# Rare-Event Estimation and Calibration for Large-Scale Stochastic Simulation Models 

Yuanlu Bai

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

## COLUMBIA UNIVERSITY

© 2023

Yuanlu Bai
All Rights Reserved

Abstract<br>Rare-Event Estimation and Calibration for Large-Scale Stochastic Simulation Models Yuanlu Bai

Stochastic simulation has been widely applied in many domains. More recently, however, the rapid surge of sophisticated problems such as safety evaluation of intelligent systems has posed various challenges to conventional statistical methods. Motivated by these challenges, in this thesis, we develop novel methodologies with theoretical guarantees and numerical applications to tackle them from different perspectives. In particular, our works can be categorized into two areas: (1) rare-event estimation (Chapters 2 to 5) where we develop approaches to estimating the probabilities of rare events via simulation; (2) model calibration (Chapters 6 and 7) where we aim at calibrating the simulation model so that it is close to reality.

In Chapter 2, we study rare-event simulation for a class of problems where the target hitting sets of interest are defined via modern machine learning tools such as neural networks and random forests. We investigate an importance sampling scheme that integrates the dominating point machinery in large deviations and sequential mixed integer programming to locate the underlying dominating points. We provide efficiency guarantees and numerical demonstration of our approach. In Chapter 3, we propose a new efficiency criterion for importance sampling, which we call probabilistic efficiency. Conventionally, an estimator is regarded as efficient if its relative error is sufficiently controlled. It is widely known that when a rare-event set contains multiple "important regions" encoded by the dominating points, importance sampling needs to account for all of them via mixing to achieve efficiency. We argue that the traditional analysis recipe could
suffer from intrinsic looseness by using relative error as an efficiency criterion. Thus, we propose the new efficiency notion to tighten this gap. In particular, we show that under the standard Gartner-Ellis large deviations regime, an importance sampling that uses only the most significant dominating points is sufficient to attain this efficiency notion. In Chapter 4, we consider the estimation of rare-event probabilities using sample proportions output by crude Monte Carlo. Due to the recent surge of sophisticated rare-event problems, efficiency-guaranteed variance reduction may face implementation challenges, which motivate one to look at naive estimators. In this chapter we construct confidence intervals for the target probability using this naive estimator from various techniques, and then analyze their validity as well as tightness respectively quantified by the coverage probability and relative half-width. In Chapter 5, we propose the use of extreme value analysis, in particular the peak-over-threshold method which is popularly employed for extremal estimation of real datasets, in the simulation setting. More specifically, we view crude Monte Carlo samples as data to fit on a generalized Pareto distribution. We test this idea on several numerical examples. The results show that in the absence of efficient variance reduction schemes, it appears to offer potential benefits to enhance crude Monte Carlo estimates. In Chapter 6, we investigate a framework to develop calibration schemes in parametric settings, which satisfies rigorous frequentist statistical guarantees via a basic notion that we call eligibility set designed to bypass non-identifiability via a set-based estimation. We investigate a feature extraction-then-aggregation approach to construct these sets that target at multivariate outputs. We demonstrate our methodology on several numerical examples, including an application to calibration of a limit order book market simulator. In Chapter 7, we study a methodology to tackle the NASA Langley Uncertainty Quantification Challenge, a model calibration problem under both aleatory and epistemic uncertainties. Our methodology is based on an integration of distributionally robust optimization and importance sampling. The main computation machinery in this integrated methodology amounts to solving sampled linear programs. We present theoretical statistical guarantees of our approach via connections to nonparametric hypothesis testing, and numerical performances including parameter calibration and downstream decision and risk evaluation tasks.

## Table of Contents

Acknowledgments ..... vii
Dedication ..... viii
Chapter 1: Introduction ..... 1
Chapter 2: Rare-Event Simulation for Neural Network and Random Forest Predictors ..... 5
2.1 Introduction ..... 5
2.2 Related Work ..... 9
2.3 Problem Setting ..... 11
2.4 Efficient Importance Sampling via Sequential Mixed Integer Programming ..... 14
2.4.1 Basics of Importance Sampling ..... 14
2.4.2 Dominating Points and Mixture Importance Samplers ..... 16
2.4.3 Efficiency Guarantees ..... 20
2.5 Tractable Optimization Formulation for Prediction Models ..... 22
2.5.1 Tractable Formulation for Random Forest ..... 22
2.5.2 Tractable Formulation for Neural Network ..... 25
2.6 Experiments ..... 30
2.6.1 Toy Problems ..... 30
2.6.2 MAGIC Gamma Telescope Data Set ..... 37
2.7 Proofs of Theorems ..... 38
2.8 Conclusion ..... 49
2.9 Supplementary A: Training Details for Random Forests ..... 50
Chapter 3: Over-Conservativeness of Variance-Based Efficiency Criteria and Probabilistic Efficiency in Rare-Event Simulation ..... 51
3.1 Introduction ..... 51
3.2 Problem Setting and Existing Framework ..... 56
3.2.1 Crude Monte Carlo and Relative Error ..... 56
3.2.2 Importance Sampling and Asymptotic Efficiency ..... 57
3.2.3 Large Deviations and Dominating Points ..... 58
3.2.4 Asymptotically (In)efficient Importance Samplers ..... 61
3.3 Motivating Experimental Results ..... 63
3.3.1 Large Deviations of an I.I.D. Sum ..... 64
3.3.2 Overshoot Probability of Random Walk ..... 65
3.3.3 Robustness Assessment for an MNIST Classification Model ..... 67
3.4 Probabilistic Efficiency ..... 70
3.5 Probabilistically Efficient Estimation in the Gartner-Ellis Regime ..... 76
3.6 Further Numerical Experiments and Discussions ..... 82
3.6.1 Verifying Conditions for Probabilistic Efficiency ..... 83
3.6.2 Illustration of Confidence Intervals ..... 88
3.7 Future Work ..... 93
3.7.1 Developing Diagnosis for Finite-Sample Under-Estimation ..... 93
3.7.2 Further Developing Theory of IS with Missed Dominating Points ..... 94
3.8 Supplementary A: Algorithms ..... 96
3.9 Supplementary B: Additional Numerical Experiments ..... 96
3.9.1 Overshoot Probability of a Random Walk. ..... 96
3.9.2 Two-sided Overshoot Probability of a Random Walk. ..... 101
3.10 Supplementary C: A New Alternative Asymptotic Regime and Theoretical Guar- antees ..... 104
3.10.1 Guarantees for General Input Distribution ..... 105
3.10.2 Guarantees for Gaussian Input Distribution ..... 109
3.11 Supplementary D: Proofs ..... 111
Chapter 4: On the Error of Naive Rare-Event Monte Carlo Estimator ..... 132
4.1 Introduction ..... 132
4.2 Problem Setting and Main Results ..... 133
4.3 Confidence Intervals under the Standard Setting ..... 137
4.3.1 Confidence Intervals via Normal Approximation ..... 138
4.3.2 Valid Confidence Intervals ..... 140
4.4 Confidence Intervals under Targeted Stopping ..... 148
4.4.1 CLT Confidence Interval and Wilson's Confidence Interval ..... 148
4.4.2 Valid Confidence Intervals ..... 149
4.5 Numerical Experiments ..... 155
4.5.1 Experiments under the Standard Setting ..... 155
4.5.2 Experiments under the Targeted Stopping Setting ..... 156
4.6 Conclusion ..... 157
4.7 Supplementary A: Proofs ..... 158
Chapter 5: Rare-Event Simulation without Variance Reduction: An Extreme Value Theory Approach ..... 164
5.1 Introduction ..... 164
5.2 Background ..... 166
5.2.1 Challenge of Crude MC ..... 166
5.2.2 Recap of POT ..... 167
5.3 Numerical Experiments ..... 169
5.3.1 Example 1: Sample Mean ..... 169
5.3.2 Example 2: Random Walk ..... 173
5.3.3 Example 3: Neural Network ..... 176
5.3.4 Example 4: Queueing System ..... 177
5.4 Conclusion ..... 178
Chapter 6: Calibrating Over-parameterized Simulation Models: A Framework via Eligibil- ity Set ..... 180
6.1 Introduction ..... 180
6.1.1 Comparisons with Existing Approaches ..... 185
6.2 Calibration Framework ..... 187
6.2.1 Remedying Non-Identifiability via Eligibility Sets ..... 188
6.2.2 Constructing Eligibility Sets: An Elementary Case ..... 192
6.3 Elementary Numerics ..... 199
6.3.1 An M/M/1 Example ..... 199
6.3.2 A G/G/1 Example ..... 201
6.4 Feature-Based Construction of Eligibility Sets ..... 201
6.4.1 Feature Extraction ..... 203
6.4.2 Feature Aggregation ..... 206
6.5 Multi-Agent Simulator Calibration ..... 215
6.5.1 Comparisons of Feature Extraction and Aggregation Techniques ..... 216
6.5.2 Realism Metrics ..... 221
6.6 Supplementary A: Proofs of Theorems ..... 224
6.7 Supplementary B: Details on Machine Learning Models and Feature Extraction ..... 233
6.8 Supplementary C: Experimental Details ..... 235
Chapter 7: Model Calibration via Distributionally Robust Optimization: On the NASA Langley Uncertainty Quantification Challenge ..... 241
7.1 Overview of Our Methodology (Problem A) ..... 242
7.2 A DRO Perspective ..... 243
7.3 Constructing Discrepancy Measures ..... 245
7.4 Determining Existence of an Aleatory Distribution ..... 247
7.5 Towards the Main Procedure ..... 250
7.6 Related Literature ..... 251
7.7 Summarizing Discrete-Time Histories using Fourier Transform ..... 254
7.8 Uncertainty Reduction (Problem B) ..... 256
7.8.1 Ranking Epistemic Parameters (B. 1 and B.2) ..... 256
7.8.2 Impact of the value of $n_{1}$ (A.2) ..... 257
7.8.3 Updated Parameter Ranking (B.3) ..... 258
7.9 Reliability of Baseline Design (Problem C) ..... 259
7.9.1 Failure Probabilities and Severity (C.1, C. 2 and C.5) ..... 259
7.9.2 Rank for Uncertainties (C.3) ..... 260
7.9.3 Representative Realizations (C.4) ..... 260
7.10 Reliability-Based Design (Problem D) ..... 263
7.11 Design Tuning (Problem E) ..... 265
7.12 Risk-Based Design (Problem F) ..... 267
7.13 Discussion ..... 269
7.14 Supplementary A: Proof of Theorem 7.2 ..... 270
References ..... 273

## Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Henry Lam. Throughout my PhD journey, his guidance has been indispensable. I am truly grateful for the opportunities he provided me to develop my skills, and also for his support and understanding. Furthermore, I am deeply grateful to Ton Dieker, Garud Iyengar, David Yao, and Bruno Tuffin (INRIA Rennes) for serving on my dissertation committee and providing invaluable feedback.

I also appreciate working with all my collaborators. In particular, I would like to thank Zhiyuan Huang (Tongji University), Ding Zhao (Carnegie Mellon University), Sebastian Engelke (University of Geneva), Haoxian Chen, Tucker Balch (JPMorgan AI Research), Danial Dervovic (JPMorgan AI Research), and Svitlana Vyetrenko (JPMorgan AI Research) for their contributions to the materials in this thesis.

I would also like to extend my heartfelt thanks to all my friends in the Columbia community. Especially, I would like to thank my boyfriend Jitong Qi for accompanying me through all the ups and downs during these years.

Finally, I want to thank my family for their unconditional love and constant belief in my potential, especially my mother Fumei Wang, my father Jingyi Bai, my grandmother Xiangfeng Zhang, and my grandfather Junxuan Wang. I would not have been able to accomplish this without their unwavering support.

To my family.

## Chapter 1: Introduction

Stochastic simulation has been widely applied in many domains such as queueing systems, finance, insurance and engineering reliability. It is especially useful in the cases where real-world data are difficult or expensive to collect. More recently, however, the rapid surge of sophisticated problems such as safety evaluation of intelligent systems has posed various challenges to conventional statistical methods. On the one hand, the model structures are becoming complicated and even black-box, and hence it is hard or impossible to solve the problem analytically. On the other hand, large-scale simulation models can be expensive to run, which causes high computational demand. Therefore, in the modern context, in many cases it is still an open problem how to efficiently obtain a reliable conclusion. Motivated by these challenges, in this thesis, we develop novel methodologies with theoretical guarantees and numerical applications to tackle them from different perspectives. In particular, our works can be categorized into two areas: (1) rare-event estimation (Chapters 2 to 5) about estimating the probabilities of rare events via simulation; (2) model calibration (Chapters 6 and 7) about calibrating the simulation model so that it is close to reality.

Risk management is an important topic in the recent applications, which is closely related to rare-event estimation. For instance, for autonomous vehicles, it is essential to carefully assess the safety by estimating the risk probabilities such as crash rate before large-scale deployment. The main challenge of applying crude Monte Carlo (MC) methods is that due to the intrinsic rarity, the event of interest seldom occurs, and hence it requires a large sample size to guarantee an accurate estimate. Traditionally, in order to improve the efficiency, various variance reduction techniques have been developed which alter the naive sampling procedure so that adequate accuracy is attained with less samples. Among them, importance sampling (IS) is one of the most popular approaches, which generates simulation samples under an alternative probability measure and then corrects the
bias with a likelihood ratio term. However, the efficiency of the IS estimator relies heavily on the choice of the IS measure. Thus, to obtain a good IS estimator, we need to carefully design the scheme based on the problem structures and also justify its efficiency.

Conventionally, an IS estimator is regarded as efficient if its relative error, namely the ratio between its standard deviation and mean, is sufficiently controlled. It is widely known that when a rare-event set contains multiple "important regions" encoded by the dominating points, IS needs to account for all of them via mixing to achieve efficiency. Following this idea, Chapter 2 studies rare-event simulation for a class of problems where the target hitting sets of interest are defined via modern machine learning tools such as neural networks and random forests. This problem is motivated from fast emerging studies on the safety evaluation of intelligent systems, robustness quantification of learning models, and other potential applications to large-scale simulation in which machine learning tools can be used to approximate complex rare-event set boundaries. We investigate an IS scheme that integrates the dominating point machinery and sequential mixed integer programming to locate the underlying dominating points. Our approach works for a range of neural network architectures including fully connected layers, rectified linear units, normalization, pooling and convolutional layers, and random forests built from standard decision trees. We provide efficiency guarantees and numerical demonstration of our approach using a classification model in the UCI Machine Learning Repository.

While mixing all the dominating points is guaranteed to satisfy the conventional efficiency criterion, in high-dimensional settings the computational effort to locate all the dominating points can be enormous. In Chapter 3, we propose a new efficiency criterion for IS, which we call probabilistic efficiency. We argue that in typical experiments, missing less significant dominating points may not necessarily cause inefficiency, and the traditional analysis recipe could suffer from intrinsic looseness by using relative error, or in turn estimation variance, as an efficiency criterion. Thus, we propose the new efficiency notion to tighten this gap. In particular, we show that under the standard Gartner-Ellis large deviations regime, an IS that uses only the most significant dominating points is sufficient to attain this efficiency notion.

While variance reduction techniques are demonstrated to be powerful in many applications, as mentioned above, they often rely on tractable problem structures that allow careful algorithmic design in order to attain good performance. Unfortunately, this requirement could still be difficult or even impossible to meet in complex practical applications. Motivated by this limitation, in Chapters 4 and 5, we investigate how to take advantage of crude MC samples when variance reduction is hard to apply. Chapter 4 considers the estimation of rare-event probabilities using sample proportions output by crude MC. Unlike using variance reduction techniques, this naive estimator does not have a priori relative efficiency guarantee. The implementation challenges of efficiency-guaranteed variance reduction motivate us to look at naive estimators. In this chapter we construct confidence intervals for the target probability using this naive estimator from various techniques, and then analyze their validity as well as tightness respectively quantified by the coverage probability and relative half-width. Chapter 5 proposes the use of extreme value analysis, in particular the peak-over-threshold (POT) method which is popularly employed for extremal estimation of real datasets, in the simulation setting. More specifically, we view crude MC samples as data to fit on a generalized Pareto distribution. We test this idea on several numerical examples. The results show that our POT estimator appears more accurate than crude MC and, while crude MC can easily give a trivial probability estimate 0 , POT outputs a non-trivial estimate with a roughly correct magnitude. Therefore, in the absence of efficient variance reduction schemes, POT appears to offer potential benefits to enhance crude MC estimates.

Even if the problem is not related to rare events, the simulation result could still be misleading if the model is far off reality. Thus, to ensure reliable decision making, model calibration is also an important task, where we calibrate and validate simulation models against real data. Conventional methods approach these tasks by assessing the model-data match via simple hypothesis tests or distance minimization in an ad hoc fashion. However, for complex simulation models that lack analytical tractability, they can encounter challenges arising from non-identifiability and high dimensionality. More specifically, first, when the simulation model is too complicated, it is possible that different input configurations result in the same output distribution and hence they are indistin-
guishable from output-level data, which is called non-identifiability. Second, when the simulation model has high-dimensional input or output, it is difficult to measure the discrepancy between the model and reality, and to calibrate the model to shrink this discrepancy.

To address these challenges, in Chapters 6 and 7, we develop effective calibration methods involving feature extraction, distance metrics and optimization. Besides theoretical guarantees, we also present applications to such complex but practical simulation models. In Chapter 6, we investigate a framework to develop calibration schemes in parametric settings, which satisfies rigorous frequentist statistical guarantees via a basic notion that we call eligibility set designed to bypass non-identifiability via a set-based estimation. We investigate a feature extraction-then-aggregation approach to construct these sets that target at high-dimensional outputs. We demonstrate our methodology on several numerical examples, including an application to a multi-agent financial market simulator in our cooperation with JPMorgan Chase. In Chapter 7, we study a methodology to tackle the NASA Langley Uncertainty Quantification Challenge, a physical system calibration problem under both aleatory and epistemic uncertainties. That is, besides parameters to be calibrated, the model also contains random factors with unknown nonparametric distributions. Our methodology is based on an integration of robust optimization, more specifically a recent line of research known as distributionally robust optimization, and IS in MC simulation. The main computation machinery in this integrated methodology amounts to solving sampled linear programs. We present theoretical statistical guarantees of our approach via connections to nonparametric hypothesis testing, and numerical performances including parameter calibration and downstream decision and risk evaluation tasks.

# Chapter 2: Rare-Event Simulation for Neural Network and Random Forest Predictors 

### 2.1 Introduction

Due to the extensive development of artificial intelligence (AI), machine learning techniques have been embedded in many safety-sensitive physical systems, including autonomous vehicles [1] and unmanned aircraft [2]. In autonomous vehicles, for instance, machine learning predictors can be applied to many tasks including perception [3, 4], path planning [5, 6], motion control [7], or end-to-end driving systems [8, 9]. In these tasks, misprediction can cause catastrophic impacts on public safety, as exemplified by the series of fatal accidents encountered by autonomous driving systems due to the failures in detecting nearby vehicles or pedestrians (e.g. [10, 11]). To reduce the risk of such catastrophe, machine learning models in these systems need to be carefully evaluated against safety, especially before their mass deployment in public.

Recent research considers using probabilistic measures to quantify the risks of machine learning predictors or entire intelligent physical systems. These measures can be defined in a variety of ways. In robustness evaluation, a prediction model, with neural network as a dominant example, is considered more robust if it is more likely to make a consistent prediction under small perturbations on the input [12]. When the perturbation is modeled via a random distribution, the robustness of neural networks is measured by the probability that the prediction value persists [13, 14, 15]. In more complex intelligent system evaluation, risks can be quantified by the occurrence probabilities of safety-critical events. These events can be defined as the violation in terms of certain safety metrics (e.g., [16] listed seven potential safety metrics for autonomous vehicles including crashes per driving hour and disengagements per scenario), and recent studies use the probabilities of crash or injury in driving tasks as safety metrics [17, 18, 19, 20]. For AI-equipped autonomous vehicles, the
evaluation target would implicitly involve a probabilistic measurement on the embedded machine learning model. Moreover, in [21], neural networks are further used to approximate sophisticated safety-critical sets defined from complex system dynamics, and the target probabilities comprise hitting sets defined via these neural network outputs.

Our study is motivated from the estimation of probabilistic risk measures described above. Due to the complexity of machine learning predictors, these probabilities are typically unamenable to analytical formulas, even when the underlying stochastic distribution is fully modeled. This thus calls for the use of Monte Carlo simulation. However, the target probabilities, which signify the risks of dangerous yet unlikely events, are tiny. The problem thus falls into the domain of rareevent simulation, in which it is widely known that crude Monte Carlo can be extremely inefficient and variance reduction is necessarily employed. Traditionally, rare-event simulation techniques (e.g. [22, 23]) have been applied in broad application areas including queueing systems [24, 25, $26,27,28,29,30,31]$, communication networks [32, 33], finance [34, 35, 36], insurance [37, 38,39 ], reliability [40, 41, 42, 43, 44], biological processes [45, 46], dynamical systems [47, 48], and combinatorics [49, 50]. The evaluation of machine learning models and intelligent physical systems that we focus on here is a new application that is propelled rapidly by the growth of AI. Our goal is to provide a first step into building rare-event simulation algorithms in these applications, which integrate tools from both the disciplines of machine learning and rare-event simulation, and which are statistically guaranteed in terms of the classical efficiency notions in the rare-event literature.

More specifically, we study importance sampling (IS) [51] to design efficient estimators. In rare-event estimation, the rarity nature of hitting set dictates that crude Monte Carlo samples have a low frequency of observing the hitting occurrence, and this inefficiency exhibits statistically as a large relative error (i.e., ratio of standard deviation to mean) in the estimation. To mitigate this issue, IS uses an alternate distribution to generate samples that can attain a higher frequency in hitting the target event, and reweights the outputs to maintain unbiasedness via the likelihood ratios. To achieve a small relative error, the new generating distribution (i.e., the IS distribution)
is carefully selected, often by analyzing the weights in interaction with the hitting set geometry and the underlying system dynamics [52,53]. In this chapter, we follow the above analysis path in the literature and use the common theoretical notion of efficiency called asymptotic optimality or logarithmic efficiency $[54,23,40]$ that we will detail in the sequel.

