfolio loss estimate has finite second moment.

In particular, consider a set of \( d \) real-valued functions \( \phi_1(\cdot), \ldots, \phi_d(\cdot) \) on the state space \( \Omega \), which we will call basis functions. The basis functions can be written as a row vector
\[
\Phi(\omega) \triangleq \left( \phi_1(\omega), \ldots, \phi_d(\omega) \right) \in \mathbb{R}^d,
\]
for each scenario \( \omega \). We seek to approximate the portfolio loss function \( L(\cdot) \) by a linear combination of these basis functions. In other words, we would like to find a column vector \( r \in \mathbb{R}^d \) so that for each scenario \( \omega \),
\[
L(\omega) \approx \Phi(\omega)r \triangleq \sum_{\ell=1}^{d} \phi_{\ell}(\omega) r_{\ell}.
\]
We will then estimate the risk measure \( \alpha \) using this approximation.

There are two requirements for this procedure to be effective: First, the basis functions should incorporate features of the state space relevant to determining the portfolio loss, so that a linear combination of these functions can accurately approximate the portfolio loss. Second, the basis functions should be fast to evaluate. Then, when using an approximation defined by the basis functions, the outer stage expectation in the risk measure can be computed quickly. In general, the intelligent selection of basis functions is problem-dependent. We will see examples of this in the numerical case studies of Section 3.4.

Given the basis functions \( \Phi \), a global approximation to the portfolio loss \( L \) can be found by solving minimum mean squared error problem
\[
r^* \in \arg\min_{r \in \mathbb{R}^d} E \left[ (L(\omega) - \Phi(\omega)r)^2 \right].
\] (3.6)
We can then approximate the risk measure by \( E[f(\Phi(\omega)r^*)] \). Given the optimal regression coefficients \( r^* \), for each scenario \( \omega \), define \( M(\omega) \) to be the model error of the approximation \( \Phi(\omega)r^* \),
\[
M(\omega) \triangleq L(\omega) - \Phi(\omega)r^*.
\] (3.7)
Here, \( M(\cdot) \) represents the residual error under the best approximation afforded by the basis functions \( \Phi \). Further, define \( \varepsilon(\omega, \zeta) \) by

\[
\varepsilon(\omega, \zeta) \triangleq \hat{L}(\omega, \zeta) - L(\omega) = \hat{L}(\omega, \zeta) - \Phi(\omega) r^* - M(\omega).
\] (3.8)

This quantity measures the discrepancy between the Monte Carlo estimate of portfolio loss in the scenario \( \omega \) and the true portfolio loss.

In order for this regression procedure to be well defined, we make the following assumption:

**Assumption 3.A2.** The second moments of \( \phi_1(\cdot), \ldots, \phi_d(\cdot) \) are finite, i.e., \( \mathbb{E}[\phi_\ell(\omega)^2] < \infty \) for each \( \ell = 1, \ldots, d \). Further, \( \phi_1(\cdot), \ldots, \phi_d(\cdot) \) are linearly independent, i.e., when \( n \geq d \),

\[
P\left( \text{rank} \begin{pmatrix} \Phi(\omega^{(1)}) \\ \vdots \\ \Phi(\omega^{(n)}) \end{pmatrix} = d \right) = 1.
\]

Without loss of generality, we can assume that the functions \( \phi_1(\cdot), \ldots, \phi_d(\cdot) \) are orthonormal, i.e., we assume that \( \mathbb{E}[\Phi(\omega)^\top \Phi(\omega)] \) is the identity matrix.

Assumption 3.A2 ensures that the basis functions are linearly independent. If this is the case, given finite second moments, there is no loss of generality in assuming that they are orthonormal. Otherwise, one could construct an equivalent orthonormal basis through the Gram-Schmidt procedure. We will assume the orthonormality for the rest of Chapter 3 as it greatly simplifies the exposition.

Given Assumptions 3.A1 and 3.A2, the optimal solution \( r^* \) to the optimization problem (3.6) exists and is unique. However, it is not possible to directly compute \( r^* \), since we cannot evaluate \( L(\cdot) \) in general. Instead, our method seeks to solve an analog of the optimization problem (3.6) that is obtained by nested simulation. In particular, in order to get a tractable problem, we will first replace the expectation in (3.6) with a sample average over scenarios. Then, the portfolio loss in each scenario can be estimated by inner stage Monte Carlo simulation. As in Section 1.2, suppose there are \( n \) scenarios \( \bar{\omega} \triangleq (\omega^{(1)}, \ldots, \omega^{(n)})^\top \). In each scenario \( \omega^{(i)} \), let \( \zeta^{(i)} \) be an i.i.d. random variable that captures the randomness of the corresponding \( m \) inner stage samples, so that \( \hat{L}(\omega^{(1)}, \zeta^{(1)}), \ldots, \hat{L}(\omega^{(n)}, \zeta^{(n)}) \) are the nested Monte Carlo portfolio loss estimates across scenarios.
Define the vector \( \vec{\zeta} \triangleq (\zeta^{(1)}, \ldots, \zeta^{(n)})^\top \). Given \( (\vec{\omega}, \vec{\zeta}) \), we solve the optimization problem

\[
\hat{r} \in \arg\min_{r \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}(\omega^{(i)}, \zeta^{(i)}) - \Phi(\omega^{(i)})r \right)^2.
\] (3.9)

Given the coefficient vector \( \hat{r} \), we estimate the risk measure \( \alpha \) by

\[
\hat{\alpha}_{\text{REG}}(m,n) \triangleq \mathbb{E}\left[ f\left( \Phi(\omega)\hat{r} \right) \bigg| \vec{\omega}, \vec{\zeta} \right].
\] (3.10)

We define \( \hat{\alpha}_{\text{REG}}(m,n) \) to be the regression estimator. Given Assumptions 3.A1 and 3.A2, the optimal solution \( \hat{r} \) in (3.9) exists and is unique almost surely. Hence, our estimator is well defined.

**Remark 3.1.** Observe that the expectation in (3.10) can be estimated via a single stage Monte Carlo simulation that only requires evaluation of the basis functions. Since we assume that the basis functions are fast to compute, it will be possible to approximate \( \hat{\alpha}_{\text{REG}}(m,n) \) to a high degree of accuracy, given modest computational effort. Indeed, if we sample \( n' \) additional outer stage scenarios, and we estimate (3.10) with

\[
\frac{1}{n'} \sum_{i=1}^{n'} f(\Phi(\omega'_i)\hat{r}),
\] (3.11)

the MSE between (3.10) and (3.11) is in the order of \( (n')^{-1} \), which is asymptotically negligible compared to the MSE of the estimator \( \hat{\alpha}_{\text{REG}}(m,n) \) as we will show in Corollary 3.2. Therefore, it is justified to use (3.11) in practice. For the purpose of discussion, we assume (3.10) can be exactly computed.$^2$

### 3.3 Analysis

In this section, we provide theoretical analysis of the regression method presented in Section 3.2. Here we are interested in the asymptotic squared error of the regression method, as the number of scenarios \( n \) tends to infinity. Our analysis consists of two separate cases, with different assumptions on the loss function \( f(\cdot) \).

---

$^2$Besides nested sampling, some extra computational budget is needed for the regression step, i.e., the computation of (3.9). According to (B.1) in the appendix, the computational budget needed here is a linear function of the number of outer stage scenarios \( n \), i.e., the extra computational burden is \( O(n) \), which is asymptotically proportional to (if \( m \) is \( O(1) \)) or dominated by (if \( m \to \infty \)) the total required computation budget \( O(k) \) in nested simulation. Therefore, the computational requirement for (3.9) is ignored in the analysis.
In Section 3.3.1 we consider the first case, where the function $f(\cdot)$ is assumed to be twice differentiable. Here, we establish that the squared error converges in probability at the rate $n^{-1}$, until it reaches an asymptotic bias level at which the error ceases to improve. In Section 3.3.2 we consider the second case, where the function $f(\cdot)$ is assumed to be Lipschitz continuous. Here, we establish that the squared error converges in mean at the rate $n^{-1+\delta}$, for any $\delta > 0$, until it reaches an asymptotic bias level. In both cases, the asymptotic level of bias is bounded by the model error associated with the basis functions. While the exact assumptions and conclusions differ in the two analyses, taken together, the spirit of these results is to suggest that our regression approach will converge at the same rate as traditional non-nested Monte Carlo simulation over a large range, given a suitably good choice of basis functions.

All proofs for this section are deferred to Appendix B.

3.3.1 Differentiable Case

In the first case, we make the following differentiability assumption:

**Assumption 3.F1.** The function $f(\cdot)$ is twice differentiable with bounded second derivative, so that there exists a scalar $U_{\text{diff}}$ with

$$|f''(L)| \leq U_{\text{diff}},$$

(3.12)

for any $L \in \mathbb{R}$.

3.3.1.1 Asymptotic Distribution of the Estimator

In order to characterize the asymptotic distribution of the regression estimator, we make the following technical assumption:

**Assumption 3.A3.** The matrix $E\left[v(\omega) \Phi(\omega)\Phi(\omega)^{\top}\right]$ is positive definite, and

$$E\left[\phi_\ell(\omega)^2M(\omega)^2\right] < \infty,$$

for each $\ell = 1, \ldots, d$.

Assumption 3.A3 is a technical assumption standard in regression theory (see, e.g., White...
To begin our analysis, we have the following lemma that characterizes the convergence of coefficients of the regression estimator.

**Lemma 3.1.** Suppose Assumptions 3.A1, 3.A2, and 3.A3 hold. As the number of scenarios \( n \to \infty \),

\[
\sqrt{n} (\hat{r} - r^*) \xrightarrow{d} N \left( 0, \Sigma_M + \frac{\Sigma_v}{m} \right),
\]

where

\[
\Sigma_M \triangleq \mathbb{E} \left[ M^2 (\omega) \Phi(\omega)^\top \Phi(\omega) \right],
\]

and

\[
\Sigma_v \triangleq \mathbb{E} \left[ v(\omega) \Phi(\omega)^\top \Phi(\omega) \right].
\]

Therefore, as \( n \to \infty \),

\[
\|\hat{r} - r^*\|_2 = O_p(1)/\sqrt{n}.
\]

According to Lemma 3.1, as the number of scenarios \( n \to \infty \), the estimated coefficients \( \hat{r} \) converge to the optimal coefficients \( r^* \) at the rate \( n^{-1/2} \) in probability. However, we are interested in not only the convergence of regression coefficients, but also in the convergence of the resulting estimated risk measure given next. To this end, we have the following result:

**Theorem 3.1.** Suppose that Assumptions 3.F1, 3.A1, 3.A2, and 3.A3 hold. Then there exists a sequence of random variables \( \{B_{M,n}\} \), for \( n = 1, 2, \ldots \), satisfying

\[
B_{M,n} \xrightarrow{p} B^*_M \triangleq \mathbb{E} \left[ f(\Phi(\omega)r^*) \right] - \alpha,
\]

so that

\[
\sqrt{n} \left( \hat{\alpha}_{REG(m,n)} - \alpha - B_{M,n} \right) \xrightarrow{d} N \left( 0, \mathbb{E} \left[ f'(L(\omega)) \Phi(\omega) \right] \left( \Sigma_M + \frac{\Sigma_v}{m} \right) \left( \mathbb{E} \left[ f'(L(\omega)) \Phi(\omega) \right] \right)^\top \right),
\]

where \( \Sigma_M \) and \( \Sigma_v \) are defined by (3.13) and (3.14). Further, the asymptotic bias \( B^*_M \) satisfies

\[
|B^*_M - \mathbb{E} \left[ f'(L(\omega)) M(\omega) \right]| \leq \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (M(\omega))^2 \right].
\]

Let \( \xi_1, \xi_2, \ldots \) be a sequence of random vectors. If there exists a vector \( \xi^* \) such that for every \( b > 0 \),

\[
P \left( \|\xi_n - \xi^*\|_2 < b \right) \to 1 \text{ as } n \to \infty,
\]

then \( \xi_n \) converges to \( \xi^* \) in probability. We write this as \( \xi_n \xrightarrow{p} \xi^* \) or \( \xi_n \xrightarrow{L_2} \xi^* \) = \( O_p(1) \). If we denote by \( F_{\xi_n} \) and \( F_{\xi^*} \) the cumulative distribution functions of random variables \( \xi_n \) and \( \xi^* \), and if \( \lim_{n \to \infty} F_{\xi_n} = F_{\xi^*} \) at all continuity points of \( F_{\xi^*} \), then \( \xi_n \) converges to \( \xi^* \) in distribution. We write this as \( \xi_n \xrightarrow{d} \xi^* \).
Theorem 3.1 establishes the following three points. First, the bounds in (3.16) indicate that the asymptotic bias $B^*_M$ depends on the model error $M(\cdot)$, i.e., the quality of the best approximation under the basis functions. Second, we expect that the error of the estimator $\hat{\alpha}_{\text{REG}(m,n)} - \alpha$ decreases at the rate $n^{-1/2}$ in probability until it hits a term that converges to the asymptotic bias level at which point the estimator ceases to improve. Third, for large $n$, the quantity $\hat{\alpha}_{\text{REG}(m,n)} - \alpha - B_{M,n}$ has mean approximately zero and variance approximately

$$\frac{1}{n} E \left[ f'(L(\omega)) \Phi(\omega) \right] \left( \Sigma_M + \frac{\Sigma_v}{m} \right) \left( E \left[ f'(L(\omega)) \Phi(\omega) \right] \right)^\top.$$ 

Given a fixed number of inner stage samples $k = mn$, it is clear that the asymptotically minimum squared error is achieved when $m = 1$ and $n = k$.

In order to further interpret Theorem 3.1, consider, as a special case, the following corollary:

**Corollary 3.1.** Suppose that Assumptions 3.F1, 3.A1, 3.A2, and 3.A3 hold. When $m = 1$, $n = k$, and the portfolio loss $L$ is in the span of the basis functions $\Phi$, then $M(\omega) \equiv 0$, and we have

$$\sqrt{k} \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right) \overset{d}{\to} N \left( 0, E \left[ f'(L(\omega)) \Phi(\omega) \right] \Sigma_v \left( E \left[ f'(L(\omega)) \Phi(\omega) \right] \right)^\top \right),$$

as $k \to \infty$.

Moreover, suppose that the conditional variance of inner samples does not depend on the scenario, i.e., $v(\omega) \equiv v$, for every scenario $\omega$. Then, there exists a constant scalar $v^*$ defined by

$$v^* \triangleq v E \left[ f'(L(\omega)) \Phi(\omega) \right] \left( E \left[ f'(L(\omega)) \Phi(\omega) \right] \right)^\top \leq v E \left[ (f'(L(\omega)))^2 \right],$$

such that, as $k \to \infty$,

$$\sqrt{k} \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right) \overset{d}{\to} N \left( 0, v^* \right).$$

According to Corollary 3.1, the error scales as $k^{-1/2}$ in probability as a function of the total number of inner samples $k$. By applying the continuous mapping theorem, we have that, as $k \to \infty$, the quantity $k \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right)^2$ converges to a random variable. This is the same rate of convergence as the convergence rate in the case of non-nested Monte Carlo simulation.

### 3.3.2 Lipschitz Continuous Case

Section 3.3.1 established theoretical results when $f(\cdot)$ has bounded second derivative everywhere. Since many risk measures of interest are not twice differentiable, in this section, we investigate the convergence of the regression estimator under the alternative assumption of Lipschitz continuity:
Assumption 3.F2. The function \( f(\cdot) \) is Lipschitz continuous, i.e., there exists a scalar \( U_{\text{Lip}} \), such that
\[
|f(L') - f(L'')| \leq U_{\text{Lip}} |L' - L''|,
\]
for any \( L', L'' \in \mathbb{R} \).

3.3.2.1 Asymptotic Distribution of the Estimator

Under this assumption, we can bound the asymptotic squared error of the regression estimator \( \hat{\alpha}_{\text{REG}(m,n)} \) as follows:

**Theorem 3.2.** Suppose that Assumptions 3.F2, 3.A1, 3.A2, and 3.A3 hold. Then as the number of scenarios \( n \to \infty \),
\[
(\hat{\alpha}_{\text{REG}(m,n)} - \alpha)^2 \leq U_{\text{Lip}}^2 \mathbb{E} \left[ (M(\omega))^2 \right] + O_P \left( \frac{1}{n} \right).
\]

According to Theorem 3.2, the squared error is bounded above by a random variable that decays at the rate \( n^{-1} \) in probability plus a constant that is a function of the model error, i.e., the quality of the basis functions. This is analogous to the situation of Theorem 3.1 discussed earlier. As in that case, given a budget \( k = mn \) on the total number of inner stage samples, the bound on the asymptotic squared error is minimized when \( m = 1 \) and \( n = k \).

3.3.2.2 Mean Squared Error of the Estimator

The result of Theorem 3.2 provides a bound on the convergence of the squared error in probability. This can be strengthened by bounding the mean squared error of the regression estimator \( \hat{\alpha}_{\text{REG}(m,n)} \). In order to do so, we will apply the methodology of Shapiro et al. (2009). To this end, we make the following technical assumptions:

**Assumption 3.A4.** The moment generating functions of \( \|\Phi(\omega)\|^2_2 \), \( (M(\omega))^2 \), and \( (\varepsilon(\omega,\zeta))^2 \) are finite-valued in a neighborhood of zero.

In our problem, define
\[
G(r, \omega, \zeta) \triangleq \left( \tilde{L}(\omega, \zeta) - \Phi(\omega) r \right)^2,
\]
and
\[
g(r) \triangleq \mathbb{E} [G(r, \omega, \zeta)].
\]
In other words, \( G(r, \omega, \zeta) \) is the squared error of the regression estimate with coefficient vector \( r \) versus the standard nested estimate in a single scenario, and \( g(r) \) is the mean squared error across all scenarios.

For any \( \rho > 0 \), define the set

\[
\mathcal{R}_\rho \triangleq \left\{ r \in \mathbb{R}^d : \| r - r^* \|_2^2 \leq \rho \right\},
\]

which is a compact and convex neighborhood of \( r^* \).