In terms of our scope of study, we focus on piecewise linear machine learning predictors, which include random forests and neural networks with common activation functions such as rectified linear units (ReLU). The former is an ensemble or weighted average of decision trees [55], and the latter is a network of neurons connected in multiple layers, via the activation functions [56]. We also assume the underlying distribution is Gaussian or mixtures of such. Under this setting, we design provably efficient IS schemes to estimate rare-event probabilities that the prediction outputs hit above certain high thresholds. We will describe how our considered setup relates to the risk quantification of AI -driven algorithms or intelligent physical systems presented earlier, where our proposed approach provides a rigorous first step towards the resulting rare-event simulation problems (see Section 2.3).

Our main methodology integrates the classical notion of dominating points for rare-event sets with sequential mixed integer programming (MIP) to attain an efficient estimator. The notion of dominating points, and the associated mixture-based IS scheme, is well-known in the literature [53, 57]. The MIP, while conceptually straightforward, requires leveraging recent formulations catered for the considered machine learning models. Let us explain the roles of these tools. Intuitively, a dominating point is the highest-density point in the rare-event set, so that using an IS distribution that shifts the mean to this point (via exponential tilting) gives rise to a distribution that hits the rare-event set more frequently, and the generated likelihood ratio contributes properly to the probability of interest, which are desirable for controlling the relative error. However, this is only a local characterization, as the simulation randomness could cause huge likelihood ratios for some generated samples. Controlling these ratios in turn requires a geometric property that, in the Gaussian case, implies the dominating point to be on the boundary of the rare-event set, and that the latter lies completely inside one of the half-spaces cut by the tangential hyperplane passing
through the dominating point (e.g., these occur when the rare-event set is convex). When this geometric property does not hold, then one needs to divide the rare-event set into a union of smaller sets each bearing its own dominating point, and an efficient IS scheme is built via a mixture of exponential tiltings targeted at all these individual dominating points [53]. The sequential MIP in our procedure serves to locate all these dominating points. It casts the search as a density maximization problem constrained by hitting sets induced from the considered machine learning model. The involved feasible regions shrink sequentially as we add more "cutting planes" to the constraints in order to remove the half-spaces that are already considered by earlier dominating points. Our MIPs are derived from the reformulation techniques that appeared recently in the machine learning literature, which leverage the geometric structures of ReLU neural networks [58] and random forests [59]. We provide a step-by-step guide in formulating random forests and different neural network architectures as suitable MIPs to be inserted into our sequential algorithm.

In terms of theoretical results, we show asymptotic optimality of our IS that targets at general piecewise polyhedrons, which apply to our considered rare-event sets in particular. Towards this, we also derive large deviations results for the associated probabilities of interest. Our results are developed under a different regime from the conventional one in the literature. More specifically, the latter typically scales the input random vector that falls into a fixed set (e.g. [53, 60]), while we let the exceedance threshold on the output of the machine learning model to scale. Our setting is more natural since the threshold provides meaning in defining the level of risk (e.g., in vehicle safety test, the relative velocity at the crash time can be used to compute a so-called Maximum Abbreviated Injury Score that predicts the severity of injuries [61], and hence the probabilities of relative velocity at the crash time exceeding different thresholds are of interest). To this end, the closest work that studies a similar regime is [62], but it only analyzes the tail probability that a Gaussian random vector is componentwise larger than a threshold, which is a simplified version of our regime without the machine learning transformation. While we leverage the results in [62], we also develop mathematical techniques to make the generalization fit our setting.

The chapter is organized as follows. Section 2.2 first provides a literature review. Section 2.3
describes and motivates our problem setting. Section 2.4 presents our algorithm and theoretical guarantees. Section 2.5 provides the MIP formulations for random forests and different neural network architectures. Section 2.6 shows numerical results. Section 2.7 contains the proofs of theorems. Section 2.8 summarizes this chapter and discusses future works.

### 2.2 Related Work

A significant line of work studies the use of large deviations to invent efficient IS procedures, which mathematically identifies the most likely path to trigger a rare event through minimizing the so-called rate function (see, e.g., the surveys [22,54, 63, 34, 23, 64]). This approach leads to the concept of dominating points and mixture IS $[53,57]$ which our work follows. Despite this utilization, our work differs from the previous works. First is that our considered machine learning models, including random forests and neural networks, deem the rare-event boundaries to be only expressible implicitly. This in turn necessitates the use of sequential MIP algorithm that can leverage such expressions in the search of the dominating points. This distinguishes our approach from [65] that similarly considers splitting rare-event sets via dominating points, but constrain the rare-event sets to be unions of half-spaces that are explicitly given. Second, we derive asymptotic results for the rare-event probability of interest and show efficiency of our algorithm, as the exceedance threshold increases, a regime subtly different from the majority of literature yet more natural in our setting. To this end, [62] appears closest to our work, with derived bounds and asymptotic results for the tail probability of Gaussian random vectors. However, our setting is considerably more complex as it involves piecewise linear machine learning predictor output, and correspondingly requires more intricate analysis coming from the geometry of the rare-event set. Next, similar to our derivations, [60] represents the asymptotic of probability on convex sets using dominating points, but they focus on a different scaling from ours. Specifically, like in standard large deviations theory, they focus on the conventional regime where the scaled componentwise maximum of Gaussian random vectors lies in a fixed convex set, while our target event is that the predictor output with Gaussian input (which is not scaled and cannot be expressed
as a componentwise maximum) exceeds an increasing threshold.
In the machine learning literature, some studies use probabilistic measures to evaluate the robustness of prediction models. Since these measures can be extremely small, rare-event simulation techniques are considered. [14] discusses an adaptive multilevel splitting approach to estimate the statistical robustness of machine learning models. [66] considers the problem of estimating agent failure probabilities and proposes to learn a failure probability predictor to approximate the minimum-variance IS distribution. [13] proposes an approach to compute the lower and upper bounds for a probabilistic robustness measure. Our work is motivated by the topics studied in these works, and can be viewed as a step towards the provision of rigorous guarantees for methodologies driven by the corresponding applications.

Another related line of research studies optimization problems with machine learning models in the objective. [59] discusses the optimization of tree ensemble models and provides treatment for large scale problems. [58] formulates the robustness verification of neural networks as MIP problems. These studies leverage the piecewise linear property of these machine learning models to turn optimization on the prediction output into tractable MIPs. Our MIP formulations for finding dominating points follow from these optimization studies.

We close this literature review by briefly discussing other IS schemes. The cross-entropy method $[67,68,69,70]$ uses sequential stochastic optimization to search for an optimal IS distribution in a parametric family. Adaptive IS [71, 72, 73, 74] updates the IS distribution iteratively between simulated replications to approach the optimal (zero-variance) IS distribution and generates non i.i.d samples for estimating the target expectation associated with finite-state discrete Markov chains. Another line of studies use techniques such as Markov-chain Monte Carlo (MCMC) to sample from the rare-event set of interest, or approximately from the conditional distribution given the occurrence of the rare event [75, 76, 77]. IS schemes have also been designed for heavy-tailed systems [78, 79, 80, 33, 81, 82, 83], in contrast to the light-tailed settings considered in this chapter. Besides IS, other competing methods for rare-event simulation include conditional Monte Carlo [84, 85] and splitting [86, 87, 88, 89, 90].

### 2.3 Problem Setting

We state our problem setting. Consider a prediction model $g(\cdot)$, with input $X \in \mathbb{R}^{d}$ and output $g(X) \in \mathbb{R}$. Suppose that the input follows a Gaussian distribution, i.e, $X \sim N(\mu, \Sigma)$, where $\Sigma$ is a $d \times d$ positive definite matrix. We want to estimate the probability $p=P(X \in S)$, where $S=\{x: g(x) \geq \gamma\}$ is a rare-event set with a threshold $\gamma \in \mathbb{R}$ that triggers the rare event. We note that the Gaussian assumption can be relaxed without much difficulty in our framework to, for instance, mixtures of Gaussians, which we will discuss later and can expand our scope of applicability.

This problem setting is motivated from risk assessments involving machine learning models, as exemplified below.

Example 2.1 (Statistical Robustness Metric). Studies on robustness of machine learning models have become increasingly prevalent in recent years. The topic was initiated in computer vision studies [12], where neural networks for image classification were found to be vulnerable to tiny perturbation to the input. Such a perturbed input is considered as an adversarial example. Studies have discussed how to find these adversarial examples [91] and to conduct adversarial learning [92] in more general machine learning tasks. The vulnerability to perturbation has caused safety and security concerns about using machine learning models in real-life applications. In order to evaluate how robust a prediction model is under potential perturbations, robustness metrics are proposed as quantitative benchmarks.

For instance, we consider a classification model $g(\cdot)$ that can correctly predict the input $x_{0}$ as category c. Intuitively, the model is "robust" at $x_{0}$ if the correct prediction remains for all $x$ such that $d\left(x, x_{0}\right) \leq \epsilon$ where $d$ denotes a certain distance and $\epsilon>0$ is a small real number. Based on this intuition, a statistical robustness metric considers $p=P(g(X) \neq c)$, where $X$ follows a distribution concentrated around $x_{0}[14,13]$. Here $p$ represents the probability that the output is inconsistent with the baseline prediction at $x_{0}$.

In particular, when $g(\cdot)$ predicts using "score functions" $g_{i}(\cdot)$ with $i=1, . ., K$ where $K$ denotes
the number of categories, the predicted output is the category that has the maximum score, i.e. the prediction at $x$ is given by $\arg _{\max }^{i} g_{i}(x)$. Then we note that $g(x) \neq c$ is equivalent to $g_{c}(x) \leq$ $\max _{i \neq c} g_{i}(x)$. Hence we can transform $p$ into $P(\tilde{g}(X) \geq 0)$ by defining $\tilde{g}(x)=\max _{i \neq c} g_{i}(x)-g_{c}(x)$, which reduces to our problem statement presented earlier.

Example 2.2 (Risk Evaluation of Intelligent Physical Systems). Many intelligent physical systems (e.g. driver assistance systems) are built in a modular structure, which divides the overall task into sub-tasks that are handled by different modules. The perception module extracts information from the environment through various sensors (e.g. LIDAR [93], camera, etc.), which provides input for the downstream tasks [94]. Nowadays, perception modules are usually integrated with machine learning models, which play crucial roles in converting raw sensor data (e.g. images, point clouds) into information that are readable by downstream modules (e.g. object class, bounding box) [95].

Consider an intelligent physical system that embeds a machine learning predictor $g$ for perception (e.g. object detection). We then represent the decision of the system given an input $x$ as $h(g(x))$. The probability $P(h(g(X)) \in S)$, where $S$ represents a risky region, can be used to measure the risk of the system decision. For instance, suppose we are evaluating a collision avoidance system via the probability of a severe injury. Here, $x$ can represent the sensor data of a collision scenario, $h(g(x))$ the relative speed when collision happens, which proxies the severity of potential injuries, and the evaluation is equivalent to estimating $P(h(g(X)) \geq \gamma)$ for some speed threshold $\gamma$.

In most cases, $h$ is random by itself and can have a different complexity structure than the function class $g$. Our setup, which drops the general random $h$, can be viewed as a simplified probability $P(g(X) \geq \gamma)$ that provides a first step of study along this direction.

Example 2.3 (Probability Evaluation for Learned Rare-Event Set). When the system that drives the rare event is a black-box or too complicated to analyze [96], an approach to retain tractability is to approximate or learn the rare-event set via machine learning tools [21]. An example in operations research is the prediction of congestion risks in sophisticated queueing systems arising in, e.g., healthcare applications [97], where the queue could have multiple classes of customers
and complex priority rules [98] and the event of interest could be a transient probability of high occupancy level. In such settings, we can collect data or run simulations for $\{X, Y\}$, where $X$ denotes the random object in the considered system and $Y \in\{0,1\}$ denotes either the occurrence of the considered rare event or the numerical outcome under input $X$. Then we train neural network $g(\cdot)$ to classify the rare-event region given $X$. The learned rare-event set is then represented by $\{x: g(x) \geq \gamma\}$, where $\gamma$ is the threshold for classifying rare-event (e.g. $\gamma=0.5$ ) or the threshold for the outcome to trigger the event. As a result, $p=P(g(X) \geq \gamma)$ provides an approximation to the rare-event probability.

Example 2.4 (Validating Classification Models With Rare Categories). In classification model validation tasks, estimating the predictive performance of the test model can be costly if the test data requires human-annotation [99]. When we are interested in the performance on a rare category, the estimation of predictive metrics, e.g. F-scores (or F-measures) [100], becomes more challenging and hence requires more efficient approaches than naive sampling [101]. Consider that the input of the test classification model, denoted by $X$, has a fixed probability distribution across the population of samples. We use $y(X) \in\{1,0\}$ to denote the correct annotation at the input $X$ and $g(X)$ to denote the prediction given by the test classifier. Suppose we are interested in the prediction accuracy of the rare category $y(X)=1$ (i.e. $P(y(X)=1$ ) is extremely small). The $F_{\alpha}$-measure of the classification model is defined by:

$$
\begin{equation*}
F_{\alpha}=\frac{P(y(X)=1, g(X)=1)}{\alpha P(g(X)=1)+(1-\alpha) P(y(X)=1)}, \tag{2.1}
\end{equation*}
$$

with $\alpha \in(0,1)$. We observe that when $y(X)=1$ is a rare event and the classifier $g$ is well trained, all three probabilities in the $F_{\alpha}$-measure can be extremely small. Therefore, accurately estimating the $F_{\alpha}$-measure is closely related to estimating the rare-event probabilities $P(y(X)=1, g(X)=1)$ and $P(g(X)=1)$, which are defined via the test prediction model $g$.

Our setup described in the beginning of this section thus relates to the four emerging examples above. Though we could not resolve all the issues in these examples, notably with restrictions on
the input distribution and model complexity, we view our study as a first step towards a rigorous use of rare-event simulation techniques developed among the stochastic simulation community in the surging domain of safety and risk evaluation of AI-driven systems.

### 2.4 Efficient Importance Sampling via Sequential Mixed Integer Programming

We present our IS methodology. Section 2.4.1 reviews IS basics. Section 2.4.2 describes how we integrate the notions of dominating points and mixture IS with a sequential MIP algorithm. Section 2.4.3 presents our theoretical efficiency guarantees. The reformulation and solution to the MIP algorithms, which utilize recent developments in machine learning, are discussed in Section 2.5.

### 2.4.1 Basics of Importance Sampling

When $p$ is small, estimation using crude Monte Carlo is challenging since, intuitively, the samples have a low frequency of hitting the target set. This is statistically manifested as a large relative error. To be more specific, suppose that we use the crude Monte Carlo estimator $\hat{p}_{N}=$ $\frac{1}{N} \sum_{i=1}^{N} I\left(g\left(X_{i}\right) \geq \gamma\right)$ to estimate $p$. Since the probability $p$ is tiny, the error of the estimator should be measured relative to the size of $p$. In other words, we would like the probability of having a large relative error to be small, i.e., $P\left(\left|\hat{p}_{N}-p\right|>\varepsilon p\right) \leq \delta$ where $\delta$ is the confidence level and $0<\varepsilon<1$. By Markov's inequality, a sufficient condition for this is

$$
N \geq \frac{\operatorname{Var}(I(g(X) \geq \gamma))}{\delta \varepsilon^{2} E[I(g(X) \geq \gamma)]^{2}}=\frac{R E^{2}}{\delta \epsilon^{2}}
$$

where $R E=\sqrt{\operatorname{Var}(I(g(X) \geq \gamma))} / E[I(g(X) \geq \gamma)]$ is the relative error. For the crude Monte Carlo estimator, the RE is given by $\sqrt{(1-p) / p}$. That is, the simulation size $N$ has to be roughly proportional to $1 / p$ in order to achieve a given relative error. Under the settings that $X$ has a Gaussian distribution and $g$ is piecewise linear (see Corollary 2.1 ), $p$ is exponentially small in the threshold level $\gamma$, and hence the required simulation size would grow exponentially in $\gamma$.

A common approach to speed up simulation in such contexts is to use IS (see, e.g. the surveys [22,54, 63, 34, 23, 64], among others). Suppose $X$ has a density $f$. The basic idea of IS is to change the sampling distribution to say $\tilde{f}$, and output

$$
\begin{equation*}
Z=I(g(\tilde{X}) \geq \gamma) \frac{f(\tilde{X})}{\tilde{f}(\tilde{X})} \tag{2.2}
\end{equation*}
$$

where $\tilde{X}$ is sampled from $\tilde{f}$. This output is unbiased if $f$ is absolutely continuous with respect to $\tilde{f}$ over the rare-event set $\{x: g(x) \geq \gamma\}$ since

$$
\begin{aligned}
\tilde{E}[Z] & =\int_{\mathbb{R}} I(g(x) \geq \gamma) \frac{f(x)}{\tilde{f}(x)} \tilde{f}(x) d x=\int_{\mathbb{R}} I(g(x) \geq \gamma) f(x) d x \\
& =E[I(g(X) \geq \gamma)]=P(g(X) \geq \gamma)
\end{aligned}
$$

By choosing $\tilde{f}$ appropriately, one can substantially reduce the simulation variance.
In order to measure the efficiency of an IS scheme, we introduce a rarity parameter, say $\gamma$, that parameterizes the rare-event probability $p_{\gamma}$ such that $p_{\gamma} \rightarrow 0$ as $\gamma \rightarrow \infty$. As discussed before, since the probability of interest is small, one should focus on the relative error of the Monte Carlo estimator with respect to the magnitude of this probability. To this end, we call an IS estimator $Z_{\gamma}$ for $p_{\gamma}$ asymptotically optimal [54, 23] if

$$
\begin{equation*}
\lim _{\gamma \rightarrow \infty} \frac{\log \tilde{E}\left[Z_{\gamma}^{2}\right]}{\log \tilde{E}\left[Z_{\gamma}\right]}=2 \tag{2.3}
\end{equation*}
$$

where $\tilde{E}$ denotes the expectation with regard to $\tilde{f}$. The notion (2.3) is equivalent to saying that $\tilde{E}\left[Z_{\gamma}^{2}\right]$ and $\tilde{E}\left[Z_{\gamma}\right]^{2}$ grow in the same exponential rate in $\gamma$. This ensures that the second moment, or the variance, does not explode exponentially relative to the probability of interest as $\gamma$ increases, thus preventing an exponentially large number of simulation replications to achieve a given relative accuracy. We will use asymptotic optimality as our efficiency criterion in this chapter. Moreover, in the large deviations settings where $p_{\gamma}=\tilde{E}\left[Z_{\gamma}\right]$ decays exponentially in $\gamma, \tilde{E}\left[Z_{\gamma}^{2}\right] / \tilde{E}\left[Z_{\gamma}\right]^{2}$ at most growing polynomially in $\gamma$ is a sufficient condition for asymptotic optimality.

Another commonly used efficiency criterion is the bounded relative error, which is defined as

$$
\limsup _{\gamma \rightarrow \infty} \frac{\tilde{E}\left[Z_{\gamma}^{2}\right]}{\tilde{E}\left[Z_{\gamma}\right]^{2}}<\infty .
$$

This is a stronger condition than asymptotic optimality. More efficiency criteria can be found in [23, 102].

### 2.4.2 Dominating Points and Mixture Importance Samplers

In the case of Gaussian input distributions, finding a good $\tilde{f}$ is particularly handy and one approach to devise good IS distributions uses the notion of so-called dominating point. As explained in the introduction, a dominating point can be understood as the highest-density point in the rareevent set that satisfies some conditions. More precisely, the collection of dominating points for a rare-event set with Gaussian distributed input is defined in Definition 2.1.

Definition 2.1. Suppose that $S \subset \mathbb{R}^{d}$ is a rare-event set. Suppose that a set $A \subset \mathbb{R}^{d}$ satisfies that $S \subset \bigcup_{a \in A}\left\{x:(a-\mu)^{T} \Sigma^{-1}(x-a) \geq 0\right\}$ and that $a=\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): x \in\right.$ $S$ and $\left.(a-\mu)^{T} \Sigma^{-1}(x-a) \geq 0\right\}$ for any $a \in A$. Moreover, suppose that the above conditions do not hold anymore if we remove any element from $A$. Then the points in $A$ are called the dominating points of $S$ with input distribution $N(\mu, \Sigma)$.

Note that minimizing $(x-\mu)^{T} \Sigma^{-1}(x-\mu)$ is equivalent to maximizing $\phi(x ; \mu, \Sigma)$, the Gaussian density with mean $\mu$ and covariance $\Sigma$. The condition $2(a-\mu)^{T} \Sigma^{-1}(x-a) \geq 0$ is the first-order condition of optimality for the optimization $\min _{x}(x-\mu)^{T} \Sigma^{-1}(x-\mu)$ over a convex set for $x$. Thus, intuitively, each dominating point in the collection $A$ can be viewed as the highest-density point in a "local" region formed by $S \cap\left\{x:(a-\mu)^{T} \Sigma^{-1}(x-a) \geq 0\right\}$. Figure 2.1 is an illustration of the dominating points. In particular, if $\{x: g(x) \geq \gamma\}$ is a convex set, then there is only one dominating point $a$. In this case, a well-known IS scheme is to use a Gaussian distribution $N(a, \Sigma)$ as the IS distribution $\tilde{f}$.

We explain intuitively why we need more than one dominating point (the highest-density point


Figure 2.1: Illustration of the dominating points. $a_{1}$ is the globally highest-density point in the rare-event set $S$, but the halfspace $\left\{x:\left(a_{1}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{1}\right) \geq 0\right\}$ does not fully cover $S$, so an additional point $a_{2}$ is needed to comprise a dominating set.
over $S$ ) and the pitfall if we omit the other ones in constructing efficient IS. Suppose that the rare-event set consists of two disconnected convex components which are nearly equi-distant with respect to the origin, and we choose the IS distribution to be centered at the dominating point of one component. Then, if a sample from the IS distribution hits the other component, a scenario that could be unlikely but possible, the resulting likelihood ratio, which now contributes to the output as the rare-event set is hit, could possibly be tremendous. This ultimately leads to an explosion of the relative error in the IS estimator. [103] presents more counterexamples which show that it is essential to find all the dominating points in constructing an efficient IS based on mixtures.

In view of the aforementioned discussions, we consider the following IS scheme. If we can split $\{x: g(x) \geq \gamma\}$ into $\mathcal{R}_{1}, \ldots, \mathcal{R}_{r}$, and for each $\mathcal{R}_{i}, i=1, \ldots, r$ there exists a dominating point $a_{i}$ such that $a_{i}=\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): x \in \mathcal{R}_{i}\right\}$ and $\mathcal{R}_{i} \subseteq\left\{x:\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{i}\right) \geq 0\right\}$, then we use a Gaussian mixture distribution with $r$ components as the IS distribution $\tilde{f}$, where the $i$ th component has mean $a_{i}$. This proposal guarantees the asymptotic optimality of the IS (see Theorem 2.1).

In our task, because the machine learning predictor $g(x)$ is nonlinear and $x$ is high-dimensional in general, splitting $\{x: g(x) \geq \gamma\}$ into $\mathcal{R}_{1}, \ldots, \mathcal{R}_{r}$ that have dominating points is challenging even with known parameters. This challenge motivates us to use Algorithm 2.1 to obtain the dominating
points $a_{1}, \ldots, a_{r}$ that constructs an efficient IS distribution. The procedure uses a sequential "cutting plane" approach to exhaustively look for all dominating points, by reducing the search space at each iteration via taking away the regions covered by found dominating points. The set $A$ in the procedure serves to store the dominating points we have located throughout the procedure. At the end of the procedure, we obtain a set $A$ that contains all the dominating points $a_{1}, \ldots, a_{r}$. Note that when $g(x) \geq \gamma$ is convex, the algorithm solves a series of convex quadratic programming problems, and it is well known that such problems could be solved efficiently in polynomial time (see [104] for more details on the complexity). In this chapter, we focus on the problems with piecewise linear $g(x)$, which leads to mixed integer convex quadratic optimization problems as shown in later discussion. Although a mixed integer quadratic optimization is NP-hard, we can solve it much more efficiently using specialized algorithms than general nonlinear MIPs [105].

Algorithm 2.1: Procedure to find all dominating points for the set $\{x: g(x) \geq \gamma\}$.
Input: Prediction model $g(x)$, threshold $\gamma$, input distribution $N(\mu, \Sigma)$.
Output: dominating point set $A$.
Start with $A=\emptyset$;
2 While $\left\{x: g(x) \geq \gamma,\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{i}\right)<0, \forall a_{i} \in A\right\} \neq \emptyset$ do
3 Find a dominating point $a$ by solving the optimization problem

$$
\begin{align*}
a=\arg \min _{x} & (x-\mu)^{T} \Sigma^{-1}(x-\mu)  \tag{2.4}\\
\text { s.t. } & g(x) \geq \gamma \\
& \left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{i}\right)<0, \forall a_{i} \in A
\end{align*}
$$

and update $A \leftarrow A \cup\{a\} ;$

## 4 End

Algorithm 2.1 gives $A=\left\{a_{1}, \ldots, a_{r}\right\}$. With this, we split $\{x: g(x) \geq \gamma\}$ into $\mathcal{R}_{1}, \ldots, \mathcal{R}_{r}$ where $\mathcal{R}_{i}=\left\{x: g(x) \geq \gamma,\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{i}\right) \geq 0,\left(a_{j}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{j}\right) \leq 0, \forall j<i\right\}$. Clearly $a_{i}=\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): x \in \mathcal{R}_{i}\right\}$ and $\left(a_{1}-\mu\right)^{T} \Sigma^{-1}\left(a_{1}-\mu\right) \leq \cdots \leq\left(a_{r}-\mu\right)^{T} \Sigma^{-1}\left(a_{r}-\mu\right)$. Moreover, we note that $\left(a_{1}-\mu\right)^{T} \Sigma^{-1}\left(a_{1}-\mu\right)=\min _{i=1, \ldots, r}\left\{\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(a_{i}-\mu\right)\right\}$.

Given the dominating point set $A$, we use a mixture distribution with density

$$
\tilde{f}(x)=\frac{1}{r} \sum_{i=1}^{r} \phi\left(x ; a_{i}, \Sigma\right)
$$

as the IS distribution. That is, the IS estimator is

$$
\begin{equation*}
Z=I(g(\tilde{X}) \geq \gamma) L(\tilde{X}) \tag{2.5}
\end{equation*}
$$

where $\tilde{X} \sim \tilde{f}$ and $L$, the likelihood ratio, is defined as

$$
L(x)=\frac{f(x)}{\tilde{f}(x)}=\frac{r e^{-\frac{1}{2}(x-\mu)^{T} \Sigma^{-1}(x-\mu)}}{e^{-\frac{1}{2}\left(x-a_{1}\right)^{T} \Sigma^{-1}\left(x-a_{1}\right)}+\cdots+e^{-\frac{1}{2}\left(x-a_{r}\right)^{T} \Sigma^{-1}\left(x-a_{r}\right)}} .
$$

Note that we have used uniform mixture weights in our IS distribution depicted above. These weights could potentially be tuned more carefully rather than simply equally assigned to further improve the efficiency, especially when an asymptotic zero-variance distribution is available (as in, e.g., $[106,107])$, though here we are contented with uniform weights and do not refine further. To sum up, after computing the dominating points $A=\left\{a_{1}, \ldots, a_{r}\right\}$ using Algorithm 2.1, we estimate the probability of interest via Algorithm 2.2.

## Algorithm 2.2: Construct the IS estimator with all the dominating points.

Input: Prediction model $g(x)$, threshold $\gamma$, dominating points $A=\left\{a_{1}, \ldots, a_{r}\right\}$, simulation size $N$.
Output: Estimated rare-event probability $\hat{p}$.
1 Generate $\tilde{X}_{1}, \ldots, \tilde{X}_{N} \sim \tilde{f}(x)=\frac{1}{r} \sum_{i=1}^{r} \phi\left(x ; a_{i}, \Sigma\right)$;
2 Compute $\hat{p}=\frac{1}{N} \sum_{i=1}^{N} I\left(g\left(\tilde{X}_{i}\right) \geq \gamma\right) L\left(\tilde{X}_{i}\right)$ where

$$
L(x)=\frac{r e^{-\frac{1}{2}(x-\mu)^{T} \Sigma^{-1}(x-\mu)}}{e^{-\frac{1}{2}\left(x-a_{1}\right)^{T} \Sigma^{-1}\left(x-a_{1}\right)}+\cdots+e^{-\frac{1}{2}\left(x-a_{r}\right)^{T} \Sigma^{-1}\left(x-a_{r}\right)}}
$$

3 End

### 2.4.3 Efficiency Guarantees

The efficiency guarantee of the proposed IS estimator (2.5) is given by:
Theorem 2.1. Suppose that the input $X \sim N(\mu, \Sigma)$ and the prediction model $g(\cdot)$ is a piecewise linear function (with finite pieces) such that $P(g(X) \geq \gamma)>0$ for any $\gamma \in \mathbb{R}$. The IS estimator $Z$ is defined in (2.5). Then we have that $\tilde{E}\left[Z^{2}\right] / \tilde{E}[Z]^{2}$ is at most polynomially growing in $\gamma$. Moreover, $Z$ is asymptotically optimal.

Theorem 2.1 is proved by constructing an upper bound for the relative error, which in turn depends on the asymptotic approximation of probability on polytope sets using dominating points. Our proof leverages the results in [62] on the tail exceedance asymptotic of $P\left(N\left(0, \Sigma_{n}\right) \geq t_{n}\right)$ where $\left\|t_{n}\right\| \rightarrow \infty$ as $n \rightarrow \infty$, but requires substantial generalization. Note that Theorem 2.1 only makes the very general assumptions that $g$ is piecewise linear and the probability $P(g(X) \geq \gamma)$ is nondegenerate (i.e., non-zero) for any $\gamma \in \mathbb{R}$. Our result applies to, for example, the probability $P(A X \geq t)$ where $A$ is a constant matrix and $t-\gamma e_{1}$ is a constant vector (here, $\left.e_{1}=(1,0, \ldots, 0)^{T}\right)$. If $A A^{T}$ is not invertible, then it is not easily reducible to the setting studied in [62]. To achieve a general result, we carefully construct a superset and a subset of the rare-event set to derive tight enough upper and lower bounds for the probability of interest, in which we analyze the involved asymptotic integrals instead of using the conditional probability representation in [62] that is not directly applicable in our setting. For the detailed proof, please refer to Section 2.7.

A by-product in proving Theorem 2.1 is the large deviations probability asymptotic for $P(g(X) \geq$ $\gamma)$ :

Corollary 2.1. Suppose that the input $X \sim N(\mu, \Sigma)$ and the prediction model $g(\cdot)$ is a piecewise linear function (with finite pieces) such that $P(g(X) \geq \gamma)>0$ for any $\gamma \in \mathbb{R}$. Denote $a=$ $\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): g(x) \geq \gamma\right\}$. Then $-\log P(g(X) \geq \gamma)=(1+o(1))(a-\mu)^{T} \Sigma^{-1}(a-\mu) / 2$ as $\gamma \rightarrow \infty$. In particular, $P(g(X) \geq \gamma)$ is exponentially small in $\gamma$.

The theoretical guarantee given by Theorem 2.1 justifies the sequential MIP algorithm for searching dominating points. The resulting mixture IS distribution is asymptotically optimal. We
point out some related works that use mixture distributions that are related to our proposed method. In [65], mixture IS distributions are constructed based on separating rare-event set with half-spaces. However, in this work, the rare-event set is restricted to be a union of half-spaces, and these halfspaces are assumed to be known. The use of Algorithm 2.1 allows us to deal with more general rare-event sets. Moreover, in relation to Corollary 2.1, we also mention the work [60] that derives an asymptotic result for Gaussian probabilities using dominating points. However, they focus on convex hitting sets where the entire set is scaled with a rarity parameter, which is different from our settings. First, our rare-event set is not necessarily convex. Second, even if we separate our rare-event set into the union of convex sets, their results still cannot be applied, since in our settings some linear constraints are allowed to be fixed instead of scaling with $\gamma$.

Finally, we close this section by noting that the proposed IS scheme can be extended to problems with Gaussian mixture inputs. Suppose the Gaussian mixture has $m$ components, so that $X \sim \sum_{j=1}^{m} \pi_{j} \phi\left(x ; \mu_{j}, \Sigma_{j}\right)$. For each component $j$, we implement Algorithm 2.1 with input distribution $N\left(\mu_{j}, \Sigma_{j}\right)$ to obtain dominating point set $A_{j}$ (with cardinality $r_{j}$ ). The proposed IS distribution is given by $\tilde{f}(x)=\sum_{j=1}^{m} \sum_{i=1}^{r_{j}} \pi_{j} / r_{j} \phi\left(x ; a_{j i}, \Sigma_{j}\right)$. We summarize the procedure as Algorithm 2.3.

## Algorithm 2.3: Procedure for Gaussian mixture distributed input.

Input: Prediction model $g(x)$, threshold $\gamma$, input distribution $\sum_{j=1}^{m} \pi_{j} \phi\left(x ; \mu_{j}, \Sigma_{j}\right)$, simulation size $N$.
Output: Estimated rare-event probability $\hat{p}$.
1 Implement Algorithm 2.1 with input distribution $N\left(\mu_{j}, \Sigma_{j}\right)$ to get $A_{j}=\left\{a_{j 1}, \ldots, a_{j r_{j}}\right\}$;
2 Generate $\tilde{X}_{1}, \ldots, \tilde{X}_{N} \sim \tilde{f}(x)=\sum_{j=1}^{m} \sum_{i=1}^{r_{j}} \pi_{j} / r_{j} \phi\left(x ; a_{j i}, \Sigma_{j}\right)$;
3 Compute $\hat{p}=\frac{1}{N} \sum_{i=1}^{N} I\left(g\left(\tilde{X}_{i}\right) \geq \gamma\right) L\left(\tilde{X}_{i}\right)$ where

$$
\begin{equation*}
L(x)=\frac{\sum_{j=1}^{m} \pi_{j}\left|\Sigma_{j}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(x-\mu_{j}\right)^{T} \Sigma_{j}^{-1}\left(x-\mu_{j}\right)}}{\sum_{j=1}^{m} \sum_{i=1}^{r_{j}} \pi_{j} / r_{j}\left|\Sigma_{j}\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\left(x-a_{j i}\right)^{T} \Sigma_{j}^{-1}\left(x-a_{j i}\right)}} \tag{2.6}
\end{equation*}
$$

## 4 End

Similar to Algorithm 2.2, we have the efficiency guarantee for Algorithm 2.3:

Corollary 2.2. Suppose that the input $X \sim \sum_{j=1}^{m} \pi_{j} \phi\left(x ; \mu_{j}, \Sigma_{j}\right)$ and the prediction model $g(\cdot)$ is a piecewise linear function (with finite pieces) such that $P(g(X) \geq \gamma)>0$ for any $\gamma \in \mathbb{R}$. The

IS estimator $Z$ is defined as $I(g(\tilde{X}) \geq \gamma) L(\tilde{X})$ where $\tilde{X} \sim \sum_{j=1}^{m} \sum_{i=1}^{r_{j}} \pi_{j} / r_{j} \phi\left(x ; a_{j i}, \Sigma_{j}\right)$ and $L(x)$ is as defined in (2.6). Then we have that $\tilde{E}\left[Z^{2}\right] / \tilde{E}[Z]^{2}$ is at most polynomially growing in $\gamma$. Moreovers, $Z$ is asymptotically optimal.

When we apply Algorithm 2.1 to find all dominating points, the key is to be able to solve the optimization problems in (2.4). We will investigate this in the next section.

### 2.5 Tractable Optimization Formulation for Prediction Models

We discuss how to formulate optimization problems in Algorithm 2.1 as an MIP with quadratic objective function and linear constraints, for random forest (Section 2.5.1) and neural network (2.5.2) structures.

### 2.5.1 Tractable Formulation for Random Forest

A random forest [55, 108] can be specified as follows. Given a set of $T$ decision trees $g_{1}, \ldots, g_{T}$ with $d$ dimensional input $x$, a random forest $g$ ensembles these trees by weightedly averaging their outputs, namely $g=\sum_{t=1}^{T} \lambda_{t} g_{t}$, where $\lambda_{t}$ denotes the weight of tree $t\left(\sum_{t=1}^{T} \lambda_{t}=1\right)$.


Figure 2.2: An example of a decision tree.

As illustrated in Figure 2.2, a decision tree consists of nodes and a branch structure. The nodes are categorized into splits (triangle node), the nodes with two child nodes, and leaves (circle node), the nodes with no child node. At each split, we execute a binary query defined by a dimension index and a split point, i.e., in the form of $x_{i} \leq a$, where $x_{i}$ denotes the $i$ th dimension of the input $x$ and $a \in \mathbb{R}$ is the split point. Starting from the root node, a sequence of queries leads the input down to a leaf node which corresponds to an output value.

To look for dominating points in a random forest, we follow the route in [59] that studies optimization over these models. Following the notations therein, we use $a_{i, j}$ to summarize the split point information from all trees in $g$, which denotes the $j$ th unique split point for the $i$ th dimension of the input $x$. Note that $a_{i, 1}<a_{i, 2}<\ldots<a_{i, K_{i}}$, where $K_{i}$ is the number of unique split points for the $i$ th dimension of $x$.

To represent the branch structure, we define leaves $(t)$ as the set of leaves (terminal nodes) of tree $t$ and $\operatorname{splits}(t)$ as the set of splits (non-terminal nodes) of tree $t$. In each split $s$, we let left $(s)$ be the set of leaves that are accessible from the left branch (the query at $s$ is true), and $\boldsymbol{\operatorname { r i g h t }}(s)$ be the set of leaves that are accessible from the right branch (the query at $s$ is false). For each node $s$, we use $\mathbf{V}(s) \in\{1, \ldots, d\}$ to denote the dimension that participates in the node and $\mathbf{C}(s) \in\left\{1, \ldots, K_{\mathbf{V}(s)}\right\}$ to denote the index of the split point on dimension $i$ that participates in the query of $s\left(\mathbf{V}(s)=i\right.$ and $\mathbf{C}(s)=j$ indicate the query $\left.x_{i} \leq a_{i, j}\right)$. For each $l \in \operatorname{leaves}(t), p_{t, l}$ denotes the output for the $l$ th leaf in tree $t$.

To formulate the random forest optimization as an MIP, we introduce decision variables $z_{i, j}$ and $y_{t, l}$. Firstly, we use $z_{i, j}$ to locate the input $x$ by linking its value to the split points $a_{i, j}$ 's, where we have

$$
\begin{equation*}
z_{i, j}=I\left(x_{i} \leq a_{i, j}\right), i=1, \ldots, d, j=1, \ldots, K_{i} . \tag{2.7}
\end{equation*}
$$

In order to convert (2.7) into mixed integer constraints, we introduce an arbitrary large number $B \in \mathbb{R}^{+}$which serves as the big- $M$ coefficient [109] in our formulation. For any given problem, all dominating points must have finite coordinates. This implies that for large enough $B$ we have $[-B, B]^{d}$ contain all dominating points. Thus, assuming we use a large enough $B$, we can let $x \in[-B, B]^{d}$ and $\left|a_{i, j}\right| \leq B$. Then (2.7) is represented by the following constraints:

$$
\begin{aligned}
& x_{i} \leq a_{i, j}+2\left(1-z_{i, j}\right) B \\
& x_{i}>a_{i, j}-2 z_{i, j} B \\
& z_{i, j}=\{0,1\} .
\end{aligned}
$$

Next we use $y_{t, l}=1$ to denote that tree $t$ outputs the prediction value $p_{t, l}$ on leaf $l$, and $y_{t, l}=0$ otherwise. This allows us to represent the output of the random forest as

$$
\sum_{t=1}^{T} \sum_{l \in \operatorname{leaves}(t)} \lambda_{t} p_{t, l} y_{t, l}
$$

with $\sum_{l \in \operatorname{leaves}(t)} y_{t, l}=1$. We use $\mathbf{z}, \mathbf{y}$ to represent the vectors of $z_{i, j}$ and $y_{t, l}$ respectively.
Lastly, we formulate the binary queries in a decision tree with these intermediate variables. This is achieved by forcing $y_{t, l}$ in the "unselected" branches to be 0 . At each split $s$, we have

$$
\begin{aligned}
& x_{\mathbf{V}(s)}>a_{\mathbf{V}(s), \mathbf{C}(s)} \Rightarrow \sum_{l \in \operatorname{left}(s)} y_{t, l}=0 \\
& x_{\mathbf{V}(s)} \leq a_{\mathbf{V}(s), \mathbf{C}(s)} \Rightarrow \sum_{l \in \operatorname{right}(s)} y_{t, l}=0
\end{aligned}
$$

which we reformulate with $\mathbf{z}$ into

$$
\begin{aligned}
& \sum_{l \in \mathbf{l e f t}(s)} y_{t, l} \leq z \mathbf{v}(s), \mathbf{C}(s), \forall t \in\{1, \ldots, T\}, s \in \operatorname{splits}(t) \\
& \sum_{l \in \operatorname{right}(s)} y_{t, l} \leq 1-z \mathbf{v}(s), \mathbf{C}(s), \forall t \in\{1, \ldots, T\}, s \in \mathbf{s p l i t s}(t) .
\end{aligned}
$$

Now we formulate (2.4) with $A=\emptyset$ as the following MIP

$$
\begin{align*}
\min _{x, \mathbf{y}, \mathbf{Z}} & (x-\mu)^{T} \Sigma^{-1}(x-\mu)  \tag{2.8}\\
\text { s.t. } & \sum_{i=1}^{T} \sum_{l \in \operatorname{leaves}(t)} \lambda_{t} p_{t, l} y_{t, l} \geq \gamma \\
& \sum_{l \in \operatorname{leaves}(t)} y_{t, l}=1, \forall t \in\{1, \ldots, T\} \\
& \sum_{l \in \operatorname{left}(s)} y_{t, l} \leq z \mathbf{V}(s), \mathbf{C}(s), \forall t \in\{1, \ldots, T\}, s \in \operatorname{splits}(t) \\
& \sum_{l \in \operatorname{right}(s)} y_{t, l} \leq 1-z \mathbf{V}(s), \mathbf{C}(s), \forall t \in\{1, \ldots, T\}, s \in \mathbf{s p l i t s}(t) \\
& z_{i, j} \leq z_{i, j+1}, \forall i \in\{1, \ldots, d\}, j \in\left\{1, \ldots, K_{i}-1\right\} \\
& z_{i, j} \in\{0,1\}, \forall i \in\{1, \ldots, d\}, j \in\left\{1, \ldots, K_{i}\right\} \\
& y_{t, l} \geq 0, \forall t \in\{1, \ldots, T\}, l \in \operatorname{leaves}(t) \\
& x_{i} \leq a_{i, j}+2\left(1-z_{i, j}\right) B, \forall i \in\{1, \ldots, d\}, j \in\left\{1, \ldots, K_{i}\right\} \\
& x_{i}>a_{i, j}-2 z_{i, j} B, \forall i \in\{1, \ldots, d\}, j \in\left\{1, \ldots, K_{i}\right\} .
\end{align*}
$$

This formulation has a quadratic objective function and linear constraints. Similarly, we can formulate (2.4) with $A \neq \emptyset$ by adding linear constraints $\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{i}\right)<0, \forall a_{i} \in A$ to (2.8). Note that both the number of decision variables and the number of constraints are linearly dependent on the total number of nodes in the random forest.