**Assumption 3.A5.** For any \( \rho > 0 \), there exists a constant \( \lambda > 0 \) such that for any \( r', r'' \in \mathcal{R}_\rho \), the moment generating function \( \Psi_{r', r''}(t) \) of the random variable

\[
Y_{r', r''} \triangleq (G(r', \omega, \zeta) - g(r')) - (G(r'', \omega, \zeta) - g(r''))
\]

satisfies

\[
\Psi_{r', r''}(t) \leq \exp \left( \rho \lambda^2 t^2 \right),
\]

for any \( t \in \mathbb{R} \).

Assumption 3.A5 requires that \( Y_{r', r''} \) has sub-Gaussian tails. The proof of the following result is given in Appendix B.

**Lemma 3.2.** Suppose that Assumptions 3.F2, 3.A1, 3.A2, 3.A4, and 3.A5 hold. Let \( \rho > 0 \) be an arbitrary constant. Then for any positive integer \( n \),

\[
P(\hat{r} \notin \mathcal{R}_\rho) \leq \left( \frac{2 \sqrt{2} C'' \Lambda_{2\rho}}{\sqrt{\rho}} \right)^d \exp \left( - \frac{\rho n}{C'' \lambda^2} \right),
\]

where \( \lambda \) is defined in Assumptions 3.A5, \( C' \) and \( C'' \) are universal constants (i.e., constants that do not depend on the problem), and

\[
\Lambda_{\rho} \triangleq (2 \sqrt{\rho} + 1) d + 2E \left[ (M(\omega))^2 \right] + 2E \left[ (\varepsilon(\omega, \zeta))^2 \right].
\]

Lemma 3.2 bounds the probability that the estimated regression coefficients \( \hat{r} \) are not in the fixed neighborhood \( \mathcal{R}_\rho \) of the optimal coefficients \( r^* \), and demonstrates that this probability decays exponentially as \( n \to \infty \). Lemma 3.2 is not only an asymptotic result, but a finite-sample result that holds for every \( n \). Given Lemma 3.2, we can establish the following theorem:
Theorem 3.3. Suppose that Assumptions 3.F2, 3.A1, 3.A2, 3.A4, and 3.A5 hold, and let \( \delta > 0 \) be an arbitrary positive constant. Then for any positive integer \( n \),

\[
E \left[ (\Phi (\omega) (\hat{r} - r^*))^2 \right] \\
= E \left[ \|\hat{r} - r^*\|^2 \right] \\
\leq \frac{1}{n^{1-\delta}} + 2^{3d} C'(C'')^d (A_2)^d \lambda^2 n^{(1-\delta)d-1} \exp \left( -\frac{n^\delta}{C'\lambda^2} \right) + \frac{2^d C'(C'')^d (A_2)^d \lambda^2}{n} \exp \left( -\frac{n}{C'\lambda^2} \right) \\
= O \left( n^{-1+\delta} \right).
\]

Theorem 3.3 establishes that the squared error between our approximation of the loss function and the best possible approximation using the same basis functions decays at the rate \( n^{-1+\delta} \) for any \( \delta > 0 \). A corollary of Theorem 3.3 is our main result, which establishes the rate of convergence of the MSE of the risk estimator:

Corollary 3.2. Suppose that Assumptions 3.F2, 3.A1, 3.A2, 3.A4, and 3.A5 hold, and let \( \delta > 0 \) be an arbitrary positive constant. Then, for any positive integer \( n \),

\[
E \left[ (\hat{\alpha}_{\text{REG}(m,n)} - \alpha)^2 \right] \\
\leq U_{\text{Lip}}^2 \left( 2^{3d} C'(C'')^d (A_2)^d \lambda^2 n^{(1-\delta)d-1} \exp \left( -\frac{n^\delta}{C'\lambda^2} \right) + \frac{2^d C'(C'')^d (A_2)^d \lambda^2}{n} \exp \left( -\frac{n}{C'\lambda^2} \right) \right) \\
+ U_{\text{Lip}}^2 \left( E \left[ (M(\omega))^2 \right] + n^{-1+\delta} \right) \\
= U_{\text{Lip}}^2 E \left[ (M(\omega))^2 \right] + O \left( n^{-1+\delta} \right).
\]

According to Corollary 3.2, first, the MSE of the regression estimator \( \hat{\alpha}_{\text{REG}(m,n)} \) decays at the rate \( n^{-1+\delta} \) for any \( \delta > 0 \) until it hits an asymptotic bias level, which is a function of the model error; this is reminiscent of Theorem 3.2 except that we have \( \delta \) in the exponent of \( n^{-1+\delta} \) here. Second, Corollary 3.2 holds for any arbitrary \( n \), which is a finite-sample result rather than an asymptotic result in Theorem 3.2. Third, Corollary 3.2 implies the convergence of the MSE, which is stronger than the convergence of the squared error in probability in Theorem 3.2.

3.4 Numerical Results

In this section we use several examples to compare the relative performance of standard nested simulation, the delta-gamma approximation, and our regression method. In Section 3.4.2 we
present a simple example where there is only one underlying asset. In Section 3.4.3, we have three examples with portfolios of multiple assets.

3.4.1 Experimental Setting

Our examples involve portfolios consisting of one or more underlying assets as well as derivatives based on them. We assume that all underlying asset prices follow geometric Brownian motion processes, and that option prices are determined according to the standard single-asset Black-Scholes model and its multi-asset generalization.

Specifically, assume risk factors \( \omega \triangleq (\omega_1, \ldots, \omega_Q) \in \Omega \subset \mathbb{R}^Q \) are distributed according to a multivariate Gaussian distribution with mean zero, variance one, and a \( Q \)-by-\( Q \) correlation matrix. Given \( \omega \) and the risk horizon \( \tau \), define \( S_\tau(\omega) \) to be the prices of underlying assets at time \( \tau \), with

\[
S_\tau(\omega) \triangleq (S_{1,\tau}(\omega), \ldots, S_{Q,\tau}(\omega)),
\]

where

\[
S_{j,\tau}(\omega) = S_{j,0} \exp \left( (\mu_j - \frac{\sigma_j^2}{2}) \tau + \sigma_j \sqrt{\tau} \omega_j \right).
\]

Here \( S_{j,0} \) is the price of the \( j \)th asset at time 0, \( \mu_j \) is the drift of the \( j \)th asset under the real-world distribution, and \( \sigma_j \) is the annual volatility of the \( j \)th asset. In this setting, asset prices are lognormally distributed and there is exactly one risk factor per asset.

To estimate the portfolio loss at time \( \tau \) we use Monte Carlo simulation of the inner stage sample paths under the risk-neutral distribution to generate asset cashflows between times \( \tau \) and \( T \). For the \( j \)th security and for each inner stage sample path \( p = 1, \ldots, m \), define \( W_{j,t}^{(p)} \) to be a Brownian motion for \( t \in [\tau, T] \) with \( W_{j,\tau}^{(p)} = 0 \). Notice that there is no correlation between \( W_{j',t}^{(p')} \) and \( W_{j'',t}^{(p'')} \) if \( p' \neq p'' \), i.e., the \( m \) inner stage sample paths are independent; on the other hand, there could exist correlation between securities on the same path. The set

\[
\zeta \triangleq \left\{ W_{j,t}^{(p)}, \text{for} \ t \in [\tau, T], \ j = 1, \ldots, Q, \ p = 1, \ldots, m \right\}
\]

represents all of the inner stage uncertainty given the outer stage scenario \( \omega \). Specifically, conditional on \( \omega \), we assume that the risk-neutral asset prices on the \( p \)th sample path are given by

\[
S_{j,t}^{(p)}(\omega, \zeta) = S_{j,\tau}(\omega) \exp \left( (\tau_j - \frac{\sigma_j^2}{2}) (t - \tau) + \sigma_j W_{j,t}^{(p)} \right),
\]
for $t \in [\tau, T]$ and $j = 1, \ldots, Q$, where $r_f$ is the continuously compounded riskless rate of interest.

For each $p = 1, \ldots, m$, the discounted portfolio cashflows along the $p$th sample path may, in general, depend on asset prices at all times between $\tau$ and $T$. Typically, however, portfolio cashflows only depend on prices at a finite number of times, so only these essential points are simulated. The portfolio loss estimate $\hat{L}(\omega, \zeta)$ is determined by the average over the $m$ sample paths. We focus on the expected excess loss risk measure

$$E \left[ (L(\omega) - c)^+ \right],$$

(3.22)

where $c$ is a threshold that will be specified in each example.

In each of the following examples, we compare the accuracy of a number of methods that estimate the risk measure. The closed form expression for the portfolio losses $L(\omega)$ given a risk factor scenario $\omega$ is known in these examples. We will precisely compute the risk measure $\alpha$ by using non-nested Monte Carlo simulation in the outer stage with a Sobol sequence (see, e.g., Sobol (1967); Glasserman (2004)). Similarly, the delta-gamma estimator $\hat{\alpha}_{DG}$ and the regression estimator $\hat{\alpha}_{REG(m,n)}$ provide approximations of portfolio losses, and also require non-nested simulation in the outer stage, and thus we also estimate the risk measure $\alpha$ by simulating the outer stage with the same Sobol sequence.

### 3.4.2 Single Asset Example

- **Model:** There is one asset with initial price $S_0 = 100$, real-world drift $\mu = 8\%$, and volatility $\sigma = 20\%$. The interest rate is $r_f = 3\%$.

- **Portfolio:** The portfolio consists of three barrier options that mature at time $T = 1/12$ year. The risk horizon is $\tau = 1/52$ year.

---

4In the delta-gamma approximation, we will approximate the loss as a quadratic function of $S_\tau(\omega)$ instead of a quadratic function of $\omega$.

5In this thesis, all barrier options are partial-time barrier options (see, e.g., Section 4.17.4, Haug (2006)) that can be knocked in or out only between times $\tau$ and $T$.

6Along each sample path, the cashflow of a barrier option depends only on minimum underlying asset price and the final asset price. We simulate these two quantities instead of sampling the entire sample path. Details can be found in Hui (1997) and Metwally and Atiya (2002).
(a) Long one down-and-out put option with strike $K_1 = 101$ and barrier $H_1 = 91$.

(b) Long one down-and-out put option with strike $K_2 = 110$ and barrier $H_2 = 100$.

(c) Short one down-and-out put option with strike $K_3 = 114.5$ and barrier $H_3 = 104.5$.

The loss threshold is set to $c = 0.3608$, the 95th percentile of $L(\omega)$.

- **Basis functions**: We test several sets of basis functions. For simplicity, denote the asset prices at time $\tau$ by $S_\tau(\omega)$.

  $\Phi^{(1)}$: This basis set includes quadratic functions of $S_\tau(\omega)$. Specifically, $\Phi^{(1)}(\omega)$ consists of 
  
  $\phi_1(\omega) = 1$, $\phi_2(\omega) = S_\tau(\omega)$, and $\phi_3(\omega) = (S_\tau(\omega))^2$.

  $\Phi^{(2)}$: This basis set includes all elements in $\Phi^{(1)}$ (i.e., $\phi_\ell(\omega)$ for $\ell = 1, 2, 3$) as well as $\phi_4(\omega) = (S_\tau(\omega) - H_1)^+$, $\phi_5(\omega) = (S_\tau(\omega) - H_2)^+$, $\phi_6(\omega) = (S_\tau(\omega) - H_3)^+$, and $\phi_7(\omega)$, $\phi_8(\omega)$, $\phi_9(\omega)$ defined as the squares of $\phi_4(\omega)$, $\phi_5(\omega)$, $\phi_6(\omega)$, respectively.

  $\Phi^{(3)}$: This basis set includes all elements of $\Phi^{(2)}$ (i.e., $\phi_\ell(\omega)$ for $\ell = 1, \ldots, 9$) as well as $\phi_{10}(\omega) = L(\omega)$, the exact expression for the loss determined using analytical formulas to value the derivatives.

The basis set $\Phi^{(1)}$ is chosen because the regression method with this set is comparable to the delta-gamma approximation. The basis set $\Phi^{(2)}$ is tested because the portfolio loss is known in advance to depend on the barrier levels. The basis set $\Phi^{(3)}$ represents an ideal case where the basis functions span the true portfolio loss function and there is no model error.

For each method, we simulate 1,000 independent trials to estimate the MSE of the estimator. Results are given in Figure 3.1. The MSE of the standard nested estimator decays at $k^{-2/3}$, which is consistent with the discussion in Section 2.1. Observe that the delta-gamma estimator is the least accurate of the five estimators. Figure 3.2 shows the local delta-gamma approximation around $S_\tau(\omega^*)$ does not provide a good global approximation. The regression method with basis set $\Phi^{(1)}$ fits a quadratic function globally and leads to a better risk estimate than the local approximation used in the delta-gamma approach.

In Figure 3.1, the MSE of the regression method with basis set $\Phi^{(1)}$ initially decays at the rate $k^{-1}$, but then stops decreasing when the MSE reaches the asymptotic bias level determined by the regression model error. Basis sets $\Phi^{(2)}$ and $\Phi^{(3)}$ perform much better, with their MSE’s decaying at
Figure 3.1: Illustration of the mean squared error in the single asset example. The vertical axis shows the mean squared error, and the horizontal axis represents the total number of inner stage samples.
\( c = 0.3608 \)

\( S_\tau(\omega^*) = 100.1540 \)

Figure 3.2: Illustration of approximations in the single asset example. The vertical axis shows the portfolio loss, either true or estimated, and the horizontal axis represents the underlying asset price \( S_\tau(\omega) \) at time \( \tau \).

the rate \( k^{-1} \) over the current range in Figure 3.1. In fact, over the current range of simulation, the regression method with basis set \( \Phi^{(2)} \), which is the basis set chosen more intelligently than \( \Phi^{(1)} \), does about as well as the regression method with \( \Phi^{(3)} \), while \( \Phi^{(3)} \) contains the actual loss function in the basis set. Although the MSE of the regression method with \( \Phi^{(2)} \) decays over a larger range than the one with \( \Phi^{(1)} \), eventually, according to our theoretical results, the MSE of the regression method with \( \Phi^{(2)} \) will also flatten out.

3.4.3 Multiple Asset Examples

In this section we test three multi-asset examples from Glasserman et al. (2000). Common properties of these three examples are summarized here. Asset prices, denoted by \( S_{j,t}(\omega) \), follow independent geometric Brownian motion processes unless otherwise stated. Initial asset prices are all \( S_{j,0} = 100 \). Assets share common real-world drifts of \( \mu = 8\% \) and annual volatilities of \( \sigma = 30\% \).
The interest rate is \( r_f = 5\% \). Derivatives all have strikes \( K = 100 \) and a common maturity \( T = 0.1 \) year. The risk horizon is \( \tau = 0.04 \) year.

1. (Ex10:) Ten underlying assets and a portfolio of derivatives with discontinuous payoff functions.

- **Portfolio:** The portfolio consists of three derivatives on each of the 10 assets.
  
  (a) Short 10 down-and-out call options with barrier \( H = 95 \).
  
  (b) Short 5 cash-or-nothing put options.
  
  (c) An amount of underlying asset that delta hedges the two derivatives above.

   The loss threshold is \( c = 306.8763 \), the 99th percentile of \( L(\omega) \).

- **Basis functions:**

  \( \Phi^{(1)} \): This basis set includes quadratic functions of \( S_{j,\tau}(\omega) \). Specifically, \( \Phi^{(1)}(\omega) \) consists of \( \phi_\ell(\omega) \) for \( \ell = 1, \ldots, 21 \), where \( \phi_1(\omega) = 1, \phi_{1+i}(\omega) = S_{i,\tau}(\omega) \) and \( \phi_{11+i}(\omega) = (\phi_{1+i}(\omega))^2 \) for \( i = 1, \ldots, 10 \).

  \( \Phi^{(2)} \): This basis set includes fifth order polynomials of \( S_{j,\tau}(\omega) \)'s and fifth order polynomials of \( (S_{j,\tau}(\omega) - H)^+ \) for \( j = 1, \ldots, 10 \). Specifically, \( \Phi^{(2)}(\omega) \) consists of \( \phi_\ell(\omega) \) for \( \ell = 1, \ldots, 101 \), where \( \phi_1(\omega) = 1, \phi_{1+i}(\omega) = S_{i,\tau}(\omega) \) and \( \phi_{11+i}(\omega) = (S_{i,\tau}(\omega) - H)^+ \) for \( i = 1, \ldots, 10 \), and \( \phi_{21+i}(\omega) = (\phi_{1+i}(\omega))^2, \phi_{41+i}(\omega) = (\phi_{1+i}(\omega))^3, \phi_{61+i}(\omega) = (\phi_{1+i}(\omega))^4, \) and \( \phi_{81+i}(\omega) = (\phi_{1+i}(\omega))^5 \) for \( i = 1, \ldots, 20 \).

  \( \Phi^{(3)} \): This basis set includes all elements in \( \Phi^{(2)} \) as well \( \phi_{102}(\omega) = L(\omega) \), the exact expression for the loss determined using the analytical formulas to value the derivatives.

2. (Ex10E) Ten underlying assets with a portfolio consisting of exchange options whose payoffs depend on pairs of underlying assets.

- **Portfolio:** Short 10 exchange options on each of five asset pairs \( j \) and \( j + 5 \), for \( j = 1, \ldots, 5 \). The payoff of the \( j \)th exchange option is \( \max(0, S_{j,T} - S_{j+5,T}) \).

  The loss threshold is \( c = 278.8783 \), the 99th percentile of \( L(\omega) \).

- **Basis functions:**
\(\Phi^{(1)}\): This basis set consists of bivariate quadratic functions of \(S_{j,\tau}(\omega)\) and \(S_{j+5,\tau}(\omega)\) for \(j = 1, \ldots, 5\). Specifically, \(\Phi^{(1)}(\omega)\) contains \(\phi_1(\omega)\) for \(\ell = 1, \ldots, 26\), where \(\phi_1(\omega) = 1\), \(\phi_{1+\iota}(\omega) = S_{\iota,\tau}(\omega)\), \(\phi_{6+\iota}(\omega) = S_{\iota+5,\tau}(\omega)\), \(\phi_{11+\iota}(\omega) = (S_{\iota,\tau}(\omega))^2\), \(\phi_{16+\iota}(\omega) = S_{\iota,\tau}(\omega) \cdot S_{\iota+5,\tau}(\omega)\), and \(\phi_{21+\iota}(\omega) = (S_{\iota+5,\tau}(\omega))^2\) for \(\iota = 1, \ldots, 5\).