### 2.5.2 Tractable Formulation for Neural Network

A neural network $g(\cdot)$ is a network that connects a large number of computational units known as neurons $[56,110]$. Depending on the task, this network bears a specific architecture that usually involves multiple layers of neurons and different operations over the neurons. For simplification, here we consider layers with consecutive architecture and each layer of the neural network only contains one specific structure.

The key part of the reformulation is to deal with the non-linearity brought by the maximum function. Our treatment of the maximum function follows from [58], which rewrites neural network structures into linear equations with binary variables.

In order to obtain tractable formulation for the constraint $g(x) \geq \gamma$, we independently handle each single layer in $g(\cdot)$. Assume we have $l$ layers in $g(\cdot)$, where $g_{i}(\cdot)$ denotes the $i$ th layer. Given input $x$, the output of the neural network can be represented as $g(x)=g_{l}\left(g_{l-1}\left(\ldots g_{1}(x)\right)\right)$. For convenience, we introduce $x_{i}$ to denote the output of the $i$ th layer (note that it is also the input for the $i+1$ th layer). In other words, for the $i$ th layer we have $x_{i}=g_{i}\left(x_{k-1}\right)$. Using these notations, we can transform the constraint $g(x) \geq \gamma$ into a sequence of constraints:

$$
\begin{aligned}
& x_{l} \geq \gamma, \\
& x_{l}=g_{l}\left(x_{l-1}\right), \\
& x_{l-1}=g_{l-1}\left(x_{l-2}\right), \\
& \ldots, \\
& x_{1}=g_{1}(x) .
\end{aligned}
$$

This transformation makes clear that the constraints altogether are tractable if the constraint for each layer (i.e. $x_{i}=g_{i}\left(x_{i-1}\right)$ ) is tractable . Note that both the number of decision variables and the number of constraints are linearly dependent on the total number of neurons in the neural network. In the rest of this section, we discuss the reformulation of neural network layers concerning different structures.

## Fully Connected Layer

In a fully connected layer, each neuron performs a linear transformation on the input. We consider a layer with $n$ neurons and the input for this layer is a vector $x \in \mathbb{R}^{m}$. We use $w_{i} \in \mathbb{R}^{m}$ and $b_{i} \in \mathbb{R}$ to denote the weight and bias respectively for the linear transformation in the $i$ th neuron. Then the output of the $i$ th neuron can be represented by $y_{i}=w_{i}^{T} x+b_{i}$. To summarize, the output
of the layer, $y=\left[y_{1}, y_{2}, \ldots, y_{n}\right] \in \mathbb{R}^{n}$, is given by

$$
y=W^{T} x+b
$$

where $W=\left[w_{1}, w_{2}, \ldots, w_{n}\right]$ and $b=\left[b_{1}, b_{2}, \ldots, b_{n}\right]$.

## ReLU Layer

In a rectified linear unit (ReLU) layer, negative elements in the input are replaced by 0's. For the $i$ th input, the output is given by $y_{i}=\max \left\{x_{i}, 0\right\}$. This can be represented by

$$
\begin{aligned}
& y_{i} \leq x_{i}-l\left(1-z_{i}\right), \\
& y_{i} \geq x_{i}, \\
& y_{i} \leq u z_{i}, \\
& y_{i} \geq 0, \\
& z_{i} \in\{0,1\},
\end{aligned}
$$

where $z_{i} \in\{0,1\}$ is a binary variable, $u$ and $l$ are the upper and lower bounds of the input respectively.

## Normalization Layer

In a normalization layer, the input is normalized and linearly transformed to make the gradient descent algorithm more efficient. Again we assume the input is $x \in \mathbb{R}^{m}$ with a given normalization parameter $\mu \in \mathbb{R}^{m}$ and $\Sigma \in \mathbb{R}^{m \times m}$. Moreover, we have the transformation matrix $\gamma \in \mathbb{R}^{m \times m}$ and bias vector $\beta \in \mathbb{R}^{m}$. The output is given by

$$
y=\gamma\left(\Sigma^{-1 / 2}(x-\mu)\right)+\beta
$$

## Pooling Layer

In a pooling layer, a "filter" that can be applied to adjacent elements in a vector or matrix goes through the input with a certain stride. Such type of layer is used to summarize "local" information and reduce the dimension of the input. Max pooling and average pooling are two types of commonly used filters.

Suppose the input is represented by matrix $x \in \mathbb{R}^{m_{1} \times m_{2}}$, where $x_{i j}$ denotes the element in the $i$ th row $j$ th column. The size of the filter is $s_{1} \times s_{2}$ with stride $\left(s_{1}, s_{2}\right)$. The output has size $y \in \mathbb{R}^{n_{1}, n_{2}}$, where $n_{1}=m_{1} / s_{1}$ and $n_{2}=m_{2} / s_{2}$. We assume that the value of $s_{1}, s_{2}$ are carefully chosen so that $n_{1}$ and $n_{2}$ are integers.

For average pooling layer, we have

$$
y_{i j}=\frac{\sum_{r=(i-1) s_{1}+1}^{i s_{1}} \sum_{c=(j-1) s_{2}+1}^{j s_{2}} x_{r c}}{s_{1} s_{2}}
$$

for $i=1, \ldots, n_{1}, j=1, \ldots, n_{2}$.
For max pooling layer, we have $y_{i j}=\max _{(r, c) \in S_{i j}} x_{r c}$ for $i=1, \ldots, n_{1}, j=1, \ldots, n_{2}$, where $S_{i j}=\left\{(r, c) \mid r=(i-1) s_{1}+1, \ldots, i s_{1}, c=(j-1) s_{2}+1, \ldots, j s_{2}\right\}$. The tractable formulation is given by

$$
\begin{array}{lr}
y_{i j} \leq x_{r c}-(u-l)\left(1-z_{r c}\right), & (r, c) \in S_{i j} \\
y_{i j} \geq x_{r c}, & (r, c) \in S_{i j} \\
\sum_{(r, c) \in S_{i j}} z_{r c}=1 & \\
z_{r c} \in\{0,1\}, & (r, c) \in S_{i j} .
\end{array}
$$

## Convolutional Layer

In a convolutional layer, several filters are used to extract features from the input. The input of the layer is $x \in \mathbb{R}^{m_{1} m_{2}}$. Suppose we have $r$ filters and assume the filters have size $s_{1} \times s_{2}$ with stride
$\left(t_{1}, t_{2}\right)$. We use $w_{i} \in \mathbb{R}^{t_{1} t_{2}}$ and $b_{i} \in \mathbb{R}^{t_{1} t_{2}}$ to denote the weight and bias for the $i$ th filter. The output is $y \in \mathbb{R}^{n_{1} \times n_{2} \times r}$, where $n_{1}=\left(m_{1}-s_{1}\right) / t_{1}$ and $n_{2}=\left(m_{2}-s_{2}\right) / t_{2}$. Again we assume the numbers are carefully chosen so that $n_{1}, n_{2}$ are integers.

Then we have

$$
\begin{aligned}
& y_{i j k}=w_{k}^{T}\left(\tilde{x}_{i j}\right)+b_{k}, \\
& \tilde{x}_{i j}=\left[x_{(i-1) t_{1}+1,(j-1) t_{2}+1}, x_{(i-1) t_{1}+2,(j-1) t_{2}+1}, \ldots, x_{(i-1) t_{1}+1,(j-1) t_{2}+2}, \ldots, x_{(i-1) t_{1}+s_{1},(j-1) t_{2}+s_{2}}\right] .
\end{aligned}
$$

for integers $1 \leq i \leq n_{1}, 1 \leq j \leq n_{2}$ and $1 \leq k \leq r$.

## Reformulation in the Output Layer

Here we discuss the reformulation of the output layer, which also provides us clues on how other more general problems in classification tasks are potentially transformable into the constraint $g(x) \geq \gamma$. Although the output layer is usually highly nonlinear, we show how to formulate it as linear mixed-integer constraints.

In classification tasks, the neural network usually uses a softmax layer as the output layer for training purposes. Suppose the classification problem has $n$ categories in total, the last layer inputs $x \in \mathbb{R}^{n}$ and outputs $y \in \mathbb{R}^{n}$ with $y_{i}=\frac{e^{x_{i}}}{\sum_{j=1}^{n} e^{x_{j}}}$. The prediction for classification is determined by the maximum value of $y_{i}$. Indeed, the result is equivalent if we determine the categories by the maximum value of $x_{i}$.

When the constraint is $g(X)=i$ or $g(X) \neq i$, we can use this equivalence to reformulate the last layer (and therefore complete the formulation for the whole network). Specifically, $g(X)=i$ can be formulated as $x_{i} \geq x_{j}$, for $j \neq i$ and $g(X) \neq i$ can be formulated as $x_{i} \leq \max _{j \neq i} x_{j}$, where $j \neq i$ denotes $j$ is an element for the set that contains all possible indexes except $i$. For tractable
form, the latter formula can be further rewritten as:

$$
\begin{aligned}
& x_{i} \leq x_{j}+\left(1-z_{j}\right)(u-l), j \neq i . \\
& \sum_{j \neq i} z_{j} \geq 1, \\
& z_{j} \in\{0,1\}, j \neq i
\end{aligned}
$$

### 2.6 Experiments

This section presents several experimental results using our Algorithm 2.1 for neural network and random forest predictors. In Section 2.6.1, we consider two simple toy examples. The first problem has one dominating point and the second problem has multiple dominating points. To illustrate the efficiency of the proposed IS scheme, we compare it with a naive IS scheme using uniform distribution. In Section 2.6.2, we consider a realistic problem generated from a classification data set with a high-dimensional feature space.

### 2.6.1 Toy Problems

We consider the rare-event set $\{x: g(x) \geq \gamma\}$ and the input $X$ follows a Gaussian distribution $N\left(0, I \sigma^{2}\right)$, where $I$ denotes the identity matrix and $\sigma^{2} \in \mathbb{R}^{+}$. The prediction model $g$ is trained with a data set with uniformly designed inputs, and labeled using a deterministic function denoted by $y$. In order to build a prediction model with reasonable quality, the inputs of the training data are generated from a bounded region $[l, u]^{d}$, where the region is chosen sufficiently large in terms of $\gamma$ that setting $g(x)$ to $-\infty$ outside the region barely affects the target probability. As a result, whether we impose this bound or not does not affect the probability materially, and we choose to impose it since this setting provides a good and simple IS scheme (i.e., uniform distribution) for comparison.

Given the above setting, we consider a uniform IS scheme as a baseline method in our experiments. Consider a problem where $X$ follows a distribution $f(x)$, and the set $\{x: g(x) \geq \gamma\}$
is known to lie inside $[l, u]^{d}$ where $d$ is the dimension of the input variable $X$. The uniform IS estimator of $P(g(x) \geq \gamma)$ is given by

$$
Z_{\text {uniform }}=I(g(X) \geq \gamma) f(X)(u-l)^{d},
$$

where $X$ is generated from a uniform distribution on $[l, u]^{d}$. This estimator has a polynomially growing relative efficiency as the magnitude of the dominating points grows [111], but the efficiency also depends significantly on the size of the bounded set, i.e., $l, u, d$.

In the first example, we use the deterministic function

$$
\begin{equation*}
y(x)=\left(x_{1}-5\right)^{3}+\left(x_{2}-4.5\right)^{3}+\left(x_{1}-1\right)^{2}+x_{2}^{2}+500 \tag{2.9}
\end{equation*}
$$

to label the training samples. We generate 2,601 samples with input $x=\left[x_{1}, x_{2}\right]$ using a uniform grid over the space with a mesh of 0.1 on each coordinate over the bounded space $[0,5]^{2}$. The dataset we obtained is denoted as $D=\left\{\left(X_{n}, Y_{n}\right)\right\} . g(x)$ is trained using $D$. We note that the region $[0,5]^{2}$ is large enough in our experiments, so that $g(x)$ can be thought of as being set to $-\infty$ outside this box. For instance, when $\sigma^{2}=1$, the ratio of the probability of falling outside $[-5,5]^{d}$ (as $[0,5]^{d}$ is almost equivalent to $[-5,5]^{d}$ here) to the probability of interest (for the first example with $\gamma=500$ or the second example with $\gamma=8$ ) is smaller than 0.05 , the largest ratio among all considered settings.

We first train a neural network predictor as $g(x)$. The neural network has 3 layers with 100 neurons in each of the 2 hidden layers, and all neurons are ReLU. To illustrate the rare-event set in the problem, we use $\gamma=500$ in this example. The defined rare-event set is presented in Figure 2.3. We observe that the set is roughly convex and should have a single dominating point. We obtain the dominating point for the set at $(3.3676,2.6051)$.

In our experiments, we first vary the value of $\gamma$ to verify the asymptotic performance of the proposed IS estimator as $\gamma$ increases. We then vary the value of $\sigma^{2}$ to create problems with different distribution setups, where a smaller $\sigma^{2}$ gives a rarer probability.


Figure 2.3: Rare-event set and dominating points for the neural network (case 1).


Figure 2.4: Rare-event set and dominating points for the random forest (case 1).

Figures 2.5 and 2.6 present the experimental results with fixed $\sigma^{2}=0.3$ and a varing $\gamma$ based on 50,000 samples. Figure 2.5 shows that the proposed IS estimator provides similar estimates as the baseline estimator, while Figure 2.6 shows our estimator provides a better confidence interval width and the advantage grows slightly as $\gamma$ increases.

In Figures 2.7 and 2.8, we present the experimental results for different variance values with $\gamma=500$. Again we observe the proposed IS scheme provides smaller relative errors in all cases and the advantage increases with smaller variance (the relative error increases from 2.5 to 10 for the proposed IS and 5 to 55 for the uniform IS in the considered range of $\sigma$ ).

Next, we investigate how the size of the predictor would affect the efficiency of our proposed estimator. We note that a neural network with a larger size results in a larger number of linear pieces in the rare-event set formulation. To obtain rare-event sets with different numbers of linear pieces, we use neural networks with different number of neurons for training and subsequently building the rare-event sets. In particular, we vary the number of neurons in the second layer and keep other parameters fixed.

Table 2.1 presents the computation time for solving the mixed integer optimization under different cases. Although the numbers of constraints and variables increase by roughly $30 \%$ (from 150 total neurons to 200) as we increase the number of second layer neurons, there is no signif-
icant increase in the computation time. In Figures 2.9 and 2.10, we present the performances of our IS estimator. We observe that our IS estimator consistently outperforms the naive estimator as evidenced by the similar estimates in Figure 2.9 and the smaller relative errors in Figure 2.10.


Figure 2.5: Probability estimation with different $\gamma$. Neural network, case 1.


Figure 2.7: Probability estimation with different distribution setups. Neural network, case 1.


Figure 2.6: 95\% confidence interval half-width with different $\gamma$. Neural network, case 1 .


Figure 2.8: Relative error with different distribution setups. Neural network, case 1 .

Table 2.1: The computation time for solving the mixed integer optimization to obtain the first dominating point in the neural network defined rare-event set in case 1.

| Number of Layer 2 Neurons | 50 | 55 | 60 | 65 | 70 | 75 | 80 | 85 | 90 | 95 | 100 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Number of Total Neurons | 150 | 155 | 160 | 165 | 170 | 175 | 180 | 185 | 190 | 195 | 200 |
| Computation Time (sec) | 0.323 | 0.390 | 0.217 | 0.218 | 0.205 | 0.379 | 0.384 | 0.436 | 0.357 | 0.235 | 0.425 |



Figure 2.9: Probability estimation with different different neural network sizes, case 1 .


Figure 2.11: Probability estimation with different distribution setups. Random forest, case 1.


Figure 2.10: Relative error with different neural network sizes, case 1 .


Figure 2.12: Relative error with different distribution setups. Random forest, case 1.

Next, we train a random forest $g(x)$, which ensembles three regression trees (see further training details in Section 2.9). The three regression trees are averaged and each of them has around 600 nodes. Again we illustrate the rare-event set with $\gamma=500$, which is presented in Figure 2.4. The dominating point is obtained by implementing Algorithm 2.1, which is located at $(3.05,2.65)$.

Figures 2.11 and 2.12 show our results with random forest. In Figure 2.11, we observe that the estimates for the two IS schemes are similar in all considered cases. On the other hand, Figure 2.12 shows the relative error for the proposed IS is smaller in all considered $\sigma$. Moreover, as the rarity increases, the relative error of the proposed IS increases from roughly 2.5 to 5 , whereas the
relative error of the uniform IS increases from 5 to 40. The slower increasing rate indicates that the proposed IS scheme is more efficient and the outperformance is stronger for rarer problems.

We now consider true output values generated according to the function

$$
\begin{equation*}
y(x)=10 \times e^{-\left(\frac{x_{1}-5}{3}\right)^{2}-\left(\frac{x_{2}-5}{4}\right)^{2}}+10 \times e^{-x_{1}^{2}-\left(x_{2}-4.5\right)^{2}} \tag{2.10}
\end{equation*}
$$

Again we use a uniform grid over $[0,5]^{2}$ with a mesh of 0.1 on each coordinate to train the predictors. The random forest ensembles three regression trees with around 600 nodes and the neural network with 2 hidden layers, 100 neurons in the first hidden layer and 50 neurons in the second hidden layer. All neurons in the neural network are ReLU.

For $\gamma=8$, the shapes of the rare-event sets are shown in Figures 2.13 and 2.14. We observe that the set now consists of two disjoint regions and therefore we expect to obtain multiple dominating points. Using Algorithm 2.1, we obtain two dominating points in each case: $(0,4.15)$ and $(3.75,3.55)$ for the random forest model; $(0.113,4.162)$ and $(4.187,3.587)$ for the neural network model. Again we vary $\gamma$ and $\sigma^{2}$ to obtain problems with different rarities and use 50,000 samples for each case.


Figure 2.13: Rare-event set and dominating point for the random forest (case 2).


Figure 2.14: Rare-event set and dominating point for the neural network (case 2).

Figures 2.15 and 2.16 shows the experiment results with fixed $\sigma^{2}=0.3$ and a varying $\gamma$. As in the first example, we observe that the IS estimator provides correct estimates with better confidence


Figure 2.15: Probability estimation with different $\gamma$. Neural network, case 2.


Figure 2.16: 95\% confidence interval halfwidth with different $\gamma$. Neural network, case 2.
intervals through all considered cases. The experimental results with fixed $\gamma=8$ varying $\sigma^{2}$ for the random forest predictor are shown in Figures 2.17 and 2.18, and the results for the neural network predictor are shown in Figures 2.19 and 2.20. Similar to the previous problem, both IS schemes give similar estimates in all the cases, as observed in Figures 2.17 and 2.19. The relative errors shown in Figures 2.18 and 2.20 illustrate that, as the probability of interest decreases, the relative error ratio between the uniform IS and the proposed IS increases from 2 to around 5-6. We can conclude that the proposed IS scheme again outperforms the uniform IS and is more preferable as the rarity increases.


Figure 2.17: Probability estimation with different distribution setups. Random forest, case 2.


Figure 2.18: Relative error with different distribution setups. Random forest, case 2.


Figure 2.19: Probability estimation with different distribution setups. Neural network, case 2.


Figure 2.20: Relative error with different distribution setups. Neural network, case 2.

### 2.6.2 MAGIC Gamma Telescope Data Set

We study a rare-event probability estimation problem from a realistic classification task. The classification problem uses the MAGIC Gamma Telescope data set in the UCI Machine Learning Repository [112]. The problem is to classify images of electromagnetic showers collected by a ground-based atmospheric Cherenkov gamma telescope. The features of the data are 10dimensional characteristic parameters of the images and the data set contains 19020 data points in total. We provide some descriptive statistics of the data set in Table 2.2. Studies [113, 114, 115] use machine learning predictors to discriminate images caused by a "signal" (primary gammas) from those initiated by the "background" (cosmic rays in the upper atmosphere).

Table 2.2: Descriptive statistics of the MAGIC Gamma Telescope Data Set. "Std" denotes the standard deviation and "CoV" denotes the coefficient of variation (ratio of the standard deviation to the mean).

| Coefficient Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Mean | 53.250 | 22.181 | 2.825 | 0.380 | 0.215 | -4.332 | 10.546 | 0.250 | 27.646 | 193.818 |
| Std | 42.365 | 18.346 | 0.473 | 0.183 | 0.111 | 59.206 | 51.000 | 20.827 | 26.104 | 74.732 |
| CoV | 0.796 | 0.827 | 0.167 | 0.481 | 0.515 | -13.668 | 4.836 | 83.401 | 0.944 | 0.386 |
| Min | 4.284 | 0.000 | 1.941 | 0.013 | 0.000 | -457.916 | -331.780 | -205.895 | 0.000 | 1.283 |
| Max | 334.177 | 256.382 | 5.323 | 0.893 | 0.675 | 575.241 | 238.321 | 179.851 | 90.000 | 495.561 |
| Median | 37.148 | 17.140 | 2.740 | 0.354 | 0.197 | 4.013 | 15.314 | 0.666 | 17.680 | 191.851 |

To train the predictors, we allocate 15,000 data points as the training set and use the remaining

4,020 data points as the testing set. All data were normalized to avoid scaling issues in training. We train a random forest that ensembles 10 random trees to achieve $85.6 \%$ testing set accuracy. For neural network, we use 2 hidden layers with 20 neurons and achieved $87 \%$ testing set accuracy.

The rare-event probability of interest is the statistical robustness metric (Example 2.1) of the two trained predictors. Specifically, we consider a testing data point, say with input $x$ and true label $y$, that is correctly predicted in both predictors (the predicted value $g(x)$ is consistent with $y)$. Then we perturb the input $x$ with a Gaussian noise $\epsilon \sim N\left(0, I \sigma^{2}\right)$ and estimate the probability of $P(g(x+\epsilon) \neq y)$, where we use uniform variance for each dimension because the input space was normalized. In our experiment, we vary the value of $\sigma^{2}$ to construct rare-event with different rarities. Note that, as discussed in Example 2.1, $P(g(x+\epsilon) \neq y)$ can be transformed into the format considered in this chapter, i.e. $P(g(X)>\gamma)$.

First, we implement Algorithm 2.1 to obtain dominating points for the rare-event sets $\{g(x+$ $\epsilon) \neq y\}$ with random forest and neural network as $g(\cdot)$ respectively. We obtain 53 dominating points for the rare-event sets associated with the random forest predictor and 217 dominating points in the neural network case. The IS distributions are constructed using these dominating points. In both problems, $\sigma^{2}$ ranges from 0.03 to 0.1 and we use 50,000 samples to estimate each target rare-event probabilities.

The experimental results for the random forest and neural network are presented in Figures 2.21 and 2.22 respectively. We observe that the estimates are very accurate in all experiments (with different rarities), which are indicated by the tight $95 \%$ confidence intervals. These results show that our proposed IS scheme performs well with large numbers of dominating points and in relatively high-dimensional problems.

### 2.7 Proofs of Theorems

Throughout this section, we write $f_{1}(\gamma) \sim f_{2}(\gamma)$ if $\lim _{\gamma \rightarrow \infty} f_{1}(\gamma) / f_{2}(\gamma)=1$ and $f_{1}(\gamma) \stackrel{\text { poly }}{\sim}$ $f_{2}(\gamma)$ if $f_{1}(\gamma) / f_{2}(\gamma)$ changes at most polynomially in $\gamma$. Unless otherwise defined, we use $x_{i}$ to denote the $i$-th component of a vector $x$. For any vectors $x, y \in \mathbb{R}^{d}$, we write $x \geq y$ if $x_{i} \geq y_{i}$ for


Figure 2.21: Probability estimation with different distribution setups. Random forest, MAGIC.


Figure 2.22: Probability estimation with different distribution setups. Neural network, MAGIC.
any $i=1, \ldots, d$. For any index sets $I, J \subset\{1, \ldots, d\}$ and any $x \in \mathbb{R}^{d}, A \in \mathbb{R}^{d \times d}$, we use $x_{I}$ to denote the subvector $\left(x_{i}\right)_{i \in I}$ and use $A_{I J}$ to denote the submatrix $\left(A_{i j}\right)_{i \in I, j \in J}$.

First of all, we adapt Theorem 4.1 in [62] to obtain the following lemma.

Lemma 2.1. Let $Y$ be a d-dimensional Gaussian random vector with zero mean and positive definite covariance matrix $\tilde{\Sigma}$. Suppose that $\tilde{s}=\tilde{s}(\gamma) \notin[-\infty, 0]^{d}$ is a vector in $[-\infty, \infty)^{d}$ such that as $\gamma \rightarrow \infty$, at least one of its components goes to $\infty$. Use $y^{*}$ to denote $\arg \min _{y \geq \tilde{s}} y^{T} \tilde{\Sigma}^{-1} y$. Then by Proposition 2.1 in [62], we know that there exists a unique set $I \subset\{1, \cdots, d\}$ such that

$$
\begin{align*}
& 1 \leq|I| \leq d  \tag{2.11a}\\
& y_{I}^{*}=\tilde{s}_{I} \neq \mathbf{0}_{I} ;  \tag{2.11b}\\
& \text { If } J:=\{1, \ldots, d\} \backslash I \neq \emptyset, \text { then } y_{J}^{*}=-\left(\tilde{\Sigma}^{-1}\right)_{J J}^{-1}\left(\tilde{\Sigma}^{-1}\right)_{J I} \tilde{s}_{I} \geq \tilde{s}_{J} ;  \tag{2.11c}\\
& \forall i \in I, e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}>0 ;  \tag{2.11d}\\
& \min _{y \geq \tilde{s}} y^{T} \tilde{\Sigma}^{-1} y=\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*}>0 . \tag{2.11e}
\end{align*}
$$

We suppose that for sufficiently large $\gamma$, the set I does not change with $\gamma$ and if $J \neq \emptyset, \lim _{\gamma \rightarrow \infty}(\tilde{s}-$ $\left.y^{*}\right)_{J}=\tilde{s}_{J}^{*}$ where $\tilde{s}^{*}$ is a constant vector in $[-\infty, \infty)^{|J|}$. Suppose further that $\forall i \in I, e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}$
either goes to $\infty$ or is a positive constant. Then as $\gamma \rightarrow \infty$, we have that

$$
P(Y \geq \tilde{s}) \sim C \frac{\exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\}}{\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{S}_{I}}
$$

where $C$ is a positive constant in $\gamma$.

Before showing the proof, we provide a brief and intuitive explanation on the index set $I$ as defined in the lemma. We minimize $y^{T} \tilde{\Sigma}^{-1} y$ subject to $y \geq \tilde{s}$. Among the constraints, $y_{I} \geq \tilde{s}_{I}$ is crucial while $y_{J} \geq \tilde{s}_{J}$ could be removed without affecting the optimal solution. Thus, the original optimization problem is equivalent to minimizing $y^{T} \tilde{\Sigma}^{-1} y$ subject to $y_{I}=\tilde{s}_{I}, y_{J} \in \mathbb{R}^{|J|}$. For example, if $d=2, \tilde{s}=(1,0)^{T}$ and $\tilde{\Sigma}$ is the identity matrix, then $I=\{1\}$ and $J=\{2\}$ since $y_{1} \geq 1$ could not be removed while $y_{2} \geq 0$ could. Now we prove the lemma:

Proof of Lemma 2.1. Given $x \in \mathbb{R}^{d}$, we define the transformation $\tilde{x}$ in the following way: $\tilde{x}_{i}=$ $\left(e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}\right)^{-1} x_{i}, \forall i \in I ; \tilde{x}_{J}=x_{J}$. Using (3.4) in [62], we know that

$$
\left(x+y^{*}\right)^{T} \tilde{\Sigma}^{-1}\left(x+y^{*}\right)=x^{T} \tilde{\Sigma}^{-1} x+2 x_{I}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}+\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*},
$$

and thus

$$
\begin{aligned}
\phi\left(\tilde{x}+y^{*} ; 0, \tilde{\Sigma}\right) & =(2 \pi)^{-\frac{d}{2}}|\tilde{\Sigma}|^{-\frac{1}{2}} \exp \left\{-\frac{1}{2}\left[\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x}+2 \tilde{x}_{I}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}+\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*}\right]\right\} \\
& =(2 \pi)^{-\frac{d}{2}}|\tilde{\Sigma}|^{-\frac{1}{2}} \exp \left\{-\frac{1}{2}\left[\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x}+2 x_{I}^{T} \mathbf{1}_{I}+\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*}\right]\right\}
\end{aligned}
$$

Then we get that

$$
\begin{aligned}
& P(Y \geq \tilde{s}) \\
= & \int_{y \geq \tilde{s}} \phi(y ; 0, \tilde{\Sigma}) \mathrm{d} y \\
= & \int_{\tilde{x} \geq \tilde{s}-y^{*}} \phi\left(\tilde{x}+y^{*} ; 0, \tilde{\Sigma}\right) \mathrm{d} \tilde{x} \\
= & \left.\left(\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{S}_{I}\right)^{-1} \int_{x \geq \tilde{s}-y^{*}} \phi\left(\tilde{x}+y^{*} ; 0, \tilde{\Sigma}\right) \mathrm{d} x \quad \text { (Here } \tilde{x} \text { can be viewed as a function of } x .\right) \\
= & (2 \pi)^{-\frac{d}{2}}|\tilde{\Sigma}|^{-\frac{1}{2}}\left(\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{S}_{I}\right)^{-1} \exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\} \int_{x \geq \tilde{s}-y^{*}} \exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x .
\end{aligned}
$$

Apparent from the above, it suffices to show that $\int_{x \geq \tilde{s}-y^{*}} \exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x$ converges to a positive constant as $\gamma \rightarrow \infty$. We will prove this result via applying the dominated convergence theorem. We first need to derive an integrable upper bound for the integrand. Indeed, using (3.6) in [62] we know that

$$
\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x}+2 \tilde{x}_{I}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}+\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} \geq \tilde{x}_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} \tilde{x}_{J}+2 \tilde{x}_{I}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}+\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*}
$$

and hence $\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} \geq \tilde{x}_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} \tilde{x}_{J}=x_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} x_{J}$. Thus

$$
\exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \leq \exp \left\{-x_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} x_{J} / 2-x_{I}^{T} \mathbf{1}_{I}\right\}
$$

Moreover, we have that

$$
\begin{aligned}
\int_{x \geq \tilde{s}_{-y^{*}}} \exp \left\{-x_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} x_{J} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x & \leq \int_{x_{I} \geq \mathbf{0}_{I}, x_{J} \in \mathbb{R}^{|J|}} \exp \left\{-x_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} x_{J} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x \\
& =\int_{\mathbb{R}^{|J|}} \exp \left\{-x_{J}^{T}\left(\tilde{\Sigma}_{J J}\right)^{-1} x_{J} / 2\right\} \mathrm{d} x_{J}<\infty .
\end{aligned}
$$

To investigate the limit of $\exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\}$, we further partition $I$ into $I_{1}=\{i \in I$ : $\left.e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{S}_{I} \rightarrow \infty\right\}$ and $I_{2}=\left\{i \in I: e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{S}_{I}\right.$ is a positive constant $\}$. By the definition, we
know that for any given $x \in \mathbb{R}^{d}, \tilde{x}_{i} \rightarrow 0$ for $i \in I_{1}$ and $\tilde{x}_{i}$ is a constant for $i \in I_{2}$ or $J$. Then we get that for any $x$,

$$
\begin{aligned}
& \lim _{\gamma \rightarrow \infty} \exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \\
= & \exp \left\{-\frac{1}{2}\left[\tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} I_{2}} \tilde{x}_{I_{2}}+2 \tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} J} \tilde{x}_{J}+\tilde{x}_{J}^{T}\left(\tilde{\Sigma}^{-1}\right)_{J J} \tilde{x}_{J}\right]-x_{I}^{T} \mathbf{1}_{I}\right\} \\
= & \exp \left\{-\frac{1}{2}\left[\tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} I_{2}} \tilde{x}_{I_{2}}+2 \tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} J} x_{J}+x_{J}^{T}\left(\tilde{\Sigma}^{-1}\right)_{J J} x_{J}\right]-x_{I}^{T} \mathbf{1}_{I}\right\} .
\end{aligned}
$$

By applying the dominated convergence theorem, we get that

$$
\begin{aligned}
& \lim _{\gamma \rightarrow \infty} \int_{x \geq \tilde{s}-y^{*}} \exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x \\
= & \iint_{x_{I} \geq \mathbf{0}_{I}, x_{J} \geq \tilde{s}_{J}^{*}} \exp \left\{-\frac{1}{2}\left[\tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} I_{2}} \tilde{x}_{I_{2}}+2 \tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} J} x_{J}+x_{J}^{T}\left(\tilde{\Sigma}^{-1}\right)_{J J} x_{J}\right]-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x_{I} \mathrm{~d} x_{J} \\
= & \iint_{x_{I_{2} \geq \mathbf{0}_{I_{2}}, x_{J} \geq \tilde{s}_{J}^{*}}} \exp \left\{-\frac{1}{2}\left[\tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} I_{2}} \tilde{x}_{I_{2}}+2 \tilde{x}_{I_{2}}^{T}\left(\tilde{\Sigma}^{-1}\right)_{I_{2} J} x_{J}+x_{J}^{T}\left(\tilde{\Sigma}^{-1}\right)_{J J} x_{J}\right]-x_{I_{2}}^{T} \mathbf{1}_{I_{2}}\right\} \mathrm{d} x_{I_{2}} \mathrm{~d} x_{J .} .
\end{aligned}
$$

This shows that $\int_{x \geq \tilde{s}-y^{*}} \exp \left\{-\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} / 2-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x$ converges to a positive constant as $\gamma \rightarrow \infty$, and hence we have proved the theorem.

Now we apply Lemma 2.1 and the techniques in its proof to derive the following result:

Lemma 2.2. Suppose that $X \sim N(\mu, \Sigma)$ where $\mu \in \mathbb{R}^{d}$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. Let $A \sim \mathbb{R}^{m \times d}$ be a constant matrix and $t \in \mathbb{R}^{m}$ be a vector. In particular, $t_{1}=\gamma+c$ for some constant $c \in \mathbb{R}$ and $t_{2}, \ldots, t_{m}$ are all constants in $\mathbb{R}$. Assume that $P(A X \geq t)>0$ for any $\gamma \in \mathbb{R}$. Define $x^{*}=\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): A x \geq t\right\}$. Then
(i) Use $A_{i}$ to denote the $i$-th row vector of $A$ and define $\mathcal{A}(x)=\left\{1 \leq i \leq m: A_{i}^{T} x=t_{i}\right\}$ for $x \in \mathbb{R}^{d}$. For sufficiently large $\gamma, \mathcal{A}\left(x^{*}\right)$ does not change with $\gamma$.
(ii) For sufficiently large $\gamma$, each component of $x^{*}$ is affine in $\gamma$.
(iii) As $\gamma \rightarrow \infty$,

$$
P(A X \geq t) \stackrel{\text { poly }}{\sim} \exp \left\{-\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right) / 2\right\}
$$

Proof of Lemma 2.2. For simplicity, we denote the polyhedron $\left\{x \in \mathbb{R}^{d}: A x \geq t\right\}$ as $P_{1}$.
(i\&ii) Note that $x^{*}$ is the optimal solution to a quadratic programming problem. It is known that

$$
\begin{equation*}
x^{*}=\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): A_{i}^{T} x=t_{i}, \forall i \in \mathcal{A}\left(x^{*}\right)\right\} . \tag{2.12}
\end{equation*}
$$

Moreover, as $\gamma$ grows, actually only the first constraint $A_{1}^{T} x \geq t_{1}=\gamma+c$ shifts with $\gamma$ while the other $m-1$ constraints keep unchanged. Thus we must have $1 \in \mathcal{A}\left(x^{*}\right)$ for sufficiently large $\gamma$. Indeed, if $1 \notin \mathcal{A}\left(x^{*}\right)$, then from (2.12), $x^{*}$ must belong to $\left\{\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): A_{i}^{T} x=\right.\right.$ $\left.\left.t_{i}, \forall i \in \bar{I}\right\}: \bar{I} \subset\{2, \ldots, m\}\right\}$, which is a finite set of constant vectors. However, we have that $A_{1}^{T} x^{*} \geq t_{1}=\gamma+c$, so when $\gamma$ is large enough, $x^{*}$ cannot be one of these constant vectors and hence $1 \in \mathcal{A}\left(x^{*}\right)$.

We consider the "candidate points" defined as follows. For any fixed index set $\tilde{I} \subset\{1, \ldots, m\}$ such that $1 \in \tilde{I},\left\{x \in \mathbb{R}^{d}: A_{i}^{T} x=t_{i}, i \in \tilde{I}\right\} \neq \emptyset$ and the constraints $A_{i}^{T} x=t_{i}, i \in \tilde{I}$ are linearly independent for sufficiently large $\gamma$ (we call such $\tilde{I}$ as valid), we solve $x^{*}(\tilde{I})=\arg \min \{(x-$ $\left.\mu)^{T} \Sigma^{-1}(x-\mu): A_{i}^{T} x=t_{i}, \forall i \in \tilde{I}\right\}$. If $x^{*}(\tilde{I})$ is feasible for the original problem, i.e. $x^{*}(\tilde{I}) \in P_{1}$, then we call $x^{*}(\tilde{I})$ a candidate point.

We note that the total number of valid $\tilde{I}$ is finite since $\tilde{I}$ is always a subset of $\{1, \ldots, m\}$. Without loss of generality, from now on we assume that $\gamma$ is large enough such that $1 \in \mathcal{A}\left(x^{*}\right)$ and for any valid $\tilde{I},\left\{x \in \mathbb{R}^{d}: A_{i}^{T} x=t_{i}, i \in \tilde{I}\right\} \neq \emptyset$ and $A_{i}^{T} x=t_{i}, i \in \tilde{I}$ are linearly independent. In this case, $x^{*}$ is the candidate point which attains the minimum objective value.

First, we show that for any valid $\tilde{I}$, each component of $x^{*}(\tilde{I})$ is affine in $\gamma$. Suppose that $\tilde{I}=\left\{i_{1}, \ldots, i_{|\tilde{I}|}\right\}$ with $i_{1}<\cdots<i_{|\tilde{I}|}$. We have that $A_{i_{j}}, j=1, \ldots,|\tilde{I}|$ are linearly independent. Let $A(\tilde{I}) \in \mathbb{R}^{d \times d}$ be a constant invertible matrix whose $j$-th row vector is $A_{i_{j}}$ for $j=1, \ldots,|\tilde{I}|$. Consider the transformation $y=A(\tilde{I})(x-\mu)$ and solve $y^{*}(\tilde{I})=\arg \min \left\{y^{T}\left(A(\tilde{I})^{-1}\right)^{T} \Sigma^{-1} A(\tilde{I})^{-1} y\right.$ : $\left.y_{j}=t_{i_{j}}-A_{i_{j}}^{T} \mu, j=1, \ldots,|\tilde{I}|\right\}$. We have that $x^{*}(\tilde{I})=A(\tilde{I})^{-1} y^{*}(\tilde{I})+\mu$. To ease the notation, we
denote $\Sigma^{\prime}=A(\tilde{I})^{T} \Sigma A(\tilde{I}), t^{\prime}=t_{\tilde{I}}-(A \mu)_{\tilde{I}}, I^{\prime}=\{1, \ldots,|\tilde{I}|\}$ and $J^{\prime}=\{1, \ldots, d\} \backslash I^{\prime}$. Then

$$
\begin{aligned}
y^{*}(\tilde{I}) & =\arg \min \left\{y^{T} \Sigma^{\prime-1} y: y_{I^{\prime}}=t^{\prime}\right\} \\
& =\arg \min \left\{y_{I^{\prime}}^{T}\left(\Sigma^{\prime-1}\right)_{I^{\prime} I^{\prime}} y_{I^{\prime}}+2 y_{I^{\prime}}^{T}\left(\Sigma^{\prime-1}\right)_{I^{\prime} J^{\prime}} y_{J^{\prime}}+y_{J^{\prime}}^{T}\left(\Sigma^{\prime-1}\right)_{J^{\prime} J^{\prime}} y_{J^{\prime}}: y_{I^{\prime}}=t^{\prime}\right\} .
\end{aligned}
$$

By solving the above problem, we get that $y^{*}(\tilde{I})_{I^{\prime}}=t^{\prime}$ and $y^{*}(\tilde{I})_{J^{\prime}}=-\left(\Sigma^{\prime-1}\right)_{J^{\prime} J^{\prime}}^{-1}\left(\Sigma^{\prime-1}\right)_{J^{\prime} I^{\prime}} t^{\prime}$. By the definition, for fixed index set $\tilde{I}, \Sigma^{\prime}$ is a constant matrix. Besides, $t_{1}^{\prime}=t_{1}-(A \mu)_{1}$ is an affine function in $\gamma$ while other components of $t^{\prime}$ are all constants. Hence, each component of $y^{*}(\tilde{I})$ is affine in $\gamma$. As a result, each component of $x^{*}(\tilde{I})$ is also affine in $\gamma$.

To check whether $x^{*}(\tilde{I})$ is feasible, it is equivalent to check whether $A_{i}^{T} x^{*}(\tilde{I}) \geq t_{i}$ for any $i \notin \tilde{I}$. We know that for $i \notin \tilde{I}$ (which implies that $i \neq 1$ ), $A_{i}^{T} x^{*}(\tilde{I})$ is affine in $\gamma$ while $t_{i}$ is a constant. Hence, for sufficiently large $\gamma$, it is determined whether $x^{*}(\tilde{I})$ is a candidate point or not. Again, the total number of valid $\tilde{I}$ is finite. Therefore, $\left\{\tilde{I}: x^{*}(\tilde{I})\right.$ is a candidate point $\}$ does not change for large $\gamma$.

Finally, for each $\tilde{I}$ such that $x^{*}(\tilde{I})$ is a candidate point for sufficiently large $\gamma$, we have that the objective value $\left(x^{*}(\tilde{I})-\mu\right)^{T} \Sigma^{-1}\left(x^{*}(\tilde{I})-\mu\right)$ is a quadratic function of $\gamma$. Recall that $x^{*}$ is the candidate point with minimum objective value. Thus, when $\gamma$ is sufficiently large, $\left\{\tilde{I}: x^{*}=x^{*}(\tilde{I})\right\}$ must be non-empty and fixed. We pick a specific $\tilde{I}$ such that $x^{*}=x^{*}(\tilde{I})$ for sufficiently large $\gamma$. Since we have proved that each component of $x^{*}(\tilde{I})$ is affine in $\gamma$, we get statement (ii). Then when $\gamma$ is large enough, for each $i$, it is determined whether $A_{i}^{T} x^{*}=t_{i}$, i.e. $i \in \mathcal{A}\left(x^{*}\right)$, which completes the proof of (i).
(iii) In this proof, we will construct a superset and a subset of $P_{1}$, and hence develop an upper bound and a lower bound for $P\left(X \in P_{1}\right)$. Then it suffices to show that both bounds are approximately $\exp \left\{-\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right) / 2\right\}$ up to polynomial factors.