\(\Phi^{(2)}\): This basis set includes bivariate fifth order polynomials of \(S_{j,\tau}(\omega)\) and \(S_{j+5,\tau}(\omega)\) for \(j = 1, \ldots, 5\). Specifically, \(\Phi^{(1)}(\omega)\) consists of \(\phi_\ell(\omega)\) for \(\ell = 1, \ldots, 101\), where \(\phi_\ell(\omega)\) for \(\ell = 1, \ldots, 26\) are defined as in \(\Phi^{(1)}(\omega)\), and for the third-, fourth- and fifth-order terms, we have the corresponding powers of \(S_{j,\tau}(\omega)\) for \(j = 1, \ldots, 10\) as well as the cross terms of \(S_{j,\tau}(\omega)\) and \(S_{j+5,\tau}(\omega)\) for \(j = 1, \ldots, 5\).

\(\Phi^{(3)}\): This basis set includes all elements in \(\Phi^{(2)}\) as well as \(\phi_{102}(\omega) = L(\omega)\), the exact expression for the loss determined using the analytical formulas to value the derivatives.

3. (Ex100) There are 100 underlying assets with non-zero correlations.

- **Model:** The assets are divided into 10 groups of 10 assets each. The correlation is 20% between assets in the same group and is 0% otherwise. The assets in the first three groups have a volatility of 50%, those in the next four groups have a volatility of 30%, and those in the last three groups have a volatility of 10%.

- **Portfolio:** Short 10 at-the-money calls and 10 at-the-money puts on each of the 100 underlying assets.

The loss threshold is \(c = 876.8636\), the 99th percentile of \(L(\omega)\).

- **Basis functions:**

  \(\Phi^{(1)}\): This basis set includes quadratic functions of \(S_{j,\tau}(\omega)\) for \(j = 1, \ldots, 100\). Specifically, \(\Phi^{(1)}(\omega)\) consists of \(\phi_\ell(\omega)\) for \(\ell = 1, \ldots, 201\), where \(\phi_1(\omega) = 1\), \(\phi_{1+\iota}(\omega) = S_{\iota,\tau}(\omega)\) and \(\phi_{101+\iota}(\omega) = (\phi_{1+\iota}(\omega))^2\) for \(\iota = 1, \ldots, 100\).

  \(\Phi^{(2)}\): This basis set includes fifth order polynomials of \(S_{j,\tau}(\omega)\) for \(j = 1, \ldots, 100\). Specifically, \(\Phi^{(1)}(\omega)\) consists of \(\phi_\ell(\omega)\) for \(\ell = 1, \ldots, 501\), where \(\phi_1(\omega) = 1\), \(\phi_{1+\iota}(\omega) = S_{\iota,\tau}(\omega)\), \(\phi_{101+\iota}(\omega) = (\phi_{1+\iota}(\omega))^2\), \(\phi_{201+\iota}(\omega) = (\phi_{1+\iota}(\omega))^3\), \(\phi_{301+\iota}(\omega) = (\phi_{1+\iota}(\omega))^4\), \(\phi_{401+\iota}(\omega) = (\phi_{1+\iota}(\omega))^5\) for \(\iota = 1, \ldots, 100\).
Φ(3): This basis set includes all elements in Φ(2) as well as \( \phi_{502}(\omega) = L(\omega) \), the exact expression for the loss determined using the analytical formulas to value the derivatives.

For each method, we simulate 100 independent trials to estimate the MSE of the estimator. Results are given in Figures 3.3–3.5. In each example the MSE of the standard nested estimator decays at \( k^{-2/3} \), as expected. The delta-gamma estimator is fast to compute but not very accurate, since the local approximation does not capture important features of the portfolio loss functions. The regression method with basis set Φ(3) works the best in each example, with the MSE decaying at rate \( k^{-1} \). With more realistic basis sets Φ(1) and Φ(2), the MSE’s decay at the rate \( k^{-1} \) until hitting the asymptotic bias levels, which is often beyond the scale of the figures. MSE results for \( k = 5,000,000 \) are summarized in Table 3.1.

Basis function sets were chosen for simplicity and consistency across examples. Better choices of basis functions, e.g., using the analytical expressions for vanilla derivative prices and powers of these expressions, may lead to smaller asymptotic bias levels.

3.5 Conclusion

In this chapter we propose a new risk estimation method based on Monte Carlo simulation and regression. Given \( n \) outer stage scenarios and \( mn \) inner stage samples, we show that the optimal choice of \( m \) and \( n \) are \( m^* = 1 \) and \( n^* = k \). With these choices, our theoretical results show that the mean squared error diminishes at a rate close to \( k^{-1} \) until a non-diminishing bias level is reached. The proposed regression method outperforms standard nested simulation and the delta-gamma method when used with a “good” set of basis functions as long as \( k \) is not too large. More importantly, before hitting the bias level, the proposed method recovers the standard \( k^{-1} \) convergence rate of non-nested unbiased simulation estimators. Numerical results showed that the regression method significantly outperforms other methods and illustrated consistency with the theoretical results.
Figure 3.3: Illustration of the mean squared error in example Ex_{10}. The vertical axis shows the mean squared error, and the horizontal axis represents the total number of inner stage samples.
Figure 3.4: Illustration of the mean squared error in example Ex_{10E}. The vertical axis shows the mean squared error, and the horizontal axis represents the total number of inner stage samples.
Figure 3.5: Illustration of the mean squared error in example Ex100. The vertical axis shows the mean squared error, and the horizontal axis represents the total number of inner stage samples.
### Table 3.1: MSE results for the four examples. The results are computed over independent trials (1,000 in the single asset example and 100 in the other three), each with a total simulation budget of $k = 5,000,000$. The last column contains normalize MSE results relative to the regression estimator with basis set $\Phi^{(3)}$. 

<table>
<thead>
<tr>
<th>Example</th>
<th>MSE Optimal Standard Nested Estimator</th>
<th>MSE Delta-Gamma</th>
<th>MSE Regression with Basis Set $\Phi^{(1)}$</th>
<th>MSE Regression with Basis Set $\Phi^{(2)}$</th>
<th>MSE Regression with Basis Set $\Phi^{(3)}$</th>
<th>MSE Norm. to Regression with Basis Set $\Phi^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single asset example</td>
<td>1.0·10^{-5}</td>
<td>333.9</td>
<td>3.9·10^{-4}</td>
<td>12,414.1</td>
<td>7.7·10^{-8}</td>
<td>2.5</td>
</tr>
<tr>
<td>Example Ex10</td>
<td>1.4·10^{-2}</td>
<td>1,107.9</td>
<td>3.5·10^{-1}</td>
<td>28,761.4</td>
<td>1.7·10^{-5}</td>
<td>1.4</td>
</tr>
<tr>
<td>Example Ex10E</td>
<td>1.4·10^{-3}</td>
<td>311.7</td>
<td>1.9·10^{-3}</td>
<td>411.2</td>
<td>8.1·10^{-6}</td>
<td>1.8</td>
</tr>
<tr>
<td>Example Ex100</td>
<td>1.5·10^{-2}</td>
<td>147.5</td>
<td>3.9·10^{-1}</td>
<td>3,853.4</td>
<td>7.1·10^{-5}</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Chapter 4

The Weighted Regression Method

This chapter follows the idea in Chapter 3 and develops a risk estimation procedure via weighted regression, which not only uses spatial information of scenarios but also emphasizes scenarios more important to the calculation of the risk measure. Similar to Chapter 3, the algorithm and analysis in this chapter are developed for the risk measures in the form of (1.1).

4.1 The Weighted Regression Algorithm

When using the regression estimator $\hat{\alpha}_{\text{REG}(m,n)}$, all outer stage scenarios are weighted uniformly in the sample mean squared error objective of (3.9). This is reasonable if we seek a globally good approximation over the entire scenario space $\Omega$. However, when dealing with specific risk measures, the portfolio loss estimates in some scenarios deserve more precise estimation than those in other scenarios. For example, when estimating the expected excess loss, scenarios with large losses contribute substantially to the estimator. Scenarios with large profits, on the other hand, do not directly impact the risk calculation, so the accuracy of their estimates is not as important. Since scenarios are not equally important, a weighted regression scheme where larger weights are assigned to more important scenarios is logical.

In order to analyze further, we need the following assumptions. Suppose $h: \Omega \rightarrow \mathbb{R}$ is a non-negative function.

Assumption 4.A1. The second moment of $\sqrt{h(\omega)L(\omega)}$ is finite, i.e., $E[h(\omega)L(\omega)^2] < \infty$. The
estimated loss $\hat{L}(\omega, \zeta)$ satisfies
\[ E \left[ \hat{L}(\omega, \zeta) \mid \omega \right] = L(\omega), \]
and
\[ \text{Var} \left( \hat{L}(\omega, \zeta) \mid \omega \right) = \frac{v(\omega)}{m} < \infty. \]
Also, $h(\omega)$ and $v(\omega)$ satisfy
\[ E[h(\omega)v(\omega)] < \infty. \]

Assumption 4.A2. The second moments of $\sqrt{h(\cdot)}\phi_1(\cdot), \ldots, \sqrt{h(\cdot)}\phi_d(\cdot)$ are finite, i.e.,
\[ E[h(\omega)\phi_\ell(\omega)^2] < \infty, \]
for $\ell = 1, \ldots, d$. Further, if $\omega^{(1)}, \ldots, \omega^{(n)}$ are i.i.d. scenarios, then the matrix with rows $\sqrt{h(\omega^{(i)})}\Phi(\omega^{(i)})$, for $i = 1, \ldots, n$, has full rank almost surely. Moreover, we assume that the functions $\phi_1(\cdot), \ldots, \phi_d(\cdot)$ are orthonormal, i.e., we assume that $E[\Phi(\omega)^\top \Phi(\omega)]$ is the identity matrix.

Notice that Assumptions 4.A1 and 4.A2 are reminiscent of Assumptions 3.A1 and 3.A2. Using $h(\cdot)$ as the weight function, we define
\[ r^*(h) \in \arg\min_{r \in \mathbb{R}^d} E \left[ h(\omega)(L(\omega) - \Phi(\omega)r)^2 \right]. \] (4.1)
For any specific $h$ that satisfies Assumptions 4.A1 and 4.A2, the optimal solution $r^*(h)$ to (4.1) exists and is unique. Given outer stage scenarios $\omega^{(1)}, \ldots, \omega^{(n)}$, we assign $h(\omega^{(i)})$ to be the weight of the scenario $\omega^{(i)}$ in the regression, i.e., with $\vec{h} \triangleq (h(\omega^{(1)}), \ldots, h(\omega^{(n)}))^\top$. Since $L(\omega^{(1)})$ is unobservable in practice, we use the Monte Carlo portfolio loss estimates $\hat{L}(\omega^{(i)}), \zeta^{(i)})$, $i = 1, \ldots, d$, and estimate the regression coefficients by
\[ \hat{r}(\vec{h}) \in \arg\min_{r \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n h(\omega^{(i)})(\hat{L}(\omega^{(i)}, \zeta^{(i)}) - \Phi(\omega^{(i)})r)^2. \] (4.2)
The weighted regression estimator is
\[ \hat{\alpha}_{\text{REG}(m,n,h)} \triangleq E \left[ f(\Phi(\omega)\hat{r}(\vec{h})) \right] \mid \vec{\omega}, \vec{\zeta}. \] (4.3)
For any specific $h$ that satisfies Assumptions 4.A1 and 4.A2, the optimal solution $\hat{r}(\vec{h})$ to (4.2) exists and is unique almost surely when $n \geq d$, so our estimator $\hat{\alpha}_{\text{REG}(m,n,h)}$ is well defined. By following the same argument as in Remark 3.1, we assume $\hat{\alpha}_{\text{REG}(m,n,h)}$ can be exactly computed.

Remark 4.1. Notice that the scale of $h(\cdot)$ has no impact on $\hat{r}(\vec{h})$, i.e., if we multiply $h(\cdot)$ by a constant factor, $\hat{\alpha}_{\text{REG}(m,n,h)}$ will not change.
4.2 Analysis

Consider the MSE of the weighted regression estimator $\hat{\alpha}_{\text{REG}(m,n,h)}$:

$$
E[(\hat{\alpha}_{\text{REG}(m,n,h)} - \alpha)^2] = E \left( \left( E[f(\Phi(\omega)\tilde{r}(\tilde{h})) | \tilde{\omega}, \tilde{\zeta}] - E[f(L(\omega))] \right)^2 \right). \tag{4.4}
$$

In order to analyze (4.4), we need the following technical assumptions.

**Assumption 4.F1.** The function $f(\cdot)$ is Lipschitz continuous, i.e., there exists a scalar $U_{\text{Lip}}$, such that

$$
|f(L') - f(L'')| \leq U_{\text{Lip}} |L' - L''|,
$$

for any $L', L'' \in \mathbb{R}$.

**Assumption 4.A3.** The moment generating functions of $h(\omega) \|\Phi(\omega)\|_2^2$, $h(\omega)(L(\omega) - \Phi(\omega)r)^2$, and $h(\omega)(\tilde{L}(\omega, \zeta) - L(\omega))^2$ are finite-valued in a neighborhood of zero.

Define

$$
G(r, \omega, \zeta, h) \triangleq h(\omega) \left( \tilde{L}(\omega, \zeta) - \Phi(\omega)r \right)^2, \tag{4.5}
$$

and

$$
g(r, h) \triangleq E[G(r, \omega, \zeta, h)]. \tag{4.6}
$$

Given $r$, the function $G(r, \omega, \zeta, h)$ is the squared error of a regression estimate $\Phi(\omega)r$ versus the Monte Carlo estimate in scenario $\omega$, and the function $g(r, h)$ is the mean over scenarios of the squared error. For any $\rho > 0$, $R_\rho$ is defined by (3.21). We make the following additional assumption.

**Assumption 4.A4.** For any $\rho > 0$, there exists a constant $\lambda > 0$ such that for any $r', r'' \in R_\rho$, the moment generating function $\Psi_{r', r''}(t)$ of the random variable

$$
Y_{r', r''} \triangleq (G(r', \omega, \zeta, h) - g(r', h)) - (G(r'', \omega, \zeta, h) - g(r'', h))
$$

satisfies

$$
\Psi_{r', r''}(t) \leq \exp(\rho \lambda t^2),
$$

for any $t \in \mathbb{R}$. 
Similar to Assumption 3.A5, Assumption 4.A4 requires that $Y_{r',r''}$ has sub-Gaussian tails. (Notice that here $Y_{r',r''}$ depends on $h$.) We have the following result, whose proof is given in Appendix C:

**Theorem 4.1.** For a specific $h$, if Assumptions 4.F1, 4.A1, 4.A2, 4.A3, and 4.A4 hold, then

$$
\lim_{n \to \infty} \mathbb{E} \left[ (\hat{\alpha}_{\text{REG}}(m,n,h) - \alpha)^2 \right] = \left( \mathbb{E} \left[ f(\Phi(\omega)r^*(h)) \right] - \mathbb{E} \left[ f(L(\omega)) \right] \right)^2.
$$

(4.7)

As the number of outer stage scenarios $n \to \infty$, Theorem 4.1 shows that the MSE of $\hat{\alpha}_{\text{REG}}(m,n,h)$ diminishes until reaching the level of the asymptotic bias, and does not vanish in the limit. If the portfolio loss $L$ is not in the span of the basis functions $\Phi$, even with large $m$ and $n$, the bias induced by the regression method exists in general, which leaves the limit of (4.7) non-zero. We make the following assumption.

**Assumption 4.F2.** The function $f(\cdot)$ has first derivative almost everywhere.

Given Assumption 4.F2, we apply Jensen’s inequality and heuristically apply a Taylor approximation,

$$
(E \left[ f(\Phi(\omega)r^*(h)) \right] - E \left[ f(L(\omega)) \right])^2 \leq E \left[ (f(\Phi(\omega)r^*(h)) - f(L(\omega)))^2 \right]
$$

$$
\approx E \left[ (f'(L(\omega)))^2 (\Phi(\omega)r^*(h) - L(\omega))^2 \right].
$$

(4.8)

Instead of finding the choice of weight function that minimizes the limit of (4.7), for reasons of tractability, we consider optimizing the upper bound given by (4.8). Here, we seek to find regression approximations that are close to the true portfolio loss emphasizing scenarios $\omega$ where $f'(L(\omega))$ is large. With (4.8) as our objective, we have the following optimization problem to determine a choice of weight function:

$$
\text{minimize} \quad E \left[ (f'(L(\omega)))^2 (\Phi(\omega)r^*(h) - L(\omega))^2 \right]
$$

subject to $E[h(\omega)] > 0,$

$$
h(\omega) \geq 0, \quad \forall \omega \in \Omega.
$$

(4.9)

In the following theorem, we find a globally optimal solution to the problem (4.9).

**Theorem 4.2.** Assume that Assumption 4.F2 holds. Define the weight function by

$$
h_{\text{opt}}(\omega) = (f'(L(\omega)))^2,
$$

(4.10)

and assume that $E \left[ (f'(L(\omega)))^2 \right] > 0$. Then $h_{\text{opt}}$ is a globally optimal solution to the optimization problem (4.9).
Proof. With \( r^* \) defined by (4.1), for any \( h \) in the domain of (4.9),

\[
E \left[ \left( f'(L(\omega)) \right)^2 (\Phi(\omega)r^*(h) - L(\omega))^2 \right] = E \left[ h_{opt}(\omega)(\Phi(\omega)r^*(h) - L(\omega))^2 \right] \geq E \left[ h_{opt}(\omega)(\Phi(\omega)r^*(h_{opt}) - L(\omega))^2 \right] = E \left[ (f'(L(\omega)))^2 (\Phi(\omega)r^*(h_{opt}) - L(\omega))^2 \right],
\]

where (4.11) and (4.13) follow from (4.10), and (4.12) follows from (4.1).

Theorem 4.2 suggests a specific choice of weight function, in particular, that the weight \( h(\omega) \) of each scenario \( \omega \) should be proportional to \( (f'(L(\omega)))^2 \).

4.3 Practical Implementation

In this section and in the Section 4.4, we consider the risk measure

\[
\alpha = E \left[ f(L(\omega)) \right] = E \left[ (L(\omega) - c)^+ \right].
\]

(4.14)

Given a loss threshold \( c \in \mathbb{R} \), \( \alpha \) is the expected excess loss over the level \( c \). From (4.10), we have the weight function

\[
h_{opt}(\omega) = \left( f'(L(\omega)) \right)^2 = \mathbb{I}_{\{L(\omega) \geq c\}}.
\]

(4.15)

This weighting is intuitively reasonable, as scenarios that have losses larger than the threshold \( c \) are assigned with more weight. In practice, however, the portfolio loss \( L(\omega) \) is unobservable, so we propose a two-pass procedure that does not depend on knowledge of \( L(\omega) \). In particular, we will first approximate \( L(\omega) \) with an unweighted regression and obtain \( \hat{r} \); from \( \hat{r} \), we will construct an approximation to \( h_{opt} \) to be used in a weighted regression. In order to describe our method, define \( M(\omega) \) to be the model error given the approximation \( \Phi(\omega)r^* \),

\[
M(\omega) \triangleq L(\omega) - \Phi(\omega)r^*.
\]

(4.16)

Using standard results from the theory of regression, Section 3 establishes the following lemma.

Lemma 3.1. Suppose Assumptions 3.A1, 3.A2, and 3.A3 hold. As the number of scenarios \( n \to \infty \),

\[
\sqrt{n}(\hat{r} - r^*) \overset{d}{\to} N \left( 0, \Sigma_M + \frac{\Sigma_\nu}{m} \right),
\]
CHAPTER 4. THE WEIGHTED REGRESSION METHOD

where

\[ \Sigma_M \triangleq \mathbb{E} \left[ M^2(\omega) \Phi(\omega)\Phi(\omega)^\top \right], \]

and

\[ \Sigma_v \triangleq \mathbb{E} \left[ v(\omega) \Phi(\omega)\Phi(\omega)^\top \right]. \]

Therefore, as \( n \to \infty \),

\[ \| \hat{r} - r^* \|_2 = \mathcal{O}_p\left( \frac{1}{\sqrt{n}} \right). \]

Remark 4.2. One implication of Lemma 3.1 is that, given a fixed number of inner stage samples \( k = mn \), the asymptotic distribution of \( \hat{r} \) has minimum variance when \( m = 1 \) and \( n = k \). We will apply this fact in our analysis and numerical experiments.

When \( m = 1 \) and \( n = k \), as \( k \to \infty \), Lemma 3.1 and (4.16) imply that, for a fixed scenario \( \omega \),

\[ \sqrt{k} (\Phi(\omega) \hat{r} + M(\omega) - L(\omega)) \xrightarrow{d} N(0, \Phi(\omega) (\Sigma_M + \Sigma_v) \Phi(\omega)^\top). \]  

Equation (4.17) suggests that when \( k \) is large, it is reasonable to approximate the Bayesian posterior distribution of the portfolio loss \( L(\omega) \) given \( \bar{\omega} \) and \( \bar{\zeta} \) by a normal distribution with mean \( \Phi(\omega) \hat{r} + M(\omega) \) and variance \( \Phi(\omega) (\Sigma_M + \Sigma_v) \Phi(\omega)^\top / k \). Given this approximation, we can approximate \( h_{opt} \) with its posterior mean:

\[ \mathbb{E} \left[ h_{opt}(\omega) \bigg| \bar{\omega}, \bar{\zeta} \right] \approx N\left( \frac{\sqrt{k} (\Phi(\omega) \hat{r} + M(\omega) - c)}{\sqrt{\Phi(\omega) (\Sigma_M + \Sigma_v) \Phi(\omega)^\top}} \right). \]  

(Notice that \( \hat{r} \) depends on \( \bar{\omega} \) and \( \bar{\zeta} \).) If the basis functions are well-chosen, the model error \( M(\omega) \) is small in magnitude relative to the unweighted regression approximation \( \Phi(\omega) \hat{r} \). Further, in practice, we observe that the denominator of the argument in the right-hand side of (4.18) does not vary by more than one order of magnitude. These observations suggest an overall approximation for the globally optimal weight function \( h_{opt} \):

\[ h(\omega) = N\left( \frac{\sqrt{k} (\Phi(\omega) \hat{r} - c)}{\Gamma} \right), \]  

for some constant \( \Gamma > 0 \). Using the weight function (4.19), we have a two-pass weighted regression procedure:
• the first pass is an unweighted regression and provides $\Phi(\omega) \hat{r}$, an approximation of $L(\omega)$ used to determine weights;

• based on $\Phi(\omega) \hat{r}$, the second pass is a weighted regression with weights defined by (4.19).

Notice that this two-pass weighted regression procedure assigns weights with a weight function inspired by $h_{opt}$, and it does not depend on information of $L(\omega)$ or any quantities that are unknown in practice. In Section 4.4, we will compare the two-pass weighted regression method with the unweighted regression method.

4.4 Numerical Results

In this section, we use numerical examples to demonstrate the two-pass weighted regression method based on (4.19). The standard nested simulation method and the unweighted regression method are also implemented as competing methods. We follow the same experimental setting as in Section 3.4.1.

4.4.1 Examples

We consider two examples: a one-dimensional problem and a higher-dimensional problem.

4.4.1.1 One-Dimensional Example

• Model. There is one asset with initial price $S_0 = 100$. The drift under the real-world distribution is $\mu = 8\%$. The annual volatility is $\sigma = 20\%$. The continuously compounded riskless rate of interest is $r_f = 3\%$.

• Portfolio. The portfolio consists of a long position in a single put option with strike $K = 95$ and maturity $T = 0.25$ years. The risk horizon is $\tau = 1/52$ years. The threshold $c = 0.859$ is the 90th percentile of the portfolio loss distribution.

4.4.1.2 Multi-Dimensional Delta-Hedged Example

• Model. There are 10 i.i.d. assets each with initial price $S_0 = 100$. The drift under the real-world distribution is $\mu = 8\%$. The annual volatility is $\sigma = 30\%$. The continuously
compounded riskless rate of interest is $r_f = 5\%$.

- **Portfolio.** The portfolio consists of three types of securities. All derivatives in the portfolio have strike $K = 100$ and maturity $T = 0.1$ years. The risk horizon is $\tau = 0.04$ years. The three types are:
  - Short 10 down-and-out call options on asset $i$ with barrier $H = 95$, for $i = 1, \ldots, 10$.
  - Short 5 cash-or-nothing put options on asset $i$, for $i = 1, \ldots, 10$.
  - An amount of asset $i$ so that the portfolio delta with respect to asset $i$ is zero, for $i = 1, \ldots, 10$.

The threshold $c = 144.007$ is the 90th percentile of the portfolio loss distribution.

### 4.4.1.3 Estimators

We test the following estimators for both examples.

- **Optimal standard nested estimator.** This is the estimator $\hat{\alpha}_{SN(m,n)}$ with the parameters $m$ and $n$ chosen optimally according to (2.1). Due to the fact described in Remark 2.1, we test a large number of choices of the parameter pair $(m, n)$, and choose the one with the minimum MSE among them. Note that in practice, this optimal choice is not achievable.

- **Unweighted quadratic regression.** The estimator $\hat{\alpha}_{REG(m,n)}$ with these basis functions:
  - In the one-dimensional problem, the underlying asset has price $S_{\tau} (\omega)$ at time $\tau$. The basis function set $\Phi$ consists of $\phi_1 (\omega) = 1$, $\phi_2 (\omega) = S_{\tau} (\omega)$, and $\phi_3 (\omega) = (S_{\tau} (\omega))^2$.
  - In the multi-dimensional problem, the underlying assets have price $S_{j,\tau} (\omega)$ for $j = 1, \ldots, 10$ at time $\tau$. The basis function set $\Phi$ includes quadratic functions of each $S_{j,\tau} (\omega)$ and quadratic functions of each $(S_{j,\tau} (\omega) - H)_+$ for $j = 1, \ldots, 10$. Specifically, $\Phi$ consists of $\phi_\ell (\omega)$ for $\ell = 1, \ldots, 41$, where $\phi_1 (\omega) = 1$, $\phi_{1+i} (\omega) = S_{\ell,\tau} (\omega)$ and $\phi_{11+i} (\omega) = (S_{\ell,\tau} (\omega) - H)_+$ for $i = 1, \ldots, 10$, and $\phi_{21+i} (\omega) = (\phi_{1+i} (\omega))^2$ for $i = 1, \ldots, 20$. There are 41 basis functions.

- **Two-pass weighted quadratic regression.** This is the estimator $\hat{\alpha}_{REG(m,n,h)}$ with basis functions as in the unweighted quadratic regression method above and the weight function (4.19).
From Remark 4.2 we use $m = 1$ with the estimators of unweighted quadratic regression and two-pass weighted quadratic regression.

4.4.2 Numerical Performance

Results for the MSE of the various estimators are given in Figures 4.1 and 4.2. We interpret these results as follows:

- **Optimal standard nested estimator.** As the total number of inner stage samples $k$ increases, the MSE converges at the rate $k^{-2/3}$, consistent with theoretical results.

- **Unweighted quadratic regression.** The MSE of this estimator converges at the rate $k^{-1}$ until reaching an asymptotic bias level which depends on the size of the regression model error $M(\cdot)$, or, equivalently, the quality of the basis functions. This is consistent with the analysis in Section 3.
Figure 4.2: Illustration of the mean squared error in the 10-dimensional delta-hedged example. The vertical axis shows the mean squared error, and the horizontal axis represents the total number of inner stage samples.


**Figure 4.3:** Illustration of approximations in the one-dimensional example. The vertical axis shows the portfolio loss, either true or estimated, and the horizontal axis represents the underlying asset price $S_\tau(\omega)$ at time $\tau$.

- **Two-pass weighted quadratic regression.** When $k$ is small, the MSE of the two-pass weighted quadratic regression method is close to that of the unweighted quadratic regression method. This is due to the weight function (4.19). When $k$ is small, the weight function is relatively flat (i.e., nearly equal weights). When $k$ is large, the weight function (4.19) becomes steep, and more weight is assigned to more important scenarios, where losses exceed the threshold $c$. With large $k$, the MSE of the two-pass weighted quadratic regression method works about one order of magnitude better than the unweighted quadratic regression method.

Figure 4.3 shows the approximations via different regression methods for large values of $n$ in the one-dimensional example. Here, the unweighted quadratic regression method does not approximate the true portfolio loss in regions with large losses. On the other hand, the two-pass weighted quadratic regression method, with the same basis functions, fits the true portfolio loss much better in the large loss region.
4.5 Conclusion

In this chapter we propose a weighted regression method for nested estimation. This method improves upon unweighted regression by assigning more weight to scenarios that are more important to the calculation of the risk measure. This approach fits the true loss curve better in the region of the most important scenarios, with the weight of each scenario \( h(\omega) \) proportional to the quantity \( (f'(L(\omega)))^2 \).

Since \( L(\omega) \) is unobservable, we propose a practically implementable two-pass weighted regression method: the first pass is an unweighted regression method and provides an approximation \( \Phi(\omega)\hat{r} \) of \( L(\omega) \), and in the second pass, we have a weighted regression with the weight \( h(\omega) \) proportional to the quantity \( (f'(\Phi(\omega)\hat{r}))^2 \). Numerical results showed that this two-pass weighted regression method works well, which outperforms the unweighted regression method asymptotically, and outperforms standard nested simulation when \( k \) is not too large.
Bibliography


Appendix A

Proofs Related to the Sequential Method

In this section, we provide proofs of Theorems 2.2, 2.3, and 2.4 in Section 2.3.2 which analyze the performance of the Threshold estimator.

A.1 Preliminaries

Consider the Threshold estimator with \( n \) scenarios and a threshold parameter \( \gamma \). Each scenario \( \omega^{(i)} \) has inner loss samples \( \hat{Z}_{i,1}, \hat{Z}_{i,2}, \ldots \), which, by Assumption 2.2, are i.i.d. normal random variables with mean \( L(\omega^{(i)}) \) and standard deviation \( \sigma(\omega^{(i)}) \). The estimator will generate \( m_i \) inner stage samples in this scenario, with

\[
m_i = \inf \left\{ m > 0 : \left| S_m^{(i)} \right| \geq \gamma \right\} .
\]

Here, for \( m \geq 0 \), the partial sum \( S_m^{(i)} \) is defined by

\[
S_m^{(i)} \triangleq \sum_{j=1}^{m} \frac{1}{\sigma(\omega^{(i)})} \left( \hat{Z}_{i,j} - c \right).
\]

Each term in this partial sum has mean \( \mu(\omega^{(i)}) \triangleq (L(\omega^{(i)}) - c)/\sigma(\omega^{(i)}) \). By considering these partial sums over all \( n \) scenarios, the Threshold estimator can be written as

\[
\hat{\alpha}_{\text{SEQ}(m,n)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{S_m^{(i)} \geq \gamma\}.
\]
We are interested in the asymptotic behavior of the bias,
\[
\bar{b}(\gamma) \triangleq E[\hat{\alpha}_{\text{SEQ}(m,n)} - \alpha] = E \left[ \mathbb{I}_{\{S_m \geq \gamma\}} - \mathbb{I}_{\{L(\omega^{(i)}) \geq c\}} \right],
\]  
(A.2)
and the expected number of inner stage samples per scenario,
\[
\bar{m}(\gamma) \triangleq E[m_i],
\]  
(A.3)
as \gamma \to \infty.

Now, given \(\mu \in \mathbb{R}\), define \(P_\mu\) to be a probability measure so that, under \(P_\mu\), the random variables \(Y_1, Y_2, \ldots\) are a collection of i.i.d. normal random variables with mean \(\mu\) and unit variance. For each \(m \geq 0\), define the partial sum
\[
S_m \triangleq \sum_{j=1}^{m} Y_j.
\]
It follows from Assumption 2.2 that, if \(\mu = \mu(\omega^{(i)})\), then \(S_m\) has the same distribution as \(S_{m}^{(i)}\).

Define
\[
m(\gamma) \triangleq \inf \{m > 0 : |S_m| \geq \gamma\}.
\]
From (A.1) and (A.3), we have that \(\bar{m}(\gamma)\), the expected number of inner stage samples for the \text{Threshold} estimator, satisfies
\[
\bar{m}(\gamma) = \int E_\mu[m(\gamma)|p(\mu)]\, d\mu.
\]  
(A.4)
Here, \(E_\mu\) denotes expectation under the distribution \(\mu\). Similarly, we can define
\[
b_+(\gamma) \triangleq \mathbb{I}_{\{m(\gamma) < \infty \text{ and } S_m(\gamma) \geq \gamma\}}, \quad b_-(\gamma) \triangleq -\mathbb{I}_{\{m(\gamma) < \infty \text{ and } S_m(\gamma) \leq -\gamma\}},
\]
\[
b(\gamma) \triangleq b_-(\gamma)\mathbb{I}_{\{\mu \geq 0\}} + b_+(\gamma)\mathbb{I}_{\{\mu < 0\}}.
\]
Then, from (A.2), we have that \(\bar{b}(\gamma)\), the bias of the \text{Threshold} estimator, satisfies
\[
\bar{b}(\gamma) = \int E_\mu[b(\gamma)|p(\mu)]\, d\mu.
\]  
(A.5)
Finally, by Assumption 2.2 define \(\delta \in (0, 1)\) so that \(p\) is continuously differentiable over the interval \([-\delta, \delta]\), and set
\[
U_0 \triangleq \max_{|\mu| \leq \delta} |p(\mu)|, \quad U_1 \triangleq \max_{|\mu| \leq \delta} |p'(\mu)|.
\]  
(A.6)
APPENDIX A. PROOFS RELATED TO THE SEQUENTIAL METHOD

A.2 Asymptotic Bias

The asymptotic bias result of Theorem 2.2 is that, as $\gamma \to \infty$, $\bar{b}(\gamma) = O(\gamma^{-2})$. We will establish this via a careful analysis of (A.5). In particular, consider the decomposition

$$|\bar{b}(\gamma)| \leq \left| \int_{|\mu| > \delta} E_{\mu}[b(\gamma)p(\mu)] d\mu \right| + \left| \int_{|\mu| \leq \delta} E_{\mu}[b(\gamma)] p(\mu) d\mu \right|$$

(A.7)

$$\leq \left| \int_{|\mu| > \delta} E_{\mu}[b(\gamma)] p(\mu) d\mu \right| + \left| \int_{|\mu| \leq \delta} E_{\mu}[b(\gamma)] \mu p'(\varsigma(\mu)) d\mu \right| + \left| \int_{|\mu| \leq \delta} E_{\mu}[b(\gamma)] \mu p'(\varsigma(\mu)) d\mu \right|.$$

(A.8)

Here, using Assumption 2.2 and Taylor’s theorem, and $\varsigma$ is a function with $|\varsigma(\mu)| \leq \delta$ for all $\mu \in [-\delta, \delta]$. By symmetry, for any $\mu$, we have that

$$E_{\mu}[b(\gamma)] = -E_{-\mu}[b(\gamma)].$$

Then,

$$|\bar{b}(\gamma)| \leq \left| \int_{|\mu| > \delta} E_{\mu}[b(\gamma)] p(\mu) d\mu \right| + \left| \int_{|\mu| \leq \delta} E_{\mu}[b(\gamma)] \mu p'(\varsigma(\mu)) d\mu \right| + \left| \int_{|\mu| \leq \delta} |E_{\mu}[b(\gamma)]| \mu d\mu \right|.$$

(A.9)

Theorem 2.2 will follow by applying Lemmas A.1 and A.2 established below, to (A.8).