First, we construct the superset of $P_{1}$ by removing constraints. Following the above proof, we can find a maximal valid index set $\tilde{I}=\left\{i_{1}, \ldots, i_{m^{\prime}}\right\}$ such that $1=i_{1}<\cdots<i_{m^{\prime}}$ and $x^{*}=x^{*}(\tilde{I})$ for sufficiently large $\gamma$. Intuitively, $\left\{A_{i_{j}}^{T} x \geq t_{i_{j}}, j=1, \ldots, m^{\prime}\right\}$ is the maximal linearly independent
subset of active constraints at $x^{*}$. If $m^{\prime}<d$, then we can add redundant constraints in the form of $x_{k_{l}} \geq-\infty, l=1, \cdots, d-m^{\prime}$ such that we get $d$ linearly independent constraints now. More specifically, let

$$
B=\left(\begin{array}{c}
A_{i_{1}}^{T} \\
\vdots \\
A_{i_{m^{\prime}}}^{T} \\
e_{k_{1}}^{T} \\
\vdots \\
e_{k_{d-m^{\prime}}}^{T}
\end{array}\right), s=\left(\begin{array}{c}
t_{i_{1}} \\
\vdots \\
t_{i_{m^{\prime}}} \\
-\infty \\
\vdots \\
-\infty
\end{array}\right) .
$$

By the construction, we get that $B$ is a $d \times d$ constant invertible matrix. Denote $P_{2}=\left\{x \in \mathbb{R}^{d}\right.$ : $B x \geq s\}$. Since $P_{2}$ is obtained by removing constraints from $P_{1}$, we have that $P_{1} \subset P_{2}$ and thus $P\left(X \in P_{1}\right) \leq P\left(X \in P_{2}\right)$. Now we develop the asymptotic result of $P\left(X \in P_{2}\right)$, where we directly apply Lemma 2.1.

We know that $Y:=B(X-\mu) \sim N(0, \tilde{\Sigma})$ where $\tilde{\Sigma}=B \Sigma B^{T}$ is positive definite. We denote $y^{*}=\arg \min \left\{y^{T} \tilde{\Sigma}^{-1} y: y \geq \tilde{s}\right\}$ where $\tilde{s}=s-B \mu$. It is easy to verify that $y^{*}=B\left(x^{*}-\mu\right)$ and $\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*}=\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right)$. From (ii), we know that each component of $y^{*}$ is also affine in $\gamma$ for large $\gamma$.

Now we verify the assumptions of Lemma 2.1. Recall that under our settings, $s_{1}=\gamma+c$ for some constant $c \in \mathbb{R}$ so $\tilde{s}_{1} \rightarrow \infty$ as $\gamma \rightarrow \infty$. We still use the symbol $I$ to denote the set that satisfies (2.11). By the definition, clearly $\left\{m^{\prime}+1, \ldots, d\right\} \subset J=\{1, \ldots, d\} \backslash I$. Basically, $I$ is the minimal subset of $\left\{1, \ldots, m^{\prime}\right\}$ such that $x^{*}=x^{*}\left(\left\{i_{j}: j \in I\right\}\right)$. Following the previous proof, we know that $1 \in I$ and $I$ does not change for sufficiently large $\gamma$. Moreover, we know that $y^{*} \geq \tilde{s}$ and each component of $y^{*}$ is affine in $\gamma$, then the limit $\lim _{\gamma \rightarrow \infty}\left(\tilde{s}-y^{*}\right)_{J}$ exists in $[-\infty, 0]^{|J|}$. Indeed, for $j \in J \cap\left\{2, \ldots, m^{\prime}\right\}, \tilde{s}_{j}$ is a constant and then $\tilde{s}_{j}-y_{j}^{*}$ converges to $-\infty$ or a nonpositive constant while for $j \in\left\{m^{\prime}+1, \ldots, d\right\}, \tilde{s}_{j}-y_{j}^{*} \equiv-\infty$. Finally, for any $i \in I$, we know that $e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}>0$ and it is an affine function of $\gamma$, and thus either it goes to $\infty$ or it is a positive constant as $\gamma \rightarrow \infty$.

In conclusion, all the assumptions of Lemma 2.1 hold in this case. Therefore, we get that

$$
P\left(X \in P_{2}\right) \sim C \frac{\exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\}}{\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}}=C \frac{\exp \left\{-\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right) / 2\right\}}{\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}}
$$

for some positive constant $C$, which implies that

$$
P\left(X \in P_{2}\right) \stackrel{\text { poly }}{\sim} \exp \left\{-\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right) / 2\right\}
$$

Clearly, if $P_{1}=P_{2}$, then we have proved statement (iii). From now on, we assume that $P_{1} \neq P_{2}$. In this case, $P_{2}$ is a relaxation of $P_{1}$ by removing inactive constraints at $x^{*}$. That is, for any $x \in \overline{P_{2} \backslash P_{1}}$, there exists $i \in\{1, \ldots, m\}$ such that $A_{i}^{T} x \leq t_{i}$ while $A_{i}^{T} x^{*}>t_{i}$. In particular, $x^{*} \notin \overline{P_{2} \backslash P_{1}}$.

Next, we construct the subset of $P_{1}$ by selecting a small neighborhood around $x^{*}$. Denote $x^{* *}=$ $\arg \min \left\{\left(x-x^{*}\right)^{T} \Sigma^{-1}\left(x-x^{*}\right): x \in \overline{P_{2} \backslash P_{1}}\right\}$. Note that $\overline{P_{2} \backslash P_{1}}$ can be expressed as the union of finite polyhedrons, each of which is formed by a shifting constraint and some constant constraints like $P_{1}$. Similar to the previous arguments, we could derive that $\left(x^{* *}-x^{*}\right)^{T} \Sigma^{-1}\left(x^{* *}-x^{*}\right) \geq 0$ is a quadratic function of $\gamma$ for large $\gamma$. Thus we know that $\left(x^{* *}-x^{*}\right)^{T} \Sigma^{-1}\left(x^{* *}-x^{*}\right)$ either goes to $\infty$ or stays a nonnegative constant as $\gamma \rightarrow \infty$. However, if $\left(x^{* *}-x^{*}\right)^{T} \Sigma^{-1}\left(x^{* *}-x^{*}\right)=0$ for sufficiently large $\gamma$, then we have that $x^{* *}=x^{*}$, which contradicts with $x^{*} \notin \overline{P_{2} \backslash P_{1}}$. Therefore, there exists a constant $0<\varepsilon<1$ such that for sufficiently large $\gamma,\left(x^{* *}-x^{*}\right)^{T} \Sigma^{-1}\left(x^{* *}-x^{*}\right)>\varepsilon^{2}$, and hence $\overline{P_{2} \backslash P_{1}} \subset\left\{x \in \mathbb{R}^{d}:\left(x-x^{*}\right)^{T} \Sigma^{-1}\left(x-x^{*}\right)>\varepsilon^{2}\right\}$. Thus, $\left\{x \in \mathbb{R}^{d}:\left(x-x^{*}\right)^{T} \Sigma^{-1}\left(x-x^{*}\right) \leq\right.$ $\left.\varepsilon^{2}\right\} \cap P_{1}=\left\{x \in \mathbb{R}^{d}:\left(x-x^{*}\right)^{T} \Sigma^{-1}\left(x-x^{*}\right) \leq \varepsilon^{2}\right\} \cap P_{2}$ for sufficiently large $\gamma$. Correspondingly, there exists $\varepsilon^{\prime}>0$ such that $\left\{x \in \mathbb{R}^{d}:\|x\|_{\infty} \leq \varepsilon^{\prime}\right\} \subseteq\left\{x \in \mathbb{R}^{d}: x^{T} \Sigma^{-1} x \leq \varepsilon^{2}\right\}$.

Still we define $Y=B(X-\mu) \sim N(0, \tilde{\Sigma})$. Then we get that

$$
\begin{aligned}
P\left(X \in P_{1}\right) & \geq P\left(\left(X-x^{*}\right)^{T} \Sigma^{-1}\left(X-x^{*}\right) \leq \varepsilon^{2}, X \in P_{1}\right) \\
& =P\left(\left(X-x^{*}\right)^{T} \Sigma^{-1}\left(X-x^{*}\right) \leq \varepsilon^{2}, X \in P_{2}\right) \\
& =P\left(\left(Y+B \mu-B x^{*}\right)^{T} \tilde{\Sigma}^{-1}\left(Y+B \mu-B x^{*}\right) \leq \varepsilon^{2}, Y \geq \tilde{s}\right) .
\end{aligned}
$$

Similar to the proof of Lemma 2.1, we have that

$$
\begin{aligned}
& P\left(\left(Y+B \mu-B x^{*}\right)^{T} \tilde{\Sigma}^{-1}\left(Y+B \mu-B x^{*}\right) \leq \varepsilon^{2}, Y \geq \tilde{s}\right) \\
= & \int_{\left(y+B \mu-B x^{*}\right)^{T} \tilde{\Sigma}^{-1}\left(y+B \mu-B x^{*}\right) \leq \varepsilon^{2}, y \geq \tilde{s}} \phi(y ; 0, \tilde{\Sigma}) \mathrm{d} y \\
= & \int_{\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} \leq \varepsilon^{2}, \tilde{x} \geq \tilde{s}-y^{*}} \phi\left(\tilde{x}+y^{*} ; 0, \tilde{\Sigma}\right) \mathrm{d} \tilde{x} \\
= & \left(\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}\right)^{-1} \int_{\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} \leq \varepsilon^{2}, \tilde{x} \geq \tilde{s}-y^{*}} \phi\left(\tilde{x}+y^{*} ; 0, \tilde{\Sigma}\right) \mathrm{d} x \quad(\tilde{x} \text { is viewed as a function of } x .) \\
= & (2 \pi)^{-\frac{d}{2}} \left\lvert\, \tilde{\Sigma}^{-\frac{1}{2}}\left(\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}\right)^{-1} \exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\} \times\right. \\
& \int_{\tilde{x}^{T} \tilde{\Sigma}-1} \tilde{\Sigma}^{2} \leq \varepsilon^{2}, \tilde{x} \geq \tilde{s}-y^{*} \\
\geq & (2 \pi)^{-\frac{d}{2}}|\tilde{\Sigma}|^{-\frac{1}{2}}\left(\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}\right)^{-1} \exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\}\left(1-\varepsilon^{2} / 2\right) \times \\
& \int_{\tilde{x}^{T} \tilde{\Sigma}^{-1} \tilde{x} \leq \varepsilon^{2}, \tilde{x} \geq \tilde{s}-y^{*}} \exp \left\{-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x \\
\geq & (2 \pi)^{-\frac{d}{2}}|\tilde{\Sigma}|^{-\frac{1}{2}}\left(\prod_{i \in I} e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}\right)^{-1} \exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\}\left(1-\varepsilon^{2} / 2\right) \int_{0 \leq \tilde{x} \leq \varepsilon^{\prime} 1} \exp \left\{-x_{I}^{T} \mathbf{1}_{I}\right\} \mathrm{d} x \\
= & (2 \pi)^{-\frac{d}{2}}|\tilde{\Sigma}|^{-\frac{1}{2}}\left(1-\varepsilon^{2} / 2\right) \varepsilon^{\prime|J|}\left(\prod_{i \in I} \frac{1-\exp \left\{-e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{S}_{I} \varepsilon^{\prime}\right\}}{e_{i}^{T}\left(\tilde{\Sigma}_{I I}\right)^{-1} \tilde{s}_{I}}\right) \exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\} \\
p o l y & \exp \left\{-\left(y^{*}\right)^{T} \tilde{\Sigma}^{-1} y^{*} / 2\right\} \\
= & \exp \left\{-\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right) / 2\right\} .
\end{aligned}
$$

Combining the upper and lower bound for $P\left(X \in P_{1}\right)$, we finally get that

$$
P\left(X \in P_{1}\right) \stackrel{\text { poly }}{\sim} \exp \left\{-\left(x^{*}-\mu\right)^{T} \Sigma^{-1}\left(x^{*}-\mu\right) / 2\right\} .
$$

Now we use the asymptotic result in Lemma 2.2 to prove Theorem 2.1.

Proof of Theorem 2.1. Suppose that $g(x)=g_{i}(x)$ for $h_{i j}(x) \geq 0, j=1, \ldots, m_{i}, i=1, \ldots, r^{\prime}$ where
$g_{i}$ 's and $h_{i j}$ 's are all affine functions. Then we can split $\left\{x \in \mathbb{R}^{d}: g(x) \geq \gamma\right\}$ into $\tilde{\mathcal{R}}_{1}, \ldots, \tilde{\mathcal{R}}_{r^{\prime}}$ where $\tilde{\mathcal{R}}_{i}=\left\{x \in \mathbb{R}^{d}: g_{i}(x) \geq \gamma, h_{i j}(x) \geq 0, j=1, \ldots, m_{i}\right\}$. Without loss of generality, we assume that $P\left(X \in \tilde{\mathcal{R}}_{i}\right)>0, \forall i=1, \ldots, r^{\prime}$ for any $\gamma \in \mathbb{R}$. We denote $\tilde{a}_{i}=\arg \min \left\{(x-\mu)^{T} \Sigma^{-1}(x-\mu): x \in \tilde{\mathcal{R}}_{i}\right\}$. Applying Lemma 2.2, we get that for any $i=1, \ldots, r^{\prime}$,

$$
P\left(X \in \tilde{\mathcal{R}}_{i}\right) \stackrel{\text { poly }}{\sim} \exp \left\{-\left(\tilde{a}_{i}-\mu\right)^{T} \Sigma^{-1}\left(\tilde{a}_{i}-\mu\right) / 2\right\}
$$

Then we get that

$$
\begin{align*}
\tilde{E}[Z] & =P(g(X) \geq \gamma)=\sum_{i=1}^{r^{\prime}} P\left(X \in \tilde{\mathcal{R}}_{i}\right)  \tag{2.13}\\
& \text { poly }  \tag{2.14}\\
\sim & \exp \left\{-\min _{i=1, \ldots, r^{\prime}}\left(\tilde{a}_{i}-\mu\right)^{T} \Sigma^{-1}\left(\tilde{a}_{i}-\mu\right) / 2\right\}=\exp \left\{-\left(a_{1}-\mu\right)^{T} \Sigma^{-1}\left(a_{1}-\mu\right) / 2\right\} .
\end{align*}
$$

On the other hand, we have that for any $i=1, \ldots, r$ and $x \in \mathcal{R}_{i} \subset\left\{x \in \mathbb{R}^{d}:\left(a_{i}-\mu\right)^{T} \Sigma^{-1}(x-\right.$ $\left.\left.a_{i}\right) \geq 0\right\}$,

$$
\begin{aligned}
L(x) & \leq \frac{r e^{-(x-\mu)^{T} \Sigma^{-1}(x-\mu) / 2}}{e^{-\left(x-a_{i}\right)^{T} \Sigma^{-1}\left(x-a_{i}\right) / 2}} \\
& =r \exp \left\{-\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(a_{i}-\mu\right) / 2-\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(x-a_{i}\right)\right\} \\
& \leq r \exp \left\{-\left(a_{i}-\mu\right)^{T} \Sigma^{-1}\left(a_{i}-\mu\right) / 2\right\} \\
& \leq r \exp \left\{-\left(a_{1}-\mu\right)^{T} \Sigma^{-1}\left(a_{1}-\mu\right) / 2\right\} .
\end{aligned}
$$

Then we get that

$$
\begin{align*}
\tilde{E}\left[Z^{2}\right] & =\tilde{E}\left[I(g(\tilde{X}) \geq \gamma) L^{2}(\tilde{X})\right]=E[I(g(X) \geq \gamma) L(X)]  \tag{2.15}\\
& \leq r \exp \left\{-\left(a_{1}-\mu\right)^{T} \Sigma^{-1}\left(a_{1}-\mu\right) / 2\right\} P(g(X) \geq \gamma) . \tag{2.16}
\end{align*}
$$

Combining (2.14) and (2.16), we finally get that $\tilde{E}\left[Z^{2}\right] / \tilde{E}[Z]^{2}$ grows at most polynomially growing in $\gamma$, and hence $Z$ is asymptotically optimal.

Proof of Corollary 2.1. See (2.14) in the proof of Theorem 2.1.

Proof of Corollary 2.2. Now we suppose that $X \sim f(x)=\sum_{j=1}^{m} \pi_{j} \phi\left(x ; \mu_{j}, \Sigma_{j}\right)$. We know that

$$
\tilde{E}[Z]=P(g(X) \geq \gamma)=\sum_{j=1}^{m} \pi_{j} P\left(g(X) \geq \gamma \mid X \sim N\left(\mu_{j}, \Sigma_{j}\right)\right)
$$

and thus from (2.14),

$$
P(g(X) \geq \gamma) \stackrel{\text { poly }}{\sim} \exp \left\{-\min _{j=1, \ldots, m}\left(a_{j 1}-\mu_{j}\right)^{T} \Sigma_{j}^{-1}\left(a_{j 1}-\mu_{j}\right) / 2\right\} .
$$

Moreover, from (2.16),

$$
\begin{aligned}
\tilde{E}\left[Z^{2}\right] & =E[I(g(X) \geq \gamma) L(X)] \\
& =\sum_{j=1}^{m} \pi_{j} E\left[I(g(X) \geq \gamma) L(X) \mid X \sim N\left(\mu_{j}, \Sigma_{j}\right)\right] \\
& \leq \sum_{j=1}^{m} \pi_{j} r_{j} \exp \left\{-\left(a_{j 1}-\mu_{j}\right)^{T} \Sigma_{j}^{-1}\left(a_{j 1}-\mu_{j}\right) / 2\right\} P\left(g(X) \geq \gamma \mid X \sim N\left(\mu_{j}, \Sigma_{j}\right)\right) \\
\leq & \max _{j=1, \ldots, m}\left\{r_{j}\right\} \exp \left\{-\min _{j=1, \ldots, m}\left(a_{j 1}-\mu_{j}\right)^{T} \Sigma_{j}^{-1}\left(a_{j 1}-\mu_{j}\right) / 2\right\} \times \\
& \sum_{j=1}^{m} \pi_{j} P\left(g(X) \geq \gamma \mid X \sim N\left(\mu_{j}, \Sigma_{j}\right)\right) \\
= & \max _{j=1, \ldots, m}\left\{r_{j}\right\} \exp \left\{-\min _{j=1, \ldots, m}\left(a_{j 1}-\mu_{j}\right)^{T} \Sigma_{j}^{-1}\left(a_{j 1}-\mu_{j}\right) / 2\right\} P(g(X) \geq \gamma) .
\end{aligned}
$$

Combining the results, we get that $\tilde{E}\left[Z^{2}\right] / \tilde{E}[Z]^{2}$ at most grows polynomially in $\gamma$ and the IS estimator $Z$ is asymptotically optimal.

### 2.8 Conclusion

In this chapter, we study rare-event simulation problems motivated from robustness certification and safety-critical applications of intelligent physical systems, which involve rare-event boundaries associated with the predictions from machine learning models. We consider especially two common predictors, random forest and neural network, and the probability of prediction ex-
ceeding a threshold that relates to or forms a building block for the motivating applications. These problems amount to rare-event simulation with piecewise linear set boundaries that are implicitly defined. Our approach merges IS schemes based on the dominating point machinery with sequential integer programming to search for these points in a manner that caters to the geometry of these rare-event sets. We develop asymptotic optimality guarantees, and demonstrate through numerical examples the efficiency of our proposed schemes. Our study can be viewed as a first step to bridge rigorous efficiency-guaranteed rare-event simulation with the emerging applications of AI and intelligent systems. Much warranted further studies include the generalization of our approach to more sophisticated rare-event sets with intricate interaction behaviors, the handling of high-dimensional problems, and the investigation on the impacts of model errors in affecting rare-event probability estimation.

### 2.9 Supplementary A: Training Details for Random Forests

In our experiments in Section 2.6, the random forests are trained using built-in functions in MATLAB. For the regression tasks in Section 2.6.1, we use the "fitcensemle" function with default setting for training random forests. The function uses bagging (also known as bootstrap aggregating) to train decision trees and ensembles them by averaging their outputs. In particular, each time we train a decision tree, a subset of the input variables is randomly selected as the inputs for prediction and a training set is resampled from the empirical distribution of the original dataset. We use mean squared error as the criterion for branching in training a single decision tree. For the classification tasks in Section 2.6.2, we use the "fitrensemle" function, which uses boosting to ensemble decision trees trained using the Gini impurity score as a criterion for branching. The function starts with training a relatively small decision tree and then sequentially reduces the prediction error by ensembling new trees that are trained to emphasize the misclassified samples. For more details, please refer to [55, 108].

# Chapter 3: Over-Conservativeness of Variance-Based Efficiency Criteria and Probabilistic Efficiency in Rare-Event Simulation 

### 3.1 Introduction

We study the problem of estimating the probabilities of rare events with Monte Carlo simulation, which falls in the domain of rare-event simulation [22, 23, 116]. Traditionally, rare-event simulation is of wide interest to a variety of areas such as queueing systems [117, 31, 118, 26, $27,25,30,24,29]$, highly dependable computer systems and communication networks [42, 119, $120,121,122]$, financial risk management $[34,36,35]$ and insurance modeling [38, 37]. More recently, with the rapid development of intelligent physical systems such as autonomous vehicles and personal assistive robots [123, 21], rare-event simulation is also applied to assess their risks before deployments in public, where the risks are often quantified by the probabilities of violations of certain safety metrics such as crash or injury rate [18, 19, 17, 20]. The latter problems typically involve complex AI-driven underlying algorithms that deem the rare-event structures rough or difficult. The current work is motivated from the importance of handling such type of rare-event problems (e.g., the U.S. National Artificial Intelligence Research and Development Strategic Plan [124] lists "developing effective evaluation methods for AI" as a top priority) and provides a step towards rigorously grounded procedures in this direction.

The starting challenge in rare-event simulation is that, by its own nature, the target rare events seldom occur in the simulation experiment when using crude Monte Carlo. In other words, to achieve an acceptable estimation accuracy relative to the target probability, the required simulation size could be huge in order to obtain sufficient hits on the target events. Statistically, this issue is manifested as a large ratio between the standard deviation (per run) to the mean, known as the relative error, that determines the order of a required sample size. In the large deviations regime
where the target probability can depend exponentially on the rarity parameter, this in particular means the required sample size is exponentially large.

To address the inefficiency of crude Monte Carlo, a range of variance reduction techniques have been developed. Among them, importance sampling (IS) [51] has been broadly applied to improve the efficiency. IS uses an alternative probability measure to generate the simulation samples, and then reweighs the outputs via likelihood ratio to guarantee unbiasedness. The goal is that by using this alternate estimator than simply counting the frequency of hits in crude Monte Carlo, one can achieve a small relative error with a much smaller sample size.

To this end, it is also widely known that IS is a "delicate" technique, in the sense that the IS probability measure needs to be carefully chosen in order to achieve a small relative error. In the typical large deviations setting, the suggestion is to tilt the probability measure to the "important region". The delicacy appears when there are more than one important regions, in which case all of them need to be accounted for. More specifically, in the light-tailed regime, these important regions are guided precisely by the so-called dominating points, which capture the most likely scenario in a local region of the rare event. That is, each of these points (and their small neighborhoods) contributes dominantly to the probability of a certain component of the rare-event set according to a well-defined decomposition. Despite the tempting approach to simply shift the distribution center to the globally most likely scenario or the most significant dominating point, it is well established that if not all the dominating points are included in the IS mixture distribution, then the resulting estimator may no longer be efficient in terms of the relative error (see, for instance, the seminal work [103]).

Our main goal in this chapter is to argue that, in potentially many light-tailed problems, the inclusion of all the dominating points in an IS could be unnecessary. Our study is motivated from high-dimensional settings where finding all dominating points could be computationally expensive or even prohibitive, yet these problems may arise in recent safety-critical applications (e.g., [21, $14,125])$.

To intuitively explain the unnecessity, let us first drill into why all the dominating points are
arguably needed in the literature in the first place. Imagine that a rare event set $\mathcal{E}$ comprises two disjoint "important" regions, say $\mathcal{E}_{1}$ and $\mathcal{E}_{2}$, and the dominating points are correspondingly $a_{1}$ and $a_{2}$, which are sufficiently "far away" from each other, and $a_{1}$ has say a higher density than $a_{2}$. Roughly speaking, if an IS scheme only focuses on $\mathcal{E}_{1}$ and tilts the distribution center towards $a_{1}$, then there is a small chance that the sample from this IS distribution hits $\mathcal{E}_{2}$, so that the contribution from this sample in the ultimate estimator is non-zero and, moreover, may constitute a large likelihood ratio and consequently elicit a large variance. This unfortunate event of falling into a secondary important region is a source of inefficiency according to the relative error criterion.

Now let's take a step back and think about the following: How likely does the "unfortunate" event above occur? In a typical Monte Carlo experiment, we argue - and we will see clearly in experiments - that this could be very unlikely, to the extent that we shouldn't be worried at all with a reasonable simulation size. Yet, according to the relative error criterion, it seems necessary to worry about this, because it contributes to the variance of the estimator in each single run. This points to that using variance to measure efficiency in rare-event simulation could be too loose to begin with. This variance measure, in turn, comes from the Chebyshev inequality that converts relative error into a sufficient relative closeness between the estimate and target probability with high confidence. In other words, this Chebyshev inequality itself could be the source of looseness.

This motivates us to propose what we call probabilistic efficiency. Different from all the efficiency criteria in the literature, including asymptotic efficiency (also known as asymptotic optimality or logarithmic efficiency) and bounded relative error [23, 102], probabilistic efficiency does not use relative error. Instead, it is a criterion on achieving relative closeness directly. A distinctive element in probabilistic efficiency is the control on the simulation size itself, that we only allow it to grow moderately with the underlying rarity parameter. This moderate simulation size, which is often the only feasible option in experiments, suppresses the occurrence of the unfortunate event of falling into a secondary important region. This way, while the variance could blow up, the high-confidence closeness between the estimate and target probability could still be retained.

With this new framework, we show that under standard assumptions in the widely used Gartner-

Ellis large deviations regime, an IS that uses only the most significant dominating points is sufficient to ensure probabilistic efficiency. The Gartner-Ellis regime has been used across different applications such as queueing [30, 29, 27], communication systems [126, 127] and finance [128, $129,36,35]$. Our results thus stipulate that in all these problems, in order to obtain a good estimate relative to the ground-truth rare-event probability, we only need to exponentially tilt to the most significant dominating point when there is only one such point, without the use of any mixture. This is a sharp contrast to the established IS recipe. Moreover, this makes the construction of IS in closer line with the large deviations theory that governs the rare-event probability asymptotic. More specifically, in large deviations, the most significant dominating point coincides with the minimizer of the so-called rate function that controls the exponential decay rate of the probability. Our theory thus postulates that to attain probabilistic efficiency, it suffices to consider only this rate function minimizer when constructing the IS.

We close this introduction with further discussions of our study in relation to existing works. First, we contrast our probabilistic efficiency with the notion of "well estimated" in [42]. The latter asserts that all the paths which contribute most significantly to the target quantity need to occur sufficiently likely under the IS distribution. While [42] focuses on highly reliable systems, his notion of well-estimated can be properly generalized so that, in the Gartner-Ellis regime that we consider, it is conceptually similar to having an IS exponentially tilting only to the most significant dominating points. Despite this similarity, our proposed notion bears important differences with well-estimatedness. To start, the main result in [42] is an elegant one-way chain relation that links variance-based criteria, including bounded relative error and bounded normal approximation, to the well-estimatedness of the target probability and the estimation variance. This shows the sufficiency of these variance-based criteria to guarantee the natural well-estimated property of IS, and also that, due to the provable one-way nature of the chain relation, being well-estimated alone does not guarantee the attainment of these variance-based criteria. In particular, as bounded normal approximation guarantees good confidence interval performance, [42] argues the importance of looking at this criterion and cautions the use of well-estimatedness alone. In contrast, our main
message in this chapter asserts that one can achieve good probability estimation, and even construct good confidence intervals, without satisfying variance-based criteria. In particular, tilting only to the most significant dominating points, or similarly well-estimatedness in a properly generalized sense, is a sufficient condition to attain our more general notion of probabilistic efficiency which in turn guarantees these good estimation performances. Notably, arguing this stipulation requires the key idea and analysis on the small likelihood of the IS hitting the secondary rare-event regions controlled by other less significant dominating points, which turns out to be true under the GartnerEllis regime and a subexponential sample size.

Second, we caution that probabilistic efficiency is not meant to replace existing variance-based efficiency criteria, but rather to complement them especially in situations where identifying all dominating points is infeasible due to problem complexity. In problems where the latter is not an issue, it remains "safer" to use existing criteria, as probabilistic efficiency relies on a more subtle sample size requirement, namely that it is not exponentially large. While this condition is reasonable for realistic problems, it would warrant future experimental diagnostics to detect violations of such a condition.

Third, we note that our variance-free approach can potentially be adapted to rare-event estimation problems to streamline IS construction beyond the considered light-tailed large deviations regime. However, for some of these problems (e.g., heavy-tailed problems; [78, 83, 81, 33]), the efficiency gain appears less dramatic than ours which possesses an exponential speed-up. Moreover, the Gartner-Ellis paradigm that we consider in this chapter is arguably the most widely used and forms the basis of analysis for many rare-event problems.

In the following, we first introduce in more detail the background of rare-event simulation and the established efficiency criteria in the literature, all of which involve estimation variance or relative error (Section 3.2). Then we show several motivating numerical examples to illustrate how excluding some dominating points in IS appears to give similar and sometimes even better performances than including all these points, the latter suggested predominantly in the literature (Section 3.3). This motivates our new notion of probabilistic efficiency to explain the observed numerical
phenomena (Section 3.4) and the analysis of efficiency guarantees using this new notion (Section 3.5). We then show further numerical results to validate our theory and performances of our estimators (Section 3.6). Finally, we give some cautionary notes about probabilistic efficiency which involve the risk of under-estimation, and suggest some future directions (Section 3.7). Proofs of all theorems and propositions, if not directly following them, are in Section 3.11.

### 3.2 Problem Setting and Existing Framework

We consider an indexed family of rare events $\left\{\mathcal{A}_{\gamma}\right\}_{\gamma}$, where $\gamma$ denotes a "rarity parameter" such that as $\gamma \rightarrow \infty$, the event $\mathcal{A}_{\gamma}$ becomes rarer so that $P\left(\mathcal{A}_{\gamma}\right) \rightarrow 0$. Our goal is to estimate $p=p(\gamma):=P\left(\mathcal{A}_{\gamma}\right)$ using Monte Carlo simulation. Here, the index $\gamma$ is introduced for modeling purpose so that we can speak of asymptotic rate, which is customary in the rare-event simulation literature.

### 3.2.1 Crude Monte Carlo and Relative Error

To motivate the various notions that we would discuss momentarily, let us consider using crude Monte Carlo to estimate $p$. This means we utilize the unbiased estimator $Z=I_{\mathcal{A}_{\gamma}}$ for $p$, where $I_{\mathcal{A}_{\gamma}}$ denotes the indicator variable on the event $\mathcal{A}_{\gamma}$. Suppose we generate the output $Z$ independently $n$ times, and construct $\hat{p}$ as their sample mean. Intuitively, when $p$ is tiny, this estimator $\hat{p}$ is most likely zero unless $n$ is a huge number, since a long trial length is needed to land at the rare event $\mathcal{A}_{\gamma}$.

To describe the above challenge mathematically, we consider the following criterion. For a given tolerance level $\varepsilon>0$ (e.g., $5 \%$ ), we would like an estimator $\hat{p}$ to satisfy

$$
\begin{equation*}
P(|\hat{p}-p|>\delta p) \leq \varepsilon \tag{3.1}
\end{equation*}
$$

for a certain $0<\delta<1$ when using a simulation size $n$. In (3.1), the closeness between $\hat{p}$ and $p$, which represents the error of $\hat{p}$ in estimating $p$, is measured relative to the magnitude of $p$ itself.

This is because for tiny $p$, the estimation is only meaningful if the error is small enough relative to this tiny quantity. Linking to crude Monte Carlo, outputting merely $\hat{p}=0$, as likely to happen thereby, would be viewed as incurring a substantial error, i.e., lying in the event $|\hat{p}-p|>\delta p$ in (3.1). Put in another way, we argue that a huge sample size $n$ is needed for crude Monte Carlo to attain (3.1). Note that by Chebyshev's inequality, we get that for any $0<\delta<1$,

$$
\begin{equation*}
P(|\hat{p}-p|>\delta p) \leq \frac{\operatorname{Var}(Z)}{n \delta^{2} p^{2}} \tag{3.2}
\end{equation*}
$$

Hence, $n \geq \frac{\operatorname{Var}(Z)}{\varepsilon \delta^{2} p^{2}}$ implies that $P(|\hat{p}-p|>\delta p) \leq \varepsilon$ for $\varepsilon>0$. This means that $\frac{\operatorname{Var}(Z)}{\varepsilon \delta^{2} p^{2}}$ is a sufficient size for $n$ to achieve (3.1). This quantity depends on the ratio between the standard deviation $\sqrt{\operatorname{Var}(Z)}$ and the mean $p$, which is known as the relative error. Here, for crude Monte Carlo the relative error is $\frac{\sqrt{\operatorname{Var}(Z)}}{p}=\frac{\sqrt{p(1-p)}}{p}=\sqrt{\frac{1-p}{p}}$ which blows up as $p \rightarrow 0$. Consequently, the sufficient size for $n$ also blows up as $p \rightarrow 0$. In particular, if $p$ decays exponentially in $\gamma-\mathrm{a}$ typical scaling in large deviations, then the required $n$ to achieve (3.1) also scales exponentially.

### 3.2.2 Importance Sampling and Asymptotic Efficiency

The above challenge motivates variance reduction techniques to reduce the sample size requirement. Among the most popular is IS. In this approach, we generate samples from an alternate measure $\tilde{P}$ where $\tilde{P}$ satisfies $P I_{\mathcal{A}_{\gamma}} \ll \tilde{P}$ (i.e., $P I_{\mathcal{A}_{\gamma}}$ is absolutely continuous with respect to $\tilde{P}$ ), and use $Z=I_{\mathcal{A}_{\gamma}} \frac{d P}{d \tilde{P}}$ as an unbiased output for $p$, where $\frac{d P}{d \tilde{P}}$ is the Radon-Nikodym derivative, or the so-called likelihood ratio, between $P$ and $\tilde{P}$. Though this output is always unbiased thanks to the likelihood ratio adjustment, the performance of the IS estimator in terms of variability heavily depends on the choice of the IS probability measure $\tilde{P}$. In the literature, several efficiency criteria for IS estimators have been developed. A common criterion is asymptotic efficiency [54, 40, 23]:

Definition 3.1 (Asymptotic efficiency). The IS estimator $Z=I_{\mathcal{A}_{\gamma}} \frac{d P}{d \tilde{P}}$ under $\tilde{P}$ is said to achieve asymptotic efficiency if $\lim _{\gamma \rightarrow \infty} \frac{\log \left(\tilde{E}\left(Z^{2}\right)\right)}{\log p}=2$ where $\tilde{E}(\cdot)$ denotes the expectation under $\tilde{P}$.

For functions $f, g: \mathbb{R} \rightarrow \mathbb{R}$, we say $g(\gamma)$ is subexponential in $f(\gamma)$ as $\gamma \rightarrow \infty$ if $\lim _{\gamma \rightarrow \infty} \frac{\log g(\gamma)}{f(\gamma)}=$

0, i.e. $g(\gamma)=\exp (f(\gamma) o(1))$. For functions $f, g, h$, clearly if $\lim _{\gamma \rightarrow \infty} \frac{\log g(\gamma)}{f(\gamma)}=\lim _{\gamma \rightarrow \infty} \frac{\log h(\gamma)}{f(\gamma)}$, then $\lim _{\gamma \rightarrow \infty} \frac{\log (g(\gamma) / h(\gamma))}{f(\gamma)}=0$ and hence $g(\gamma) / h(\gamma)$ is subexponential in $f(\gamma)$ as $\gamma \rightarrow \infty$. By taking $g(\gamma)=\tilde{E}\left(Z^{2}\right), h(\gamma)=p^{2}$ and $f(\gamma)=\log p$, we see that the condition in Definition 3.1 is equivalent to the condition that $\tilde{E}\left(Z^{2}\right) / p^{2}\left(\right.$ or $\left.\widetilde{\operatorname{Var}}(Z) / p^{2}\right)$ is subexponential in $-\log p$ as $\gamma \rightarrow \infty$.

From (3.2) and its subsequent discussion, asymptotic efficiency implies that the required simulation size $n$ to attain a prefixed relative error grows only subexponentially in $-\log p$. As a stronger requirement, $Z$ is said to have a bounded relative error if $\lim \sup _{\gamma \rightarrow \infty} \frac{\widetilde{\operatorname{Var}}(Z)}{p^{2}}<\infty$, which implies that the required simulation size remains bounded no matter how small $p$ is. The criterion of bounded relative error is sometimes too strict to achieve, so we focus on asymptotic efficiency in this chapter. More efficiency criteria could be found in [23, 102, 64].

### 3.2.3 Large Deviations and Dominating Points

In the large deviations setting, the classical notion of dominating points is used to guarantee asymptotic efficiency of IS [53]. To explain, let us first recall the so-called rate function in the large deviations theory which, intuitively speaking, measures the likelihood of hitting each point on an exponential scale. More specifically, we consider the standard Gartner-Ellis regime as follows [130, 22]. Without loss of generality consider $\gamma>0$. Suppose that $\mathcal{A}_{\gamma}=\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right\}$ where $\left\{X_{\gamma}\right\}_{\gamma}$ are $\mathbb{R}^{d}$-valued random variables and $\mathcal{E}$ is a fixed Borel set in $\mathbb{R}^{d}$. Throughout this chapter, we use $A^{\top}$ to denote the transpose of any matrix $A$, and use $\mathcal{E}^{\circ}, \overline{\mathcal{E}}, \partial \mathcal{E}$ to denote the interior, closure and boundary of any set $\mathcal{E}$.

We define $\mu_{\gamma}(x)=\frac{1}{\gamma} \log E\left(e^{x^{\top} X_{\gamma}}\right), x \in \mathbb{R}^{d}$ as the scaled logarithmic moment generating function. We denote $\mathcal{D}(f)=\{x: f(x)<\infty\}$ as the domain of a function $f$. With these, we assume the following:

Assumption 3.1. $\mu_{\gamma}(x)$ satisfies the following conditions:

1. $\mu(x)=\lim _{\gamma \rightarrow \infty} \mu_{\gamma}(x)$ exists for any $x \in \mathbb{R}^{d}$, where we allow $\infty$ both as a limit value and as an element of the sequence $\left\{\mu_{\gamma}(x)\right\}$;
2. $0 \in \mathcal{D}(\mu)^{\circ}$;
3. $\mu$ is essentially smooth, i.e., $\mathcal{D}(\mu)^{\circ}$ is non-empty, $\mu$ is differentiable everywhere in $\mathcal{D}(\mu)^{\circ}$ and $\mu$ is steep.

Then we define the rate function $I(y)=\sup _{x \in \mathbb{R}^{d}}\left\{x^{\top} y-\mu(x)\right\}, y \in \mathbb{R}^{d}$ as the Legendre transform of $\mu$. For any set $\mathcal{E} \subset \mathbb{R}^{d}$, we denote $I(\mathcal{E})=\inf _{y \in \mathcal{E}} I(y)$. We make the following assumptions for the set $\mathcal{E}$ :

Assumption 3.2. $\mathcal{E} \subset \mathbb{R}^{d}$ is a Borel set such that $\overline{\mathcal{E}}=\overline{\mathcal{E}^{\circ}}, \mathcal{E}^{\circ} \cap \mathcal{D}(I)^{\circ} \neq \emptyset$ and $I(\mathcal{E})>0$.

Under these assumptions, we have the following result (e.g., adapted from [130] Theorem 3.3.3.6):

Theorem 3.1 (Gartner-Ellis Theorem). Suppose that Assumption 3.1 holds. For any Borel set $\mathcal{E} \subset \mathbb{R}^{d}$, we have that

$$
-I\left(\mathcal{E}^{\circ}\right) \leq \liminf _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) \leq \limsup _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) \leq-I(\overline{\mathcal{E}})
$$

If additionally Assumption 3.2 holds, then

$$
\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)=-I(\mathcal{E})
$$

Assumptions 3.1 and 3.2 are standard light-tailed conditions on $\mathcal{A}_{\gamma}$, which guarantee the considered probability $P\left(\mathcal{A}_{\gamma}\right)$ to decay exponentially in $\gamma$ with decay rate $I(\mathcal{E})$. The rate function $I(y)$ can be viewed as a measurement on the likelihood of hitting $y$ in the exponential scale. The most likely point to hit among $\mathcal{E}$ is hence given by the minimizer of $I(y)$ over $\mathcal{E}$, resulting in the overall exponential decay rate $I(\mathcal{E})$. Note that, by Theorem 3.1, we have $-\log p=\Theta(\gamma)$ as $\gamma \rightarrow \infty$, and hence subexponential (or exponential) in $-\log p$ is equivalent to subexponential (or exponential) in $\gamma$.

Now we present the concept of dominating points and sets:

Definition 3.2 (Dominating Set). Suppose that Assumptions 3.1 and 3.2 hold. We call $A=$ $\left\{a_{1}, \ldots, a_{r}\right\} \subset \partial \mathcal{E}$ a dominating set for $\mathcal{E}$ if

1. For each $i, a_{i} \in \mathcal{D}(I)^{\circ}$ and there exists a unique $s_{a_{i}} \in \mathbb{R}^{d}$ such that $\nabla \mu\left(s_{a_{i}}\right)=a_{i}$;
2. $\mathcal{E} \subset \bigcup_{i=1}^{r}\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$;
3. For any $i, A \backslash\left\{a_{i}\right\}$ does not satisfy Condition 2.

We call any point in A a dominating point. For two dominating points a and $a^{\prime}$, we say a is more significant than $a^{\prime}$ if $I(a)<I\left(a^{\prime}\right)$.


Figure 3.1: Illustration of rare-event set $\mathcal{E}=\mathcal{E}_{1} \cup \mathcal{E}_{2}$ and dominating points.

Dominating points can be understood as the "local minimizers" of the rate function $I$ in the sense that $a_{i}$ is the minimizer of $I$ in $\mathcal{E} \cap\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$. To understand this, first, Condition 1 in Definition 3.2 stipulates that $s_{a_{i}}$ is the gradient of $I(x)$ at the point $a_{i}$. Then, in Condition 2, $a_{i}$ can be seen as the minimizer of $I(x)$ over the set $\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$, a fact that follows from the first-order optimality condition in convex function minimization. Thus, Condition 2 stipulates that any point in the rare-event set $\mathcal{E}$ has an $I$ value not less than one of the $a_{i}$ 's. Geometrically, any points in the set $\mathcal{E}$ must lie in the half-space tangentially cut by one of the $a_{i}$ 's (i.e., the "backyard" of the $a_{i}$ ). Figure 3.1 is an illustration of a rare-event set $\mathcal{E}=\mathcal{E}_{1} \cup \mathcal{E}_{2}$ and the dominating set $\left\{a_{1}, a_{2}\right\}$. Here, $a_{1}$ is the global minimum rate point in $\mathcal{E}$, but $\mathcal{E}_{2}$ is not covered by $\left\{x: s_{a_{1}}^{\top}\left(x-a_{1}\right) \geq 0\right\}$, so $a_{2}$ is included in the dominating set. Finally, we note that Condition 3 in Definition 3.2 enforces the dominating set to be the minimal set of points such that the geometric properties in Conditions 1 and 2 are satisfied.

Note that dominating points may not be local minimizers of the rate function $I$ in $\mathcal{E}$ (even though they are minimizers in $\mathcal{E} \cap\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$ as discussed above). Nonetheless, the most significant dominating points are indeed global minimizers of $I$ in $\mathcal{E}$. This is presented in the following theorem.

Theorem 3.2. Suppose that Assumptions 3.1 and 3.2 hold, and $A=\left\{a_{1}, \ldots, a_{r}\right\}$ is a dominating set for $\mathcal{E}$. Then $I(\mathcal{E})=\min _{i=1, \ldots, r} I\left(a_{i}\right)$. That is to say, the most significant dominating points are global minimizers of I in $\mathcal{E}$.