We begin with a preliminary proposition:

**Proposition A.1.** If $\mu < 0$, then

$$e^{2\mu \gamma} E_{-\mu} \left[ (1 + 2\mu(S_{m(\gamma)} - \gamma))^2 \mathbb{I}_{\{S_{m(\gamma)} \geq \gamma\}} \right] \leq P_{\mu}(S_{m(\gamma)} \geq \gamma) \leq e^{2\mu \gamma} P_{-\mu}(S_{m(\gamma)} \geq \gamma).$$

If $\mu > 0$, then

$$e^{-2\mu \gamma} E_{-\mu} \left[ (1 + 2\mu(S_{m(\gamma)} + \gamma))^2 \mathbb{I}_{\{S_{m(\gamma)} \leq -\gamma\}} \right] \leq P_{\mu}(S_{m(\gamma)} \leq -\gamma) \leq e^{-2\mu \gamma} P_{-\mu}(S_{m(\gamma)} \leq -\gamma).$$

**Proof.** Consider the case where $\mu < 0$. Let $F_{\mu}$ denote the $N(\mu, 1)$ distribution. Note that the Radon-Nikodym derivative between the $F_{\mu}$ and $F_{-\mu}$ is given by

$$\frac{dF_{\mu}(y)}{dF_{-\mu}(y)} = e^{2\mu y}.$$

Then,

$$P_{\mu}(S_{m(\gamma)} \geq \gamma) = E_{\mu} \left[ 1_{\{S_{m(\gamma)} \geq \gamma\}} \right] = E_{-\mu} \left[ e^{2\mu S_{m(\gamma)}} 1_{\{S_{m(\gamma)} \geq \gamma\}} \right] = e^{2\mu \gamma} E_{-\mu} \left[ e^{2\mu(S_{m(\gamma)} - \gamma)} 1_{\{S_{m(\gamma)} \geq \gamma\}} \right].$$

(A.9)
For \( x > 0 \), we have that \( 1 - x \leq e^{-x} \leq 1 \). Thus,

\[
(1 + 2\mu(S_m(\gamma) - \gamma))I\{S_m(\gamma) \geq \gamma\} \leq e^{2\mu(S_m(\gamma) - \gamma)}I\{S_m(\gamma) \geq \gamma\} \leq I\{S_m(\gamma) \geq \gamma\}.
\]

The result follows after taking an expectation with respect to \( P_{-\mu} \), and applying (A.9). The case where \( \mu > 0 \) is handled similarly.

The following lemma will bound the first term of (A.8).

**Lemma A.1.** As \( \gamma \to \infty \),

\[
\int_{|\mu| > \delta} E_{\mu}[b(\gamma)]p(\mu) \, d\mu = o(\gamma^{-2}).
\]

**Proof.** Note that

\[
\left| \int_{|\mu| > \delta} E_{\mu}[b(\gamma)]p(\mu) \, d\mu \right| = \left| \int_{\delta}^{\infty} E_{\mu}[b_-(\gamma)]p(\mu) \, d\mu + \int_{-\infty}^{-\delta} E_{\mu}[b_+(\gamma)]p(\mu) \, d\mu \right|
\leq \int_{\delta}^{\infty} \left| E_{\mu}[b_-()] \right|p(\mu) \, d\mu + \int_{-\infty}^{-\delta} \left| E_{\mu}[b_+]() \right|p(\mu) \, d\mu
\leq \int_{\delta}^{\infty} P_{\mu}(S_m(\gamma) \leq -\gamma)p(\mu) \, d\mu + \int_{-\infty}^{-\delta} P_{\mu}(S_m(\gamma) \geq \gamma)p(\mu) \, d\mu.
\]

By Proposition A.1

\[
\left| \int_{|\mu| > \delta} E_{\mu}[b(\gamma)]p(\mu) \, d\mu \right| \leq \int_{\delta}^{\infty} e^{-2\mu\gamma}P_{-\mu}(S_m(\gamma) \leq -\gamma)p(\mu) \, d\mu + \int_{-\infty}^{-\delta} e^{2\mu\gamma}P_{-\mu}(S_m(\gamma) \geq \gamma)p(\mu) \, d\mu
\leq \int_{\delta}^{\infty} e^{-2\mu\gamma}p(\mu) \, d\mu + \int_{-\infty}^{-\delta} e^{2\mu\gamma}p(\mu) \, d\mu
\leq e^{-2\delta\gamma} \int_{|\mu| > \delta} p(\mu) \, d\mu
\leq o(\gamma^{-2}).
\]

Now we bound the second term of (A.8).

**Lemma A.2.** As \( \gamma \to \infty \),

\[
\int_{|\mu| \leq \delta} |E_{\mu}[b(\gamma)]\mu| \, d\mu = O(\gamma^{-2}).
\]
Proof. Notice that, using Proposition [A.1],
\[
\int_{|\mu| \leq \delta} |E_{\mu}[b(\gamma)]\mu| \, d\mu \leq \int_{-\delta}^{0} |E_{\mu}[b_+(\gamma)]\mu| \, d\mu + \int_{0}^{\delta} |E_{\mu}[b_-(\gamma)]\mu| \, d\mu
\]
\[
\leq \int_{-\delta}^{0} e^{2\mu \gamma} P_{-\mu}(S_{m(\gamma)} \geq \gamma) \mu \, d\mu + \int_{0}^{\delta} e^{-2\mu \gamma} P_{-\mu}(S_{m(\gamma)} \leq -\gamma) \mu \, d\mu
\]
\[
\leq \int_{-\delta}^{0} e^{2\mu \gamma} |\mu\mu| \, d\mu + \int_{0}^{\delta} e^{-2\mu \gamma} |\mu\mu| \, d\mu
\]
\[
= 2 \left( -\frac{e^{-2\mu \gamma}}{4\gamma^2} - \frac{\mu e^{-2\mu \gamma}}{2\gamma} \right)|_{\mu=0}^\delta
\]
\[
= \frac{1}{2\gamma} - \frac{e^{-2\delta \gamma}}{2\gamma^2} - \frac{\delta e^{-2\delta \gamma}}{\gamma}.
\]
The result follows. ■

A.3 Expected Number of Inner Samples

The asymptotic characterization of the number of inner samples provided by Theorem 2.3 is that, as \( \gamma \to \infty \), \( \bar{m}(\gamma) = O(\gamma \log \gamma) \). We will establish this via an analysis of (A.4). In particular, we have that
\[
\bar{m}(\gamma) = \int_{|\mu| > \gamma} E_{\mu}[m(\gamma)]p(\mu) \, d\mu + \int_{|\mu| \leq \gamma} E_{\mu}[m(\gamma)]p(\mu) \, d\mu.
\] (A.10)

Theorem 2.3 will follow by applying Lemmas A.4 and A.5, established below, to (A.10).

To this end, the following result will be helpful.

Lemma A.3. Suppose \( Y_1, Y_2, \ldots \) are i.i.d. random variables under the probability measure \( P_{\mu} \), with \( E_{\mu}[Y_1] = \mu \) and \( E_{\mu}[Y_1^2] < \infty \). Define, for \( m \geq 0 \), the partial sum
\[
S_m \triangleq \sum_{j=1}^{m} Y_j,
\]
and, for \( \gamma > 0 \), the hitting times
\[
m_+(\gamma) \triangleq \inf \{ m > 0 : S_m > \gamma \}, \quad m_-(\gamma) \triangleq \inf \{ m > 0 : |S_m| > \gamma \}.
\]
(i) [Lorden (1970)] Suppose that \( \mu > 0 \). Then,
\[
\sup_{\gamma > 0} E_{\mu}[S_{m_+(\gamma)} - \gamma] \leq \frac{E_{\mu}[(Y_1^+)^2]}{\mu},
\]
where \( x^+ \triangleq \max(x,0) \) denotes the positive part of \( x \).
(ii) (Pruitt (1981)) There exist constants $V_1$ and $V_2$ (independent of the distribution of $Y_1$) such that
\[
E_{\mu}[m_{\pm}(\gamma)] \leq \frac{V_1}{K_{\mu}(\gamma)}, \quad P_{\mu}\left(\max_{1 \leq m \leq n} |S_m| \leq \gamma \right) \leq \frac{V_2}{(nK_{\mu}(\gamma))^3},
\]
where
\[
K_{\mu}(\gamma) \triangleq \gamma^{-2}E_{\mu}[|Y_1|^2I_{\{|Y_1| \leq \gamma\}}].
\]

(iii) (Gut (1974))
\[
E\left[(Y_{m_{\pm}(\gamma)}^+)\right]^2 \leq E[m_{\pm}(\gamma)]E[(Y_1^+)^2].
\]

The following lemma will bound the first term of (A.10).

Lemma A.4. As $\gamma \to \infty$,
\[
\int_{|\mu| > \gamma^{-1}} E_{\mu}[m(\gamma)]p(\mu) \, d\mu = O(\gamma \log \gamma).
\]

Proof. Note that, since $Y_1$ has mean $\mu$ and unit variance under the distribution $P_{\mu}$,
\[
\frac{E_{\mu}[Y_1^+]}{\mu} \leq \frac{E_{\mu}[|Y_1|^2]}{\mu} = \frac{1 + \mu^2}{\mu}. \tag{A.11}
\]
Further, define the one-sided hitting times $m_{\pm}(\gamma)$ as in Lemma A.3. By the optional stopping theorem, if $\mu > 0$, then
\[
E_{\mu}[S_{m_{+}(\gamma)}] = \mu E_{\mu}[m_{+}(\gamma)],
\]
and if $\mu < 0$,
\[
E_{\mu}[S_{m_{-}(\gamma)}] = \mu E_{\mu}[m_{-}(\gamma)].
\]
Then, since $m(\gamma) \leq m_{+}(\gamma)$ and $m(\gamma) \leq m_{-}(\gamma)$, we have that
\[
\int_{|\mu| > \gamma^{-1}} E_{\mu}[m(\gamma)]p(\mu) \, d\mu \leq \int_{-\infty}^{-\gamma^{-1}} E_{\mu}[m_{-}(\gamma)]p(\mu) \, d\mu + \int_{\gamma^{-1}}^{\infty} E_{\mu}[m_{+}(\gamma)]p(\mu) \, d\mu
\]
\[
= \int_{-\infty}^{-\gamma^{-1}} \frac{E_{\mu}[S_{m_{-}(\gamma)}]}{\mu} p(\mu) \, d\mu + \int_{\gamma^{-1}}^{\infty} \frac{E_{\mu}[S_{m_{+}(\gamma)}]}{\mu} p(\mu) \, d\mu
\]
\[
\leq \int_{-\infty}^{-\gamma^{-1}} \left(\frac{1 + \mu^2}{\mu^2} - \frac{\gamma}{\mu}\right) p(\mu) \, d\mu + \int_{\gamma^{-1}}^{\infty} \left(\frac{1 + \mu^2}{\mu^2} + \frac{\gamma}{\mu}\right) p(\mu) \, d\mu
\]
\[
= \int_{|\mu| > \gamma^{-1}} \left(1 + \frac{1}{\mu^2} + \frac{\gamma}{|\mu|}\right) p(\mu) \, d\mu. \tag{A.12}
\]
Here, the final inequality follows from (A.11) and Part [iii] of Lemma A.3.

Now, without loss of generality, assume that \( \gamma > \delta^{-1} \). Recalling \( U_0 \) from (A.6), we have that

\[
\int_{|\mu| > \gamma^{-1}} E_\mu[m(\gamma)] p(\mu) \, d\mu \leq \int_{\delta \leq |\mu| > \gamma^{-1}} \left( 1 + \frac{1}{\mu^2} + \frac{\gamma}{|\mu|} \right) p(\mu) \, d\mu + \int_{|\mu| > \delta} \left( 1 + \frac{1}{\mu^2} + \frac{\gamma}{|\mu|} \right) p(\mu) \, d\mu \leq 2U_0 \int_{\gamma^{-1}}^{\delta} \left( 1 + \frac{1}{\mu^2} + \frac{\gamma}{\mu} \right) \, d\mu + \left( 1 + \frac{1}{\delta^2} + \frac{\gamma}{\delta} \right) \int_{|\mu| > \delta} p(\mu) \, d\mu \leq 2U_0 \left( \delta - \gamma^{-1} + \gamma - \delta^{-1} + \gamma \log \delta + \gamma \log \gamma \right) + 1 + \delta^{-2} + \gamma \delta^{-1} = O(\gamma \log \gamma).
\]

The result follows.

Now we bound the second term of (A.10).

**Lemma A.5.** As \( \gamma \to \infty \),

\[
\int_{|\mu| \leq \gamma^{-1}} E_\mu[m(\gamma)] p(\mu) \, d\mu = O(\gamma).
\]

**Proof.** Here, we will apply Part [ii] of Lemma A.3. Without loss of generality, assume that \( \gamma > 1 \). Then, \( |\mu| < 1 \) in the region of integration and thus \( \gamma - |\mu| > 0 \). The variable \( K_\mu(\gamma) \) from Lemma A.3 satisfies

\[
K_\mu(\gamma) = \gamma^{-2} E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq \gamma} \right] = \gamma^{-2} E_\mu \left[ |Y_1 + \mu|^{2} 1_{|Y_1| \leq |\mu|} \right] \geq \gamma^{-2} E_\mu \left[ |Y_1 + \mu|^{2} 1_{|Y_1| \leq |\mu|} \right] \geq \gamma^{-2} \left( E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right] \right) + 2\mu E_\mu \left[ Y_1 1_{|Y_1| \leq |\mu|} \right] = \gamma^{-2} E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right].
\]

Here, we have used the fact that under \( P_0 \), \( Y_1 \sim N(0, 1) \), hence \( E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right] = 0 \). Then, from Part [ii] of Lemma A.3, since \( m(\gamma) \leq m_{\pm}(\gamma) \),

\[
E_\mu[m(\gamma)] \leq E_\mu[m_{\pm}(\gamma)] \leq \frac{V_1 \gamma^2}{E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right]}.
\]

Without loss of generality, assume that \( \gamma > \delta^{-1} \), and recall \( U_0 \) from (A.6). Then,

\[
\int_{|\mu| \leq \gamma^{-1}} E_\mu[m(\gamma)] p(\mu) \, d\mu \leq \int_{|\mu| \leq \gamma^{-1}} \frac{V_1 \gamma^2}{E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right]} p(\mu) \, d\mu \leq \frac{V_1 \gamma^2}{E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right]} \int_{|\mu| \leq \gamma^{-1}} p(\mu) \, d\mu \leq \frac{2U_0 V_1 \gamma}{E_\mu \left[ |Y_1|^2 1_{|Y_1| \leq |\mu|} \right]}.
\]
Notice that $\gamma > 1 > \delta$ is assumed before. By the monotone convergence theorem,

$$
\lim_{\gamma \to \infty} E_0 \left[ |Y_1|^2 \mathbb{I}_{|Y_1| \leq \gamma - \delta} \right] = E_0 \left[ |Y_1|^2 \right] = 1.
$$

(A.16)

The result follows. ■

### A.4 Realized Number of Inner Samples

In this section, we will establish Theorem 2.4, which provides a probabilistic bound on the realized number of inner stage samples per scenario. Our proof relies on the following lemma, which bounds the second moment of the number of inner stage samples per scenario.

**Lemma A.6.** As $\gamma \to \infty$,

$$
E [m(\gamma)^2] = O(\gamma^3).
$$

We will defer the proof of Lemma A.6 for the moment, and first employ this lemma to prove Theorem 2.4.