We should also point out that dominating set defined according to Definition 3.2 may not be unique. Advantageously, the theory and estimators we present will flexibly apply to any such dominating set.

### 3.2.4 Asymptotically (In)efficient Importance Samplers

We are now ready to describe the main message of this section, which is the established recipe in constructing efficient IS. The standard proposal is to use a mixture of exponentially tilted distributions, where each exponential tilting is with respect to each dominating point. In particular, suppose that $A=\left\{a_{1}, \ldots, a_{r}\right\}$ is a dominating set. Then the IS distribution is $\tilde{P}$ such that

$$
\begin{equation*}
\frac{d \tilde{P}}{d P}=\sum_{i=1}^{r} \alpha_{i} e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)} \tag{3.3}
\end{equation*}
$$

with $\sum_{i=1}^{r} \alpha_{i}=1, \alpha_{i}>0, \forall i$. Here, $e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)} d P$ is the exponential tilting towards the dominating point $a_{i}$ and $\alpha_{i}$ 's are the mixing weights. The IS (3.3) is well known to be asymptotically efficient:

Proposition 3.1 (Mixture IS is asymptotically efficient). Suppose Assumptions 3.1 and 3.2 hold, and the dominating set has finite cardinality, i.e., $r<\infty$. Then the IS distribution (3.3) with any fixed $\alpha_{i}$ 's is asymptotically efficient.

While the proof of Proposition 3.1 is standard, we include it in Section 3.11 for self-containedness. Here, we describe the key intuition in justifying the necessity of mixture. First, the likelihood ratio
in the considered mixture IS is

$$
L=\frac{d P}{d \tilde{P}}=\frac{1}{\sum_{i=1}^{r} \alpha_{i} e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)}}
$$

and it satisfies that for any $i$,

$$
\begin{equation*}
L \leq \frac{1}{\alpha_{i} e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)}}=\frac{1}{\alpha_{i}} e^{-s_{a_{i}}^{\top}\left(X_{\gamma}-\gamma a_{i}\right)-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)} . \tag{3.4}
\end{equation*}
$$

In the exponent in the rightmost expression of (3.4), the second term $\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)$ is approximately $\gamma I\left(a_{i}\right)$, and the first term is the "overshoot" of the sampled $X_{\gamma}$ compared to the dominating point $a_{i}$. That is, if $X_{\gamma}$ is in the "backyard" of $a_{i}$, then this term $s_{a_{i}}^{\top}\left(X_{\gamma}-\gamma a_{i}\right) \geq 0$. The definition of dominating set, especially Condition 2 in Definition 3.2, guarantees any $(1 / \gamma) X_{\gamma}$ in $\mathcal{E}$ must have $s_{a_{i}}^{\top}\left(X_{\gamma}-\gamma a_{i}\right) \geq 0$ for at least one of the $i$ 's. Thus, by decomposing the second moment of $Z=I_{\mathcal{A}_{\gamma}} L$ according to the backyards of $a_{i}$ 's, we can ensure that the magnitude of the likelihood ratio, when $X_{\gamma}$ lies inside the rare event set, is properly controlled. More precisely, write $\mathcal{E}=\bigcup_{i=1}^{r} \mathcal{E}_{i}$ where each $\mathcal{E}_{i} \subset\left\{x: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$. Then the second moment of the IS satisfies

$$
\tilde{E}\left[I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L^{2}\right] \leq \sum_{i=1}^{r} \tilde{E}\left[I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{i}\right) L^{2}\right] \leq \sum_{i=1}^{r} \tilde{E}\left[I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{i}\right) \frac{1}{\alpha_{i}^{2}} e^{-2 \gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right]
$$

which is approximately bounded by $\sum_{i=1}^{r} e^{-2 \gamma I\left(a_{i}\right)} / \alpha_{i}^{2}$ and hence $e^{-2 \gamma I(\mathcal{E})}$ in the exponential scale, thus verifying asymptotic efficiency.

On the other hand, if we miss some dominating points in the construction of the mixture IS, then asymptotic efficiency may fail to be attained. Below we give a simple example to demonstrate this.

Proposition 3.2 (Missed dominating point leads to violation of asymptotic efficiency). Suppose that we want to estimate $p=P\left(\frac{1}{\gamma} X_{\gamma} \in(-\infty,-2] \cup[1, \infty)\right)$ where $X_{\gamma} \sim N(0, \gamma)$ under $P$. If the IS distribution is chosen as $X_{\gamma} \sim N(\gamma, \gamma)$, then $\tilde{E}\left(Z^{2}\right) / p^{2}=\Theta\left(\sqrt{\gamma} e^{3 \gamma / 2}\right)$ grows exponentially in $-\log p=\Theta(\gamma)$, and hence $Z$ is not asymptotically efficient by definition.

In this example, the dominating points are 1 and -2 , and $N(\gamma, \gamma)$ is the exponential tilt towards the first dominating point 1 (for Gaussian distribution, exponential tilting amounts to a mean shift). Here, by considering only this point, it is possible that a generated $X_{\gamma}$ satisfies $(1 / \gamma) X_{\gamma} \in(-\infty,-2]$ while the overshoot $s_{1}^{\top}\left(X_{\gamma}-\gamma 1\right)$, as explained for (3.4), takes a very negative value. This scenario contributes significantly to the overall variance and ultimately violates asymptotic efficiency.

Our main insight in this chapter is a rebuke of the above viewpoint. More specifically, we argue that missing inferior dominating point, such as the example in Proposition 3.2, can still result in a good IS according to our beginning criterion (3.1). The reason why we are interested in studying such an assertion is that for some problems, especially those involving modern large-scale models, it is computationally challenging to locate all the dominating points. A core ingredient of our assertion is to question the use of asymptotic efficiency, or more generally variance-based efficiency criteria. Before delving into the theory, let us first present some numerical results to shed light on how much difference it makes to use different numbers of dominating points in the IS mixture. This is the focus of our next section.

### 3.3 Motivating Experimental Results

We run three numerical examples to demonstrate that missing dominating points in IS construction, while provably leads to asymptotic inefficiency, could perform well empirically. This thus suggests an inadequacy in using asymptotic efficiency, or more generally variance-based criteria, to measure the performances of rare-event estimators. Besides, the computationally demanding example in Section 3.3.3 justifies the motivation why we seek to reduce the number of used dominating points in the IS mixture.

### 3.3.1 Large Deviations of an I.I.D. Sum

We consider the problem of estimating the tail probability involving a sum of random variables, where $Y_{1}, Y_{2}, \ldots$ are i.i.d and we are interested in

$$
P\left(\left|S_{m}\right| \geq a m\right)
$$

where $S_{m}=\sum_{i=1}^{m} Y_{i}$. We consider $m$ as the rarity parameter $\gamma$ presented in Section 3.2. Using the notation in the Gartner-Ellis regime, we have $X_{\gamma}=S_{m}$ and $\mathcal{E}=(-\infty,-a] \cup[a, \infty)$. Then $\mu(x)=\log E\left(e^{x Y_{1}}\right)$ and we suppose Assumption 3.1 is satisfied. By Theorem 3.1, when $\left|E Y_{1}\right|<a$, if $s_{a}$ and $s_{-a}$ satisfy $\nabla \mu\left(s_{a}\right)=a$ and $\nabla \mu\left(s_{-a}\right)=-a$, then we have

$$
\begin{equation*}
-\lim _{m \rightarrow \infty} \frac{1}{m} \log P\left(S_{m} \geq a m\right)=s_{a} a-\mu\left(s_{a}\right)=I(a) \tag{3.5}
\end{equation*}
$$

and

$$
\begin{equation*}
-\lim _{m \rightarrow \infty} \frac{1}{m} \log P\left(S_{m} \leq-a m\right)=-s_{-a} a-\mu\left(s_{-a}\right)=I(-a) . \tag{3.6}
\end{equation*}
$$

For this problem, [103] Section 3 provides two estimators, $\hat{\alpha}(m)$ and $\hat{\beta}(m)$. Specifically, we have $\hat{\alpha}(m)=\exp \left(-s_{a} S_{m}+m \mu\left(s_{a}\right)\right) I_{\left\{\left|S_{m}\right| \geq a m\right\}}$ with samples of $Y_{i}$ generated from exponentially tilted distribution using $s_{a}$. The estimator $\hat{\beta}(m)=\exp \left(-s_{a} S_{m}+m \mu\left(s_{a}\right)\right) I_{\left\{S_{m} \geq a m\right\}}+\exp \left(-s_{-a} S_{m}^{\prime}+\right.$ $\left.m \mu\left(s_{-a}\right)\right) I_{\left\{S_{m}^{\prime} \leq-a m\right\}}$ with $S_{m}$ and $S_{m}^{\prime}$ constructed from independent sequences of i.i.d. $Y_{i}$ 's and $Y_{i}^{\prime}$ 's generated from exponentially tilted distributions using $s_{a}$ and and $s_{-a}$ respectively. That is, $\hat{\beta}(m)$ attempts to estimate $P\left(S_{m} \geq a m\right)$ and $P\left(S_{m} \leq-a m\right)$ separately using different IS samples and sum up these estimates. Here, $\hat{\alpha}(m)$ only uses one dominating point whereas $\hat{\beta}(m)$ uses both points (note that even though $\hat{\beta}(m)$ does not use the mixture IS scheme in (3.3), the idea is similar in that it accounts for both dominating points). In our experiment, we follow [103] to set $Y_{1}=A-B$ with $A \sim N(1.5,1), B \sim \operatorname{Exp}(1)$ and $A, B$ independent, and $a=1.5$ (in this case, $s_{a}=(\sqrt{5}-1) / 2$, $s_{-a}=-2+\sqrt{2}$, and $\left.I(a) \approx 0.2902, I(-a) \approx 0.7044\right)$.

We run numerical experiments with $m=10,30,50$ and 100 . The results using $10^{4}$ samples
are shown in Table 3.1. By comparing the numbers in the second and third rows, we observe that $\hat{\alpha}(m)$ and $\hat{\beta}(m)$ have very similar empirical performances. However, note that:

Proposition 3.3. Under the problem specification above, $\hat{\beta}(m)$ is asymptotically efficient while $\hat{\alpha}(m)$ is not. In fact, $\tilde{E}\left(\hat{\alpha}^{2}(m)\right) \rightarrow \infty$ as $m \rightarrow \infty$ where $\tilde{E}$ denotes the expectation under the exponential tilting towards $a$.

In view of Proposition 3.3, $\hat{\alpha}(m)$ is arguably a very poor estimator as it bears an exploding variance. We therefore see an apparent discrepancy between empirical performances and theoretical guidance - The theoretically bad variance does not result in poor empirical performances. Proposition 3.3 is proved in [103], where the asymptotic efficiency of $\hat{\beta}(m)$ follows from their Proposition 1, while the variance behavior of $\hat{\alpha}(m)$ appears in their Theorem 1.

Table 3.1: Point estimates (and 95\% CI) using IS estimators for the tail probability with different $m$.

| $m$ | 10 | 30 | 50 | 100 |
| :--- | :--- | :--- | :--- | :--- |
| $\hat{\alpha}(m)$ | $8.22( \pm 0.26) \times 10^{-3}$ | $1.60( \pm 0.07) \times 10^{-5}$ | $3.77( \pm 0.18) \times 10^{-8}$ | $1.34( \pm 0.08) \times 10^{-14}$ |
| $\hat{\beta}(m)$ | $8.29( \pm 0.26) \times 10^{-3}$ | $1.60( \pm 0.07) \times 10^{-5}$ | $3.77( \pm 0.18) \times 10^{-8}$ | $1.34( \pm 0.08) \times 10^{-14}$ |

Finally, we note that the confidence intervals presented in Table 3.1 are constructed based on standard normality intervals which utilize sample variances. As we demonstrate later in Theorem 3.8 along with practical validations in Section 3.6.2, even though the estimation variance explodes in this example, such confidence intervals are in fact asymptotically valid.

### 3.3.2 Overshoot Probability of Random Walk

We consider the problem of estimating the overshoot probability of the finite-horizon maximum of a random walk. We define the probability of interest as

$$
p=P\left(\max _{m=1, \ldots, d} S_{m} \geq a\right)
$$

where $S_{m}=\sum_{i=1}^{m} Y_{i}$ and $Y_{i}$ 's are Gaussian distributed with mean 0 , standard deviation $\sigma$, and pairwise correlation -0.02 , i.e., $\operatorname{corr}\left(Y_{i}, Y_{j}\right)=-0.02$ for any $i, j \in\{1, \ldots, d\}$ with $i \neq j$. We
use $\Sigma$ to denote the covariance matrix of $\left(Y_{1}, \ldots, Y_{m}\right)$. Suppose that the rarity parameter is $\gamma=$ $1 / \sigma^{2} \rightarrow \infty$. We note that we can reformulate this target rare event as $\left\{\frac{1}{\gamma} X_{\gamma} \in \bigcup_{m=1}^{d} \mathcal{H}_{m}\right\}$ where $X_{\gamma}=\gamma\left(Y_{1}, \ldots, Y_{d}\right)^{\top}$ and $\mathcal{H}_{m}=\left\{x \in \mathbb{R}^{d}: \sum_{i=1}^{m} x_{i} \geq a\right\}$ with $x_{i}$ denoting the $i$ th element in $x$. This decomposition allows us to construct an IS estimator using the $d$ dominating points corresponding to each half-space $\mathcal{H}_{m}$. More specifically, in this example, $I(y)=\frac{1}{2} y^{\top} \Sigma^{-1} y$, and the dominating points, ranking from the most to the least significant (i.e., increasing rate function value), are $a_{1}=\frac{a \Sigma e_{d}}{e_{d}^{\tau} \Sigma e_{d}}, a_{2}=\frac{a \Sigma e_{d-1}}{e_{d-1}^{T} \Sigma e_{d-1}}, \ldots, a_{d}=\frac{a \Sigma e_{1}}{e_{1}^{\tau} \Sigma e_{1}}$, where $e_{i}$ denotes the vector with 1 in the first $i$ elements and 0 for the rest.

In our experiments, we fix $a$ and vary $\sigma$ for different rarity levels. In addition, we set $d=10$. We generate $10^{4}$ samples from IS distributions using a varying, partial list of dominating points. That is, we choose the IS distributions for $X_{\gamma}$ as $\frac{1}{k} \sum_{i=1}^{k} \phi\left(x ; \gamma a_{i}, \gamma \Sigma\right)$, where $\phi(x ; a, b)$ denotes the Gaussian density with mean $a$ and variance $b$, for $k=1, \ldots, d$. The performances of these IS estimators are shown in Table 3.2. We observe that these estimators with different numbers of dominating points all perform similarly. In particular, we present the cases with $\sigma=0.2$ and $\sigma=0.3$ respectively in Figure 3.2. We observe that in both cases the performances of the IS estimators are almost independent of the number of used dominating points, with the probability estimates all comparable while using more dominating points slightly increases the CI width.

On the other hand, Proposition 3.1 implies that the IS using all dominating points is asymptotically efficient while we have:

Proposition 3.4. Under the problem specification above, the IS estimator that exponentially tilts towards the most significant dominating point $a_{1}$, i.e., $X_{\gamma}$ distributed as $N\left(\gamma a_{1}, \gamma \Sigma\right)$, is not asymptotically efficient.

Thus, like in Section 3.3.1, there appears a mismatch between theoretical guidance and empirical observation. The asymptotic inefficiency of simple exponential tilting towards only the most significant dominating point does not result in a poor experimental performance.


Figure 3.2: Simulation results for the random walk experiment. Point estimates and CI widths for IS estimators using different numbers of dominating points.

Table 3.2: Point estimates (and 95\% CIs) from IS estimators using different numbers of dominating points for the overshoot probability. \# denotes the number of dominating points, starting from the most significant one, used in the IS estimator.

| $\sigma$ | 0.2 | 0.22 | 0.24 | 0.26 | 0.28 | 0.3 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $\#$ | prob (with CI) | prob (with CI) | prob $($ with CI) | prob (with CI) | prob (with CI) | prob (with CI) |
| 1 | $9.15( \pm 0.52) \times 10^{-8}$ | $1.24( \pm 0.19) \times 10^{-6}$ | $7.96( \pm 0.85) \times 10^{-6}$ | $3.52( \pm 0.29) \times 10^{-5}$ | $1.15( \pm 0.08) \times 10^{-4}$ | $3.11( \pm 0.20) \times 10^{-4}$ |
| 2 | $9.63( \pm 0.52) \times 10^{-8}$ | $1.24( \pm 0.08) \times 10^{-6}$ | $8.38( \pm 0.52) \times 10^{-6}$ | $3.71( \pm 0.20) \times 10^{-5}$ | $1.23( \pm 0.06) \times 10^{-4}$ | $3.31( \pm 0.16) \times 10^{-4}$ |
| 3 | $9.36( \pm 0.60) \times 10^{-8}$ | $1.15( \pm 0.07) \times 10^{-6}$ | $7.87( \pm 0.43) \times 10^{-6}$ | $3.61( \pm 0.19) \times 10^{-5}$ | $1.21( \pm 0.06) \times 10^{-4}$ | $3.21( \pm 0.15) \times 10^{-4}$ |
| 4 | $9.69( \pm 0.72) \times 10^{-8}$ | $1.18( \pm 0.08) \times 10^{-6}$ | $7.98( \pm 0.49) \times 10^{-6}$ | $3.64( \pm 0.20) \times 10^{-5}$ | $1.21( \pm 0.06) \times 10^{-4}$ | $3.25( \pm 0.16) \times 10^{-4}$ |
| 5 | $9.68( \pm 0.80) \times 10^{-8}$ | $1.17( \pm 0.09) \times 10^{-6}$ | $7.92( \pm 0.54) \times 10^{-6}$ | $3.59( \pm 0.22) \times 10^{-5}$ | $1.20( \pm 0.07) \times 10^{-4}$ | $3.20( \pm 0.17) \times 10^{-4}$ |
| 6 | $9.50( \pm 0.89) \times 10^{-8}$ | $1.15( \pm 0.10) \times 10^{-6}$ | $7.79( \pm 0.60) \times 10^{-6}$ | $3.55( \pm 0.25) \times 10^{-5}$ | $1.19( \pm 0.08) \times 10^{-4}$ | $3.17( \pm 0.19) \times 10^{-4}$ |
| 7 | $9.89( \pm 0.97) \times 10^{-8}$ | $1.20( \pm 0.11) \times 10^{-6}$ | $8.13( \pm 0.66) \times 10^{-6}$ | $3.69( \pm 0.27) \times 10^{-5}$ | $1.22( \pm 0.08) \times 10^{-4}$ | $3.26( \pm 0.21) \times 10^{-4}$ |
| 8 | $9.44( \pm 1.02) \times 10^{-8}$ | $1.16( \pm 0.11) \times 10^{-6}$ | $7.93( \pm 0.70) \times 10^{-6}$ | $3.62( \pm 0.29) \times 10^{-5}$ | $1.20( \pm 0.09) \times 10^{-4}$ | $3.21( \pm 0.22) \times 10^{-4}$ |
| 9 | $8.97( \pm 1.03) \times 10^{-8}$ | $1.11( \pm 0.12) \times 10^{-6}$ | $7.63( \pm 0.72) \times 10^{-6}$ | $3.48( \pm 0.30) \times 10^{-5}$ | $1.16( \pm 0.09) \times 10^{-4}$ | $3.11( \pm 0.23) \times 10^{-4}$ |
| 10 | $9.23( \pm 1.12) \times 10^{-8}$ | $1.15( \pm 0.13) \times 10^{-6}$ | $7.87( \pm 0.77) \times 10^{-6}$ | $3.55( \pm 0.32) \times 10^{-5}$ | $1.19( \pm 0.10) \times 10^{-4}$ | $3.20( \pm 0.25) \times 10^{-4}$ |

### 3.3.3 Robustness Assessment for an MNIST Classification Model

We consider a rare-event probability estimation problem from an image classification task. Our goal is to estimate the probability of misclassification when the input of a prediction model is perturbed by tiny noise. This probability estimate is of interest as a robustness measure of the prediction model [14]. More specifically, suppose that the prediction model $g$ is able to predict the label of input $x_{0}$, i.e. $g\left(x_{0}\right)=c$ where $c$ is the true label of $x_{0}$. Then $P\left(g\left(x_{0}+\varepsilon\right) \neq c\right)$ where $\varepsilon$ is a random perturbation can be used to measure the robustness of $g$.

In particular, we consider the classification problem on MNIST dataset which contains 70,000
images of handwritten digits and each image consists of $28 \times 28$ pixels. We train a 2-ReLUlayer neural network with 20 neurons in each layer using 60,000 training data, which achieves approximately $95 \%$ of testing data accuracy in predicting the digits. We perturb a fixed input (that is correctly predicted) with a Gaussian noise with mean 0 and standard deviation $\sigma$ on each of the 784 dimensions to assess the robustness of the prediction. Note that the rarity of this problem is determined by the value of $\sigma$, and we let the rarity parameter $\gamma=1 / \sigma^{2} \rightarrow \infty$. The target rare event can be reformulated as $\left.\left\{\frac{1}{\gamma} X_{\gamma} \in\left\{x: g\left(x_{0}+x\right) \neq c\right)\right\}\right\}$ where $X_{\gamma}=\gamma \varepsilon$.

We apply mixtures of exponential tiltings as IS estimators for this problem, namely by considering the IS distribution $\frac{1}{k} \sum_{i=1}^{k} \phi\left(x ; \gamma a_{i}, \gamma I\right)$, where $\phi(x ; a, b)$ denotes the Gaussian density with mean $a$ and variance $b$ as in Section 3.3.2, for $k=1,2, \ldots$. Here $a_{i}, i=1,2, \ldots$ denote the dominating points. In order to compute these points, we apply