**Proof of Theorem 2.4**. Fix $\epsilon > 0$ and suppose that $\gamma \geq \gamma_0$. Then, by Chebyshev’s inequality,

$$
P \left( \frac{1}{n} \sum_{i=1}^{n} m_i \geq \left(C_0 + \epsilon\right) \gamma \log \gamma \right) \leq P \left( \left| \frac{1}{n} \sum_{i=1}^{n} (m_i - \bar{m}(\gamma)) \right| \geq \epsilon \gamma \log \gamma \right)
$$

$$
\leq \frac{\text{Var}(m(\gamma))}{n (\epsilon \gamma \log \gamma)^2} \leq \frac{E \left[ m(\gamma)^2 \right]}{n (\epsilon \gamma \log \gamma)^2}.
$$

By Lemma A.6, there exist constants $C_0', \gamma_0', \gamma_1 > 0$ so that if $\gamma \geq \gamma_0'$, $E \left[ m(\gamma)^2 \right] \leq C_0' \gamma^3$. Then, if $\gamma \geq \max\{\gamma_0, \gamma_0', \gamma_1\}$, we have that

$$
P \left( \frac{1}{n} \sum_{i=1}^{n} m_i \geq \left(C_0 + \epsilon\right) \gamma \log \gamma \right) \leq \frac{C_0'}{C_1 (\epsilon \log \gamma)^2},
$$

which can be made arbitrarily small with sufficiently large $\gamma$. ■

To prove Lemma A.6, consider the decomposition

$$
E \left[ m(\gamma)^2 \right] = \int_{|\mu| > \gamma^{-1}} E_{\mu} \left[ m(\gamma)^2 \right] p(\mu) \, d\mu + \int_{|\mu| \leq \gamma^{-1}} E_{\mu} \left[ m(\gamma)^2 \right] p(\mu) \, d\mu.
$$

(A.17)

Lemma A.6 will follow by applying Lemmas A.7 and A.8 established below, to (A.17).
Lemma A.7. As $\gamma \to \infty$,
\[
\int_{|\mu| > \gamma^{-1}} E_\mu [m(\gamma)^2] \, p(\mu) \, d\mu = O(\gamma^3).
\]

Proof. We proceed as in the proof of Lemma A.4. Using the stopping times $m_+(\gamma)$ and $m_-(\gamma)$ defined there, we have
\[
\int_{|\mu| > \gamma^{-1}} E_\mu [m(\gamma)^2] \, p(\mu) \, d\mu \leq \int_{-\infty}^{-\gamma^{-1}} E_\mu [m_-(\gamma)^2] \, p(\mu) \, d\mu + \int_{\gamma^{-1}}^{\infty} E_\mu [m_+(\gamma)^2] \, p(\mu) \, d\mu.
\]

First, consider the case when $\mu > 0$. By the optional stopping theorem applied to the quadratic martingale $(S_m - \mu m)^2 - m$, we have that
\[
E_\mu \left[ (S_{m_+(\gamma)} - \mu m_+(\gamma))^2 \right] = E_\mu [m_+(\gamma)].
\]

Now, for any real numbers $a, b \in \mathbb{R}$, we have that $(a + b)^2 \leq 2(a^2 + b^2)$. Therefore,
\[
E_\mu [m_+(\gamma)^2] \leq \frac{2}{\mu^2} \left( E_\mu [m_+(\gamma)] + E_\mu \left[ (\mu m_+(\gamma))^2 \right] \right)
\]
\[
= \frac{2}{\mu^2} \left( E_\mu [m_+(\gamma)] + E_\mu \left[ S_{m_+(\gamma)}^2 \right] \right).
\]

Using the fact that $S_{m_+(\gamma)} \leq \gamma + Y_{m_+(\gamma)}^+$, Part (iii) of Lemma A.3, and (A.11),
\[
E_\mu [m_+(\gamma)^2] \leq \frac{2}{\mu^2} \left( E_\mu [m_+(\gamma)] + E_\mu \left[ (\gamma + Y_{m_+(\gamma)}^+)^2 \right] \right)
\]
\[
\leq \frac{2}{\mu^2} \left( E_\mu [m_+(\gamma)] + 2\gamma^2 + 2E_\mu \left[ (Y_{m_+(\gamma)}^+)^2 \right] \right)
\]
\[
\leq \frac{2}{\mu^2} \left( E_\mu [m_+(\gamma)] + 2\gamma^2 + 2E_\mu [m_+(\gamma)] E_\mu \left[ (Y_{m_+(\gamma)}^+)^2 \right] \right)
\]
\[
\leq \frac{2}{\mu^2} \left( E_\mu [m_+(\gamma)] + 2\gamma^2 + 2(\mu^2 + 1)E_\mu [m_+(\gamma)] \right)
\]
\[
= \left( \frac{6}{\mu^2} + 4 \right) E_\mu [m_+(\gamma)] + \frac{4\gamma^2}{\mu^2}.
\]

By similar consideration of the symmetric case, we have that when $\mu < 0$,
\[
E_\mu [m_-(\gamma)^2] \leq \left( \frac{6}{\mu^2} + 4 \right) E_\mu [m_-(\gamma)] + \frac{4\gamma^2}{\mu^2}.
\]

Therefore, repeating the calculation in (A.12),
\[
\int_{|\mu| > \gamma^{-1}} E_\mu [m(\gamma)^2] \, p(\mu) \, d\mu \leq \int_{|\mu| > \gamma^{-1}} \left[ \left( \frac{6}{\mu^2} + 4 \right) \left( 1 + \frac{1}{\mu^2} + \frac{\gamma}{|\mu|} \right) + \frac{4\gamma^2}{\mu^2} \right] \, p(\mu) \, d\mu.
\]
Without loss of generality, assume that $\gamma > \delta^{-1}$. Then, as in (A.13),

$$
\int_{|\mu| > \gamma^{-1}} E_{\mu} [m(\gamma)^2] p(\mu) \, d\mu \leq 2U_0 \int_{\gamma^{-1}}^{\delta} \left[ \left( \frac{6}{\delta^2} + 4 \right) \left( 1 + \frac{1}{\mu^2} + \frac{\gamma}{\mu} \right) + \frac{4\gamma^2}{\mu^2} \right] \, d\mu \\
+ \left( \frac{6}{\delta^2} + 4 \right) \left( 1 + \frac{1}{\delta^2} + \frac{\gamma}{\delta} \right) + \frac{4\gamma^2}{\delta^2} \\
= O(\gamma^3).
$$

Lemma A.8. As $\gamma \to \infty$,

$$
\int_{|\mu| \leq \gamma^{-1}} E_{\mu} [m(\gamma)^2] p(\mu) \, d\mu = O(\gamma^3).
$$

Proof. We proceed as in Lemma A.5. Without loss of generality, assume that $\gamma > 1$. Observe that $m(\gamma) \leq m_{\pm}(\gamma)$, since the latter is an exit time for a larger set than the former. Then, using summation by parts,

$$
E_{\mu} [m(\gamma)^2] \leq E_{\mu} [m_{\pm}(\gamma)^2] = \sum_{n=1}^{\infty} n^2 P_{\mu}(m_{\pm}(\gamma) = n) \\
= 1 + \sum_{n=1}^{\infty} (2n + 1) P_{\mu}(m_{\pm}(\gamma) > n) \\
= 1 + \sum_{n=1}^{\infty} (2n + 1) P_{\mu}\left( \max_{1 \leq m \leq n} |S_m| \leq \gamma \right).
$$

Using Part (ii) of Lemma A.3 for any integer $A \geq 1$,

$$
E_{\mu} [m(\gamma)^2] \leq \sum_{n=0}^{A-1} (2n + 1) + \frac{V_2}{K_{\mu}(\gamma)^3} \sum_{n=A}^{\infty} \frac{2n + 1}{n^3} \\
\leq A^2 + \frac{3V_2}{K_{\mu}(\gamma)^3} \sum_{n=A}^{\infty} \frac{1}{n^2} \leq A^2 + \frac{3V_2}{K_{\mu}(\gamma)^3(A-1)}.
$$

Since $K_{\mu}(\gamma) \leq 1$, we may take

$$
A \triangleq \left\lfloor \frac{3}{K_{\mu}(\gamma)} \right\rfloor,
$$

so that $A - 1 \geq 1/K_{\mu}(\gamma)$. Then, there exists a constant $W_0$ so that

$$
E_{\mu} [m(\gamma)^2] \leq \frac{W_0}{K_{\mu}(\gamma)^2} \leq \frac{W_0 \gamma^4}{\left( E_{0} \left[ |Y_1|^2 \mathbb{I}_{(|Y_1| \leq \gamma - |\mu|)} \right] \right)^2},
$$

using (A.14).
Without loss of generality, assume that $\gamma > \delta^{-1}$. Then, as in (A.15),

\[
\int_{|\mu| \leq \gamma^{-1}} E_{\mu} [m(\gamma)^2] p(\mu) \, d\mu \leq \int_{|\mu| \leq \gamma^{-1}} \frac{W_0 \gamma^4}{(E_0 [Y_1^2 I_{\{|Y_1| \leq \gamma - |\mu|\}}])^2} p(\mu) \, d\mu \\
\leq \frac{W_0 \gamma^4}{(E_0 [Y_1^2 I_{\{|Y_1| \leq \gamma - \delta\}}])^2} \int_{|\mu| \leq \gamma^{-1}} p(\mu) \, d\mu \\
\leq \frac{2U_0 W_0 \gamma^3}{(E_0 [Y_1^2 I_{\{|Y_1| \leq \gamma - \delta\}}])^2}.
\]

The result follows from (A.16).
Appendix B

Proofs Related to the Unweighted Regression Method

This section presents the proofs of the results in Section 3.3, where the asymptotic analysis of the regression estimator has been established. First, we show the following lemma, which establishes the existence and uniqueness of the optimal solutions to (3.6) and (3.9).

Lemma B.1. Given Assumptions 3.A1 and 3.A2, the function $g(r)$ is strictly convex over $\mathbb{R}^d$, and the function $(1/n) \sum_{i=1}^{n} G(\cdot, \omega^{(i)}, \zeta^{(i)})$ is strictly convex over $\mathbb{R}^d$ almost surely.

Proof. As a function of $r$, the Hessian matrix of $(1/n) \sum_{i=1}^{n} G(\cdot, \omega^{(i)}, \zeta^{(i)})$ is

$$
\nabla^2 \left( \frac{1}{n} \sum_{i=1}^{n} G \left( r, \omega^{(i)}, \zeta^{(i)} \right) \right) = \nabla^2 \left( \frac{1}{n} \sum_{i=1}^{n} \left( \hat{L}(\omega^{(i)}, \zeta^{(i)}) - \Phi(\omega^{(i)}) r \right)^2 \right) 
$$

$$
= \frac{2}{n} \sum_{i=1}^{n} \Phi(\omega^{(i)})^\top \Phi(\omega^{(i)}),
$$

which is positive semidefinite almost surely under Assumption 3.A2. Therefore, $(1/n) \sum_{i=1}^{n} G(\cdot, \omega^{(i)}, \zeta^{(i)})$ is strictly convex almost surely.

The Hessian matrix of $g(r, \omega)$ is

$$
\nabla^2 (g(r, \omega)) = \nabla^2 \left( \mathbb{E} \left[ \left( \hat{L}(\omega, \zeta) - \Phi(\omega) r \right)^2 \right] \right) = \nabla^2 \left( r \mathbb{E} \left[ \Phi(\omega)^\top \Phi(\omega) \right] r \right) = 2I_d.
$$

Therefore, $g(r)$ is strictly convex.

\[ \blacksquare \]
APPENDIX B. PROOFS RELATED TO THE UNWEIGHTED REGRESSION METHOD

Given Assumptions 3.A1 and 3.A2 according to Lemma B.1, the optimal solutions to (3.6) and (3.9) exist and are unique almost surely.

We use the following notation: \( \bar{\omega} \) denotes the \( n \) outer stage scenarios,
\[
\bar{\omega} \triangleq (\omega^{(1)}, \ldots, \omega^{(n)})^\top,
\]
\( \tilde{\zeta} \) denotes the inner stage uncertainty,
\[
\tilde{\zeta} \triangleq (\zeta^{(1)}, \ldots, \zeta^{(n)})^\top,
\]
\( \Phi (\bar{\omega}) \) is an \( n \)-by-\( d \) matrix,
\[
\Phi (\bar{\omega}) \triangleq \begin{pmatrix}
\phi_1(\omega^{(1)}) & \cdots & \phi_d(\omega^{(1)}) \\
\vdots & \ddots & \vdots \\
\phi_1(\omega^{(n)}) & \cdots & \phi_d(\omega^{(n)})
\end{pmatrix},
\]
and \( \hat{L}(\bar{\omega}, \tilde{\zeta}) \) is an \( n \)-by-1 column vector,
\[
\hat{L}(\bar{\omega}, \tilde{\zeta}) \triangleq (\hat{L}(\omega^{(1)}, \zeta^{(1)}), \ldots, \hat{L}(\omega^{(n)}, \zeta^{(n)}))^\top.
\]

From regression theory, the unique optimal solution to (3.9) takes the form
\[
\hat{r} = (\Phi (\bar{\omega})^\top \Phi (\bar{\omega}))^{-1} \Phi (\bar{\omega})^\top \hat{L}(\bar{\omega}, \tilde{\zeta}).
\] (B.1)

Then we estimate the risk measure by
\[
\hat{\alpha}_{\text{REG}}(m, n) \triangleq \mathbb{E} \left[ f(\Phi (\omega) \hat{r}) \mid \bar{\omega}, \tilde{\zeta} \right].
\]

From Assumption 3.A1 the disturbance term \( \varepsilon(\omega, \zeta) \) of (3.8) satisfies
\[
\mathbb{E} [\varepsilon(\omega, \zeta) \mid \omega] = 0,
\] (B.2)
and
\[
\text{Var} (\varepsilon(\omega, \zeta) \mid \omega) = \frac{v(\omega)}{m}.
\] (B.3)

We define
\[
\varepsilon(\bar{\omega}, \tilde{\zeta}) \triangleq (\varepsilon(\omega^{(1)}, \zeta^{(1)}), \ldots, \varepsilon(\omega^{(n)}, \zeta^{(n)}))^\top.
\]
From the definition of the model error $M(\cdot)$ by (3.7) and the projection theorem, the basis functions $\phi_1(\cdot), \ldots, \phi_d(\cdot)$ are orthogonal to $M(\cdot)$, i.e.,

$$
E [\phi_\ell (\omega) M (\omega)] = 0,
$$

for $\ell = 1, \ldots, d$. We define

$$
M (\vec{\omega}) \triangleq \left( M (\omega^{(1)}), \ldots, M (\omega^{(n)}) \right)^\top.
$$

### B.1 Differentiable Case

**Lemma 3.1.** Suppose Assumptions 3.A1, 3.A2, and 3.A3 hold. As the number of scenarios $n \to \infty$,

$$
\sqrt{n} (\hat{r} - r^*) \xrightarrow{d} N \left( \mathbf{0}, \Sigma_M + \frac{\Sigma_v}{m} \right),
$$

where

$$
\Sigma_M \triangleq E \left[ M^2 (\omega) \Phi (\omega)^\top \Phi (\omega) \right],
$$

and

$$
\Sigma_v \triangleq E \left[ v (\omega) \Phi (\omega)^\top \Phi (\omega) \right].
$$

Therefore, as $n \to \infty$,

$$
\|\hat{r} - r^*\|_2 = O_p (1).
$$

**Proof.** From Theorem 5.3 in White (2001), we have

$$
\left( \text{Cov} \left( n^{-\frac{1}{2}} \Phi^\top (\vec{\omega}) (M(\vec{\omega}) + \varepsilon(\vec{\omega}, \vec{\zeta})) \right) \right)^{-\frac{1}{2}} \sqrt{n} (\hat{r} - r^*) \xrightarrow{d} N (\mathbf{0}, I_d).
$$

Note that using (B.2) and (B.4),

$$
\text{Cov} \left( n^{-\frac{1}{2}} \Phi^\top (\vec{\omega}) (M(\vec{\omega}) + \varepsilon(\vec{\omega}, \vec{\zeta})) \right)
= \frac{1}{n} E \left[ \Phi^\top (\vec{\omega}) (M(\vec{\omega}) + \varepsilon(\vec{\omega}, \vec{\zeta})) \right] \left( \Phi^\top (\vec{\omega}) (M(\vec{\omega}) + \varepsilon(\vec{\omega}, \vec{\zeta})) \right)^\top
= \frac{1}{n} E \left[ \Phi^\top (\vec{\omega}) M (\vec{\omega}) M^\top (\vec{\omega}) \Phi (\vec{\omega}) \right] + \frac{1}{n} E \left[ \Phi^\top (\vec{\omega}) E \left[ \varepsilon(\vec{\omega}, \vec{\zeta}) \varepsilon^\top (\vec{\omega}, \vec{\zeta}) \big| \vec{\omega}, \vec{\zeta} \right] \Phi (\vec{\omega}) \right]
+ \frac{2}{n} E \left[ \Phi^\top (\vec{\omega}) M (\vec{\omega}) E \left[ \varepsilon^\top (\vec{\omega}, \vec{\zeta}) \big| \vec{\omega}, \vec{\zeta} \right] \Phi (\vec{\omega}) \right].
$$
From (B.3),
\[
E \left[ \epsilon(\bar{\omega}, \bar{\zeta}) \epsilon^\top(\bar{\omega}, \bar{\zeta}) \right] = \begin{pmatrix}
\frac{v(\omega^{(1)})}{m} & 0 & \cdots & 0 \\
0 & \frac{v(\omega^{(2)})}{m} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & \frac{v(\omega^{(n)})}{m} \\
\end{pmatrix}.
\]

Further,
\[
\frac{1}{n} E \left[ \Phi^\top(\bar{\omega}) M(\bar{\omega}) \right] = \frac{1}{n} \begin{pmatrix}
E \left[ \sum_{i=1}^{n} \phi_1^2(\omega^{(i)}) M^2(\omega^{(i)}) \right] & \cdots & E \left[ \sum_{i=1}^{n} \phi_1(\omega^{(i)}) \phi_d(\omega^{(i)}) M^2(\omega^{(i)}) \right] \\
\vdots & \ddots & \vdots \\
E \left[ \sum_{i=1}^{n} \phi_d(\omega^{(i)}) \phi_1(\omega^{(i)}) M^2(\omega^{(i)}) \right] & \cdots & E \left[ \sum_{i=1}^{n} \phi_d^2(\omega^{(i)}) M^2(\omega^{(i)}) \right] \\
\end{pmatrix}
= \begin{pmatrix}
E \left[ \phi_1^2(\omega) M^2(\omega) \right] & \cdots & E \left[ \phi_1(\omega) \phi_d(\omega) M^2(\omega) \right] \\
\vdots & \ddots & \vdots \\
E \left[ \phi_d(\omega) \phi_1(\omega) M^2(\omega) \right] & \cdots & E \left[ \phi_d^2(\omega) M^2(\omega) \right] \\
\end{pmatrix} = E \left[ M^2(\omega) \Phi(\omega)^\top \Phi(\omega) \right].
\]

Therefore,
\[
\text{Cov} \left( \frac{1}{n} \Phi^\top(\bar{\omega})(M(\bar{\omega}) + \epsilon(\bar{\omega}, \bar{\zeta})) \right) = \frac{1}{n} E \left[ M^2(\omega) \Phi(\omega)^\top \Phi(\omega) \right] + \frac{1}{m} E \left[ v(\omega) \Phi(\omega)^\top \Phi(\omega) \right].
\]

From (B.5) and (B.6), we have that
\[
\left( \Sigma_M + \frac{1}{m} \Sigma_v \right)^{-\frac{1}{2}} \sqrt{n} (\hat{r} - r^*) \xrightarrow{d} N(0, I_d),
\]
where \(\Sigma_M\) and \(\Sigma_v\) are defined by (3.13) and (3.14). Based on the result above, we have that
\[
n \left\| \hat{r} - r^* \right\|_2^2 = \left( \sqrt{n} (\hat{r} - r^*) \right)^\top \left( \sqrt{n} (\hat{r} - r^*) \right) = O_P(1),\]
as \(n \to \infty\), and the result follows.

\[\blacksquare\]

**Theorem 3.1.** Suppose that Assumptions 3.F1, 3.A1, 3.A2, and 3.A3 hold. Then there exists a sequence of random variables \(\{B_{M,n}\}\), for \(n = 1, 2, \ldots\), satisfying
\[
B_{M,n} \xrightarrow{P} B_M^* \triangleq E \left[ f(\Phi(\omega) r^*) \right] - \alpha,
\]
so that
\[
\sqrt{n} (\hat{\alpha}_{\text{REG}(m,n)} - \alpha - B_{M,n}) \xrightarrow{d} \mathcal{N} \left( 0, \mathbb{E} \left[ f' (L(\omega)) \Phi (\omega) \right] \left( \Sigma_M + \frac{\Sigma_v}{m} \right) \left( \mathbb{E} \left[ f' (L(\omega)) \Phi (\omega) \right] \right)^\top \right),
\]
where \(\Sigma_M\) and \(\Sigma_v\) are defined by (3.13) and (3.14). Further, the asymptotic bias \(B_M^*\) satisfies
\[
|B_M^* - \mathbb{E} \left[ f' (L(\omega)) M(\omega) \right]| \leq \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (M(\omega))^2 \right].
\]

**Proof.** By Taylor’s theorem,
\[
f (\Phi (\omega) \hat{r}) - f (L(\omega)) = f' (L(\omega)) (\Phi (\omega) \hat{r} - L(\omega)) + \frac{1}{2} f'' (L(\omega) + \theta \cdot (\Phi (\omega) \hat{r} - L(\omega))) (\Phi (\omega) \hat{r} - L(\omega))^2,
\]
where \(\theta \in (0,1)\) is a random variable. Then
\[
\hat{\alpha}_{\text{REG}(m,n)} - \alpha = \mathbb{E} \left[ f (\Phi (\omega) \hat{r}) | \bar{\omega}, \bar{\zeta} \right] - \mathbb{E} [f (L(\omega))]
\]
\[
= \mathbb{E} \left[ f' (L(\omega)) (\Phi (\omega) \hat{r} - L(\omega)) | \bar{\omega}, \bar{\zeta} \right] + \mathbb{E} \left[ \frac{1}{2} f'' (L(\omega) + \theta \cdot (\Phi (\omega) \hat{r} - L(\omega))) (\Phi (\omega) \hat{r} - L(\omega))^2 | \bar{\omega}, \bar{\zeta} \right].
\]

For the first term in (B.7),
\[
\mathbb{E} \left[ f' (L(\omega)) (\Phi (\omega) \hat{r} - L(\omega)) | \bar{\omega}, \bar{\zeta} \right] = \mathbb{E} [f' (L(\omega)) \Phi (\omega)] (\hat{r} - r^*) - \mathbb{E} [f' (L(\omega)) M(\omega)].
\]

From Lemma 3.1, we have that
\[
\sqrt{n} (\mathbb{E} \left[ f' (L(\omega)) \Phi (\omega) \right]) (\hat{r} - r^*) \xrightarrow{d} \mathcal{N} \left( 0, \mathbb{E} \left[ f' (L(\omega)) \Phi (\omega) \right] \left( \Sigma_M + \frac{\Sigma_v}{m} \right) \left( \mathbb{E} \left[ f' (L(\omega)) \Phi (\omega) \right] \right)^\top \right),
\]
as \(n \to \infty\).

Combining (B.7) and (B.8), and letting
\[
B_{M,n} \triangleq \mathbb{E} \left[ f' (L(\omega)) M(\omega) \right] - \mathbb{E} \left[ \frac{1}{2} f'' (L(\omega) + \theta \cdot (\Phi (\omega) \hat{r} - L(\omega))) (\Phi (\omega) \hat{r} - L(\omega))^2 | \bar{\omega}, \bar{\zeta} \right],
\]
equation (3.15) follows. From Lemma 3.1
\[
\hat{r} \xrightarrow{p} r^*.
\]
and then by the continuous mapping theorem,

\[
\hat{\alpha}_{\text{REG}(m,n)} = \mathbb{E} \left[ f(\Phi(\omega)\hat{r}) \left| \vec{\omega}, \vec{\zeta} \right. \right] \overset{\mathbb{P}}{\rightarrow} \mathbb{E} \left[ f(\Phi(\omega)r^*) \right]. \tag{B.10}
\]

Also notice that equation (3.15) implies

\[
\hat{\alpha}_{\text{REG}(m,n)} - \alpha - B_{M,n} \overset{\mathbb{P}}{\rightarrow} 0. \tag{B.11}
\]

Combining (B.10) and (B.11),

\[
B_{M,n} \overset{\mathbb{P}}{\rightarrow} B^*_{M} = E \left[ f(\Phi(\omega)r^*) \right] - \alpha. \tag{B.12}
\]

Given (3.12) and (B.9),

\[
\left| B_{M,n} - E \left[ f'(L(\omega))M(\omega) \right] \right| \leq \mathbb{E} \left[ \frac{1}{2} f''(L(\omega) + \theta \cdot (\Phi(\omega)\hat{r} - L(\omega))) (\Phi(\omega)\hat{r} - L(\omega))^2 \left| \vec{\omega}, \vec{\zeta} \right. \right] \leq \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (\Phi(\omega)\hat{r} - L(\omega))^2 \left| \vec{\omega}, \vec{\zeta} \right. \right] = \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (\Phi(\omega)\hat{r} - \Phi(\omega)r^*)^2 \left| \vec{\omega}, \vec{\zeta} \right. \right] + \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (M(\omega))^2 \right], \tag{B.13}
\]

where we have used (B.4) in equation (B.13). From Lemma 3.1, we have

\[
\frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (\Phi(\omega)\hat{r} - \Phi(\omega)r^*)^2 \left| \vec{\omega}, \vec{\zeta} \right. \right] = \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (\hat{r} - r^*)^\top \Phi(\omega)^\top \Phi(\omega) (\hat{r} - r^*) \right] = \frac{U_{\text{diff}}}{2} \|\hat{r} - r^*\|_2^2 = \frac{O_p(1)}{n}. \tag{B.14}
\]

From (B.12), (B.13), and (B.14), we have

\[
\left| B^*_{M} - E \left[ f'(L(\omega))M(\omega) \right] \right| \leq \frac{U_{\text{diff}}}{2} \mathbb{E} \left[ (M(\omega))^2 \right].
\]

\[\square\]

### B.2 Lipschitz Continuous Case

**Theorem 3.2.** Suppose that Assumptions 3.F2, 3.A1, 3.A2, and 3.A3 hold. Then as the number of scenarios \( n \to \infty \),

\[
(\hat{\alpha}_{\text{REG}(m,n)} - \alpha)^2 \leq U_{Lip}^2 \mathbb{E} \left[ (M(\omega))^2 \right] + O_p \left( \frac{1}{n} \right).
\]
Proof. Note that
\[ \hat{\alpha}_{\text{REG}(m,n)} - \alpha = E\left[ f(\Phi(\omega) \hat{r}) \right] - E\left[ f(L(\omega)) \right] = E\left[ f(\Phi(\omega) \hat{r}) - f(L(\omega)) \right]. \]

From the Lipschitz continuity condition (3.18) and Jensen’s inequality,
\[ \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right)^2 \leq U_{\text{Lip}}^2 E\left[ (\Phi(\omega)(\hat{r} - r^*))^2 \right] + U_{\text{Lip}}^2 E\left[ (M(\omega))^2 \right], \]

where we have used (B.4). Then, by the orthonormality of \( \Phi(\cdot) \),
\[ \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right)^2 \leq U_{\text{Lip}}^2 E\left[ (\Phi(\omega)(\hat{r} - r^*))^2 \right] + U_{\text{Lip}}^2 E\left[ (M(\omega))^2 \right]. \]

From Lemma 3.1 as \( n \to \infty \),
\[ \sqrt{n}(\hat{r} - r^*) \overset{d}{\to} N\left(0, \Sigma_M + \frac{\Sigma_v}{m}\right). \]

From the continuous mapping theorem, as \( n \to \infty \), \( nU_{\text{Lip}}^2 \| \hat{r} - r^* \|_2^2 \) converges to a generalized chi-square distribution. Therefore, for any \( \epsilon > 0 \), there exist \( \Delta_\epsilon > 0 \) and \( N_\epsilon > 0 \), such that for any \( n > N_\epsilon \),
\[ P\left(nU_{\text{Lip}}^2 \| \hat{r} - r^* \|_2^2 > \Delta_\epsilon \right) < \epsilon, \]

which implies that
\[ U_{\text{Lip}}^2 \| \hat{r} - r^* \|_2^2 = O_P\left(\frac{1}{n}\right). \]

With (B.15), the result follows.

B.2.1 Mean Squared Error of the Estimator

In order to prove Lemma 3.2, Theorem 3.3, and Corollary 3.2 we need the following lemmas.
Lemma B.2. For any $r \in \mathbb{R}^d$,
\[ g(r) - g(r^*) = \|r - r^*\|^2. \]

Therefore, for any $r \in \mathbb{R}^d$ and $R_\rho$ defined by (3.21), we have
\[ R_\rho = \left\{ r \in \mathbb{R}^d : g(r) \leq g(r^*) + \rho \right\}. \]

Proof. By the projection theorem and the fact that $\Phi$ is orthonormal,
\[
\begin{aligned}
g(r) &= \mathbb{E} \left[ \left( \hat{L}(\omega, \zeta) - \Phi(\omega) r \right)^2 \right] \\
&= \mathbb{E} \left[ \left( \hat{L}(\omega, \zeta) - \Phi(\omega) r^* + \Phi(\omega) r^* - \Phi(\omega) r \right)^2 \right] \\
&= \mathbb{E} \left[ (\Phi(\omega)(r - r^*))^2 \right] + \mathbb{E} \left[ \left( \hat{L}(\omega, \zeta) - \Phi(\omega) r^* \right)^2 \right] - 2\mathbb{E} \left[ (\Phi(\omega)(r - r^*)) \left( \hat{L}(\omega, \zeta) - \Phi(\omega) r^* \right) \right] \\
&= \|r - r^*\|^2 + g(r^*) - 2(r - r^*)^\top \mathbb{E} \left[ \Phi(\omega)^\top \left( \hat{L}(\omega, \zeta) - \Phi(\omega) r^* \right) \right] \\
&= \|r - r^*\|^2 + g(r^*). 
\end{aligned}
\]

Given $R_\rho$, define
\[
\hat{r}_\rho \in \arg\min_{r \in R_\rho} \frac{1}{n} \sum_{i=1}^n G\left( r, \omega^{(i)}, \zeta^{(i)} \right), \tag{B.16}
\]
which is the sample optimal solution of $r$ over $R_\rho$. Under Assumptions 3.A1 and 3.A2 according to Lemma B.1, the optimal solution $\hat{r}_\rho$ exists and is unique almost surely.

Lemma B.3. For $\rho > 0$, $\hat{r}_{2\rho} \in R_\rho$ if and only if $\hat{r} \in R_\rho$.

Proof. Notice that $\hat{r} \in R_\rho$ implies $\hat{r}_{2\rho} = \hat{r}$, and thus $\hat{r} \in R_\rho$ implies $\hat{r}_{2\rho} \in R_\rho$.

On the other hand, if $\hat{r}_{2\rho} \in R_\rho$ and $\hat{r} \notin R_\rho$, we must have $\hat{r} \notin R_{2\rho}$. Therefore, we can have
\[
\frac{1}{n} \sum_{i=1}^n G\left( \hat{r}, \omega^{(i)}, \zeta^{(i)} \right) < \frac{1}{n} \sum_{i=1}^n G\left( \hat{r}_{2\rho}, \omega^{(i)}, \zeta^{(i)} \right). 
\]

Then for any $\varphi \in (0, 1)$, by the convexity of $(1/n) \sum_{i=1}^n G(\cdot, \omega^{(i)}, \zeta^{(i)})$,
\[
\frac{1}{n} \sum_{i=1}^n G\left( \varphi \hat{r}_{2\rho} + (1 - \varphi) \hat{r}, \omega^{(i)}, \zeta^{(i)} \right) \leq \varphi \frac{1}{n} \sum_{i=1}^n G\left( \hat{r}_{2\rho}, \omega^{(i)}, \zeta^{(i)} \right) + (1 - \varphi) \frac{1}{n} \sum_{i=1}^n G\left( \hat{r}, \omega^{(i)}, \zeta^{(i)} \right) \\
< \frac{1}{n} \sum_{i=1}^n G\left( \hat{r}_{2\rho}, \omega^{(i)}, \zeta^{(i)} \right),
\]
APPENDIX B. PROOFS RELATED TO THE UNWEIGHTED REGRESSION METHOD

and thus \( \phi \hat{r}_{2\rho} + (1 - \phi) \hat{r} \notin \mathcal{R}_{2\rho} \), i.e.,

\[
g(\phi \hat{r}_{2\rho} + (1 - \phi) \hat{r}) > g(r^*) + 2\rho.
\]

However, we know that \( g(\cdot) \) is convex and thus continuous, and then

\[
\rho + g(r^*) = g(\hat{r}_{2\rho}) = \lim_{\phi \to 1} g(\phi \hat{r}_{2\rho} + (1 - \phi) \hat{r}) \geq g(r^*) + 2\rho,
\]

which is a contradiction.

\[\square\]

Lemma B.4. Suppose that Assumptions 3.F2, 3.A1, 3.A2, 3.A4, and 3.A5 hold. Let \( \rho > 0 \) be an arbitrary constant. Then consider any \( \theta \in (0, \infty) \) and suppose that \( n \geq C' \lambda^2 \rho \left( d \ln \left( \frac{2\sqrt{2C''\Lambda_{2\rho}}}{\sqrt{\rho}} \right) + \ln \left( \frac{1}{\theta} \right) \right) \), (B.17)

where \( \lambda \) is defined in Assumption 3.A5. \( C' \) and \( C'' \) are universal constants, i.e., they do not depend on the problem, and

\[
\Lambda_{\rho} \triangleq (2\sqrt{\rho} + 1) d + 2E[(M(\omega))^2] + 2E[(\varepsilon(\omega, \zeta))^2].
\]

Then,

\[ P(\hat{r} \notin \mathcal{R}_{\rho}) \leq \theta. \]

Proof. When \( \theta \geq 1 \), the result is trivial. When \( \theta \in (0, 1) \), the result follows from Corollary 5.20 in Shapiro et al. (2009) and Lemma B.3 above. In the setting here, we let \( r = 2\rho, \varepsilon = \rho, \delta = 0, \) and then \( a = 2\rho. \) Also notice that in our problem, \( \gamma = 2, c = 1, \) and \( D_a = D_{2\rho}^* = 2\sqrt{2\rho}. \) Further, compared to the notation of Shapiro et al. (2009), we use \( G(\cdot) \) as \( F(\cdot), \) \( g(\cdot) \) as \( f(\cdot), \) \( \Psi_{r,r''}(\cdot) \) as \( M_{x',x}(\cdot), \) \( n \) as \( N, \) \( d \) as \( n, \) and \( \Lambda_{2\rho} \) as \( L. \)

Assumption (M5) in Shapiro et al. (2009) is from our Assumption 3.A4, i.e., the moment generating functions of \( \|\Phi(\omega)\|^2_2, (M(\omega))^2, \) and \( (\varepsilon(\omega, \zeta))^2 \) are finite-valued in a neighborhood of
zero. In particular,
\[
|G(r', \omega, \zeta) - G(r'', \omega, \zeta)|
\]
\[
= \left| \left( \hat{L}(\omega, \zeta) - \Phi(\omega)r' \right)^2 - \left( \hat{L}(\omega, \zeta) - \Phi(\omega)r'' \right)^2 \right|
\]
\[
= \left| \left( \Phi(\omega)r' + \Phi(\omega)r'' - 2\hat{L}(\omega) \right) \Phi(\omega)(r' - r'') \right|
\]
\[
\leq \Phi(\omega)(r' - r^*) + \Phi(\omega)(r'' - r^*) + 2\Phi(\omega)r^* - 2\hat{L}(\omega) \| \Phi(\omega) \|_2 \| r' - r'' \|_2
\]
\[
\leq \left( \| \Phi(\omega) \|_2 \| r' - r^* \|_2 + \| \Phi(\omega) \|_2 \| r'' - r^* \|_2 + 2 \| \hat{L}(\omega, \zeta) - \Phi(\omega)r^* \| \right) \| \Phi(\omega) \|_2 \| r' - r'' \|_2
\]
\[
\leq \left( \| \Phi(\omega) \|_2 \sqrt{2\rho} \| \Phi(\omega) \|_2 \sqrt{2\rho} + 2 \| M(\omega) + \epsilon(\omega, \zeta) \| \| \Phi(\omega) \|_2 \right) \| r' - r'' \|_2
\]
\[
= \left( 2\sqrt{2\rho} \| \Phi(\omega) \|_2^2 + 2 \| M(\omega) + \epsilon(\omega, \zeta) \| \| \Phi(\omega) \|_2^2 \right) \| r' - r'' \|_2
\]
\[
\leq \left( (2\sqrt{2\rho} + 1) \| \Phi(\omega) \|_2^2 + 2 \| M(\omega) \|^2 + 2 \| \epsilon(\omega, \zeta) \|^2 \right) \| r' - r'' \|_2
\]
\[
\leq \left( (2\sqrt{2\rho} + 1) \| \Phi(\omega) \|_2^2 + 2 (M(\omega))^2 + 2 (\epsilon(\omega, \zeta))^2 \right) \| r' - r'' \|_2.
\]

Since in a neighborhood of zero, the finiteness of the moment generating functions of $\| \Phi(\omega) \|_2^2$, $(M(\omega))^2$, and $(\epsilon(\omega, \zeta))^2$ implies the finiteness of the moment generating function of
\[
(2\sqrt{2\rho} + 1) \| \Phi(\omega) \|_2^2 + 2 (M(\omega))^2 + 2 (\epsilon(\omega, \zeta))^2,
\]
Assumption (M5) in \cite{Shapiro2009} is satisfied.

Assumption (M6) in \cite{Shapiro2009} is from the Assumption 3.A5. Notice that Assumption 3.A5 is weaker than Assumption (M6) in \cite{Shapiro2009}, but according to the discussion after Assumption (M6) in \cite{Shapiro2009}, Assumption 3.A5 here is sufficient.

\[\blacksquare\]

**Lemma 3.2.** Suppose that Assumptions 3.F2, 3.A1, 3.A2, 3.A4, and 3.A5 hold. Let $\rho > 0$ be an arbitrary constant. Then for any positive integer $n$,
\[
P(\hat{r} \notin \mathcal{R}_\rho) \leq \left( \frac{2\sqrt{2}C'' \Lambda_2}{\sqrt{\rho}} \right)^d \exp \left( -\frac{\rho n}{C'C''^2} \right),
\]
where $\lambda$ is defined in Assumptions 3.A5, $C'$ and $C''$ are universal constants (i.e., constants that do not depend on the problem), and
\[
\Lambda_\rho \triangleq (2\sqrt{\rho} + 1) d + 2\mathbb{E} \left[ (M(\omega))^2 \right] + 2\mathbb{E} \left[ (\epsilon(\omega, \zeta))^2 \right].
\]
APPENDIX B. PROOFS RELATED TO THE UNWEIGHTED REGRESSION METHOD

Proof. Define

\[ \theta \triangleq \left( \frac{2\sqrt{2}C'' \Lambda_2 \rho}{\sqrt{\rho}} \right)^d \exp\left( -\frac{mn}{C'' \lambda^2} \right). \]

Then,

\[ n = \frac{C' \lambda^2}{\rho} \left( d \ln \left( \frac{2\sqrt{2}C'' \Lambda_2 \rho}{\sqrt{\rho}} \right) + \ln \left( \frac{1}{\theta} \right) \right), \]

which satisfies \( (B.17) \), and thus, by Lemma \( B.4 \),

\[ P (\hat{r}_2 \rho \notin R) \leq \left( \frac{2\sqrt{2}C'' \Lambda_2 \rho}{\sqrt{\rho}} \right)^d \exp\left( -\frac{mn}{C'' \lambda^2} \right). \]

From Lemma \( B.3 \) the result follows.

\[ \square \]

Lemma B.5. For any \( \rho \geq 2 \),

\[ \Lambda_\rho \leq \frac{\sqrt{\rho}}{\sqrt{2}} \Lambda_2. \]

Proof. Notice that for any \( \rho \geq 2 \),

\[
\Lambda_\rho = (2 \sqrt{\rho} + 1) d + 2E \left[ (M(\omega))^2 \right] + 2E \left[ (\varepsilon(\omega, \zeta))^2 \right] \\
= \sqrt{\rho} \left( 2 + \frac{1}{\sqrt{\rho}} \right) d + \frac{2E \left[ (M(\omega))^2 \right]}{\sqrt{\rho}} + \frac{2E \left[ (\varepsilon(\omega, \zeta))^2 \right]}{\sqrt{\rho}} \\
\leq \sqrt{\rho} \left( 2 + \frac{1}{\sqrt{2}} \right) d + \frac{2E \left[ (M(\omega))^2 \right]}{\sqrt{2}} + \frac{2E \left[ (\varepsilon(\omega, \zeta))^2 \right]}{\sqrt{2}} \\
= \frac{\sqrt{\rho}}{\sqrt{2}} \left( (2 \sqrt{2} + 1) d + 2E \left[ (M(\omega))^2 \right] + 2E \left[ (\varepsilon(\omega, \zeta))^2 \right] \right) \\
= \frac{\sqrt{\rho}}{\sqrt{2}} \Lambda_2.
\]

\[ \square \]

Theorem 3.3. Suppose that Assumptions 3.F2, 3.A1, 3.A2, 3.A4, and 3.A5 hold, and let \( \delta > 0 \) be
APPENDIX B. PROOFS RELATED TO THE UNWEIGHTED REGRESSION METHOD

an arbitrary positive constant. Then for any positive integer $n,$

$$E \left[ (\Phi(\omega)(\hat{r} - r^*))^2 \right] = E \left[ \|\hat{r} - r^*\|_2^2 \right] \leq \frac{1}{n^{1-\delta}} + 2\frac{\sqrt{2}}{\sqrt{n}} C' (C''^d (\Lambda_2)^{d/2} n^{(1-s)d-1} \exp \left( -\frac{n\delta}{C'\lambda^2} \right) + \frac{2^d C' (C''^d (\Lambda_2)^d \lambda^2}{n} \exp \left( -\frac{n}{C'\lambda^2} \right) \right) = O \left( n^{-1+\delta} \right).$$

**Proof.** Notice that

$$E \left[ (\Phi(\omega)(\hat{r} - r^*))^2 \right] = E \left[ E \left[ (\Phi(\omega)(\hat{r} - r^*))^2 \mid \tilde{\omega}, \zeta \right] \right] = E \left[ (\hat{r} - r^*)^\top E \left[ \Phi(\omega) \Phi(\omega) \right] (\hat{r} - r^*) \right] = E \left[ \|\hat{r} - r^*\|_2^2 \right] = \int_0^\infty P \left( \|\hat{r} - r^*\|_2^2 > \rho \right) d\rho. \quad (B.18)$$

where we have used Lemma B.2.

Without loss of generality, we consider an arbitrary positive constant $\delta \in (0, 1),$

$$E \left[ \|\hat{r} - r^*\|_2^2 \right] = \int_0^{\frac{1}{n^{1-\delta}}} P \left( \hat{r} \notin \mathcal{R}_\rho \right) d\rho + \int_{\frac{1}{n^{1-\delta}}}^1 P \left( \hat{r} \notin \mathcal{R}_\rho \right) d\rho + \int_1^\infty P \left( \hat{r} \notin \mathcal{R}_\rho \right) d\rho. \quad (B.19)$$

In order to bound (B.19), we bound each term separately. For the first term in (B.19),

$$\int_0^{\frac{1}{n^{1-\delta}}} P \left( \hat{r} \notin \mathcal{R}_\rho \right) d\rho \leq \frac{1}{n^{1-\delta}}.$$

For the second term in (B.19), from Lemma 3.2 and Lemma B.5

$$\int_{\frac{1}{n^{1-\delta}}}^1 P \left( \hat{r} \notin \mathcal{R}_\rho \right) d\rho \leq \int_{\frac{1}{n^{1-\delta}}}^1 \left( \frac{2\sqrt{2} C''^d \Lambda_2}{\sqrt{p}} \right)^d \exp \left( -\frac{\rho m}{C'\lambda^2} \right) d\rho \leq \left( 2\sqrt{2} C''^d \Lambda_2 \right)^d \int_{\frac{1}{n^{1-\delta}}}^1 \rho^{-\frac{d}{2}} \exp \left( -\frac{\rho m}{C'\lambda^2} \right) d\rho.$$
Define $\rho' \triangleq n^{1-\delta} \rho$, i.e., $\rho = \rho' / (n^{1-\delta})$. Then,

\[
\int_{\frac{1}{n^{1-\delta}}}^{1} P(\hat{r} \notin R_\rho) d\rho \leq \left(2\sqrt{2}C''A_2\right)^d \frac{1}{n^{1-\delta}} \int_{1}^{\infty} \left(\rho' / n^{1-\delta}\right)^{-\frac{d}{2}} \exp\left(-\frac{\rho' \lambda^2}{C' \lambda^2}\right) d\rho' \\
= \left(2\sqrt{2}C''A_2\right)^d \frac{1}{(n^{1-\delta})^{1-\frac{d}{2}}} \int_{1}^{\infty} \left(\rho' / n^{1-\delta}\right)^{-\frac{d}{2}} \exp\left(-\frac{\rho' \lambda^2}{C' \lambda^2}\right) d\rho' \\
\leq \left(2\sqrt{2}C''A_2\right)^d \frac{1}{(n^{1-\delta})^{1-\frac{d}{2}}} \int_{1}^{\infty} \exp\left(-\frac{\rho' \lambda^2}{C' \lambda^2}\right) d\rho' \\
\leq \left(2\sqrt{2}C''A_2\right)^d \frac{1}{(n^{1-\delta})^{1-\frac{d}{2}}} \frac{C' \lambda^2}{n^{1-\delta}} \int_{1}^{\infty} \exp\left(-\frac{\rho' \lambda^2}{C' \lambda^2}\right) d\rho' \\
= \left(2\sqrt{2}C''A_2\right)^d \frac{1}{(n^{1-\delta})^{1-\frac{d}{2}}} \frac{C' \lambda^2}{n^{1-\delta}} \exp\left(-\frac{\rho' \lambda^2}{C' \lambda^2}\right).
\]

For the third term in (B.19), with Lemmas 3.2 and B.5,

\[
\int_{1}^{\infty} P(\hat{r} \notin R_\rho) d\rho \leq \int_{1}^{\infty} \left(2\sqrt{2}C''A_2\rho\right)^d \frac{1}{\sqrt{\rho}} \exp\left(-\frac{\rho n^2}{C' \lambda^2}\right) d\rho \\
\leq \int_{1}^{\infty} \left(2C''A_2\right)^d \frac{1}{\sqrt{\rho}} \exp\left(-\frac{\rho n^2}{C' \lambda^2}\right) d\rho \\
= \left(2C''A_2\right)^d \int_{1}^{\infty} \exp\left(-\frac{\rho n^2}{C' \lambda^2}\right) d\rho \\
= \left(2C''A_2\right)^d \frac{C' \lambda^2}{n} \exp\left(-\frac{n^2}{C' \lambda^2}\right).
\]

Therefore, (B.19) becomes

\[
E\left[\|\hat{r} - r^*\|^2\right] \leq \frac{1}{n^{1-\delta}} + \left(2\sqrt{2}C''A_2\right)^d \frac{1}{(n^{1-\delta})^{1-\frac{d}{2}}} \frac{C' \lambda^2}{n^{1-\delta}} \exp\left(-\frac{n^2}{C' \lambda^2}\right) + \left(2C''A_2\right)^d \frac{C' \lambda^2}{n} \exp\left(-\frac{n^2}{C' \lambda^2}\right) \\
= \frac{1}{n^{1-\delta}} + 2^{3d} C' (C'')^d (A_2)^d \lambda^2 n^{(1-\delta)d-1} \exp\left(-\frac{n^2}{C' \lambda^2}\right) + 2^{d} C' (C'')^d (A_2)^d \lambda^2 \frac{C' \lambda^2}{n} \exp\left(-\frac{n^2}{C' \lambda^2}\right).\]

\[\blacksquare\]

**Corollary 3.2.** Suppose that Assumptions 3.A2, 3.A4, and 3.A5 hold, and let $\delta > 0$ be
an arbitrary positive constant. Then, for any positive integer \( n \),

\[
E \left[ \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right)^2 \right] \\
\leq U_{\text{Lip}}^2 \left( 2^{3d} C' (C'')^d (\Lambda_2)^d \lambda^2 n^{(1-\delta)d-1} \exp \left( -\frac{n \delta}{C' \lambda^2} \right) + \frac{2^d C' (C'')^d (\Lambda_2)^d \lambda^2}{n} \exp \left( -\frac{n}{C' \lambda^2} \right) \right) \\
+ U_{\text{Lip}}^2 \left( E \left[ (M(\omega))^2 \right] + n^{-1+\delta} \right) \\
= U_{\text{Lip}}^2 E \left[ (M(\omega))^2 \right] + O \left( n^{-1+\delta} \right).
\]

**Proof.** From (B.15) and Theorem 3.3 we have that

\[
E \left[ \left( \hat{\alpha}_{\text{REG}(m,n)} - \alpha \right)^2 \right] \\
\leq U_{\text{Lip}}^2 E \left[ (M(\omega))^2 \right] + U_{\text{Lip}}^2 E \left[ \|\hat{r} - r^*\|^2 \right] \\
\leq U_{\text{Lip}}^2 \left( E \left[ (M(\omega))^2 \right] \right) \\
+ \frac{1}{n^{1-\delta}} + 2^{3d} C' (C'')^d (\Lambda_2)^d \lambda^2 n^{(1-\delta)d-1} \exp \left( -\frac{n \delta}{C' \lambda^2} \right) + \frac{2^d C' (C'')^d (\Lambda_2)^d \lambda^2}{n} \exp \left( -\frac{n}{C' \lambda^2} \right) .
\]
Appendix C

Proofs Related to the Weighted Regression Method

This section presents the proofs of Theorem 4.1 in Section 4.2. Beside the notations defined in Appendix B, we define some extra notations as follows: \( L(\vec{\omega}) \) is an \( n \times 1 \) column vector,

\[
L(\vec{\omega}) \triangleq (L(\omega^{(1)}), \ldots, L(\omega^{(n)}))^\top,
\]

and \( \vec{H} \) is an \( n \times n \) diagonal matrix,

\[
\vec{H} \triangleq \begin{pmatrix}
h(\omega^{(1)}) & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & h(\omega^{(n)})
\end{pmatrix}.
\]

C.1 Mean Squared Error of the Estimator

We need the following lemma to prove Theorem 4.1

Lemma C.1. For a specific \( h \), if Assumptions 4.F1, 4.A1, 4.A2, 4.A3, and 4.A4 hold, then as \( n \to \infty \),

\[
E \left[ (\Phi(\omega) (\hat{r}(h)) - r^*(h))^2 \right] = O \left( n^{1-\delta} \right),
\]

where \( \delta > 0 \) is an arbitrary positive constant.
Proof. Following the analysis of Section 3 for unweighted regression, under appropriate technical assumptions, as \( n \to \infty \),
\[
E \left[ \| \hat{r} - r^* \|^2 \right] = O \left( n^{-1+\delta} \right), \quad (C.1)
\]
where \( \delta > 0 \) is an arbitrary positive constant.

Substituting \( L(\omega) \) with \( \sqrt{h(\omega)}L(\omega, \zeta) \), \( \hat{L}(\omega, \zeta) \) with \( \sqrt{h(\omega)}\hat{L}(\omega, \zeta) \), and \( \Phi(\omega) \) with \( \sqrt{h(\omega)}\Phi(\omega) \), the assumptions in Section 3 become Assumptions 4.F1, 4.A1, 4.A2, 4.A3, and 4.A4 here, and the regression coefficients \( r^* \) and \( \hat{r} \) become \( r^*(h) \) and \( \hat{r}(\tilde{h}) \). Then we can directly apply (C.1) and derive that, as \( n \to \infty \),
\[
E \left[ \| \hat{r}(\tilde{h}) - r^*(h) \|^2 \right] = O \left( n^{-1+\delta} \right).
\]
Moreover,
\[
E \left[ (\Phi(\omega)(\hat{r}(\tilde{h}) - r^*(h)))^2 \right] = E \left[ E \left[ (\Phi(\omega)(\hat{r}(\tilde{h}) - r^*(h)))^2 \mid \omega, \zeta^2 \right] \right]
\]
\[
= E \left[ (\hat{r}(\tilde{h}) - r^*(h))^\top E \left[ \Phi(\omega)^\top \Phi(\omega) \right] (\hat{r}(\tilde{h}) - r^*(h)) \right]
\]
\[
= E \left[ \| \hat{r}(\tilde{h}) - r^*(h) \|^2 \right]
\]
\[
= O \left( n^{-1+\delta} \right),
\]
where \( E[\Phi(\omega)^\top \Phi(\omega)] = 1 \) is from the orthonormality assumed in Assumption 4.A2.

\[ \blacksquare \]

Lemma C.1 establishes that the mean squared error between our approximation and the best approximation decays at the rate \( n^{-1+\delta} \) for any \( \delta > 0 \). Now we have the following theorem showing the limit of (4.4).

Theorem 4.1. For a specific \( h \), if Assumptions 4.F1, 4.A1, 4.A2, 4.A3, and 4.A4 hold, then
\[
\lim_{n \to \infty} E \left[ ((\hat{\alpha} \text{REG}(m,n,h)) - \alpha)^2 \right] = \left( E \left[ f(\Phi(\omega)r^*(h)) \right] - E \left[ f(L(\omega)) \right] \right)^2.
\]
Proof. Decomposing the MSE of \( \hat{\alpha}_{\text{REG}(m,n,h)} \), we have

\[
E \left[ (\hat{\alpha}_{\text{REG}(m,n,h)} - \alpha)^2 \right] = E \left[ \left( E \left[ f(\Phi(\omega)\hat{r}(\vec{h})) | \vec{\omega}, \vec{\zeta} \right] - E \left[ f(\Phi(\omega)r^*(h)) \right] \right)^2 \right] + E \left[ \left( E \left[ f(\Phi(\omega)r^*(h)) \right] \right)^2 \right] + 2E \left[ E \left[ f(\Phi(\omega)r^*(h)) \right] \left( E \left[ f(\Phi(\omega)r^*(h)) \right] - E \left[ f(L(\omega)) \right] \right) \right].
\]

We analyze the three terms in (C.2) separately. The first term in (C.2) satisfies

\[
E \left[ \left( E \left[ f(\Phi(\omega)\hat{r}(\vec{h})) | \vec{\omega}, \vec{\zeta} \right] - E \left[ f(\Phi(\omega)r^*(h)) \right] \right)^2 \right] \leq U_Lip^2 E \left[ \left( E \left[ f(\Phi(\omega)(\hat{r}(\vec{h}) - r^*(h)) | \vec{\omega}, \vec{\zeta} \right] \right)^2 \right] \leq U_Lip^2 E \left[ (\Phi(\omega)(\hat{r}(\vec{h}) - r^*(h)))^2 \right].
\]

Then by Lemma C.1 as \( n \to \infty \), this term vanishes. The third term in (C.2) satisfies

\[
2E \left[ E \left[ f(\Phi(\omega)r^*(h)) \right] \left( E \left[ f(\Phi(\omega)r^*(h)) \right] - E \left[ f(L(\omega)) \right] \right) \right] \leq 2U_Lip E \left[ \left( E \left[ f(\Phi(\omega)(\hat{r}(\vec{h}) - r^*(h))) \right] \right)^2 \right] \leq 2U_Lip \sqrt{E \left[ (\Phi(\omega)(\hat{r}(\vec{h}) - r^*(h)))^2 \right] E \left[ f(\Phi(\omega)r^*(h)) \right] - E \left[ f(L(\omega)) \right]}.\]

Then by Lemma C.1 as \( n \to \infty \), this term also vanishes. Since the second term in (C.2) is independent of \( n \), the theorem follows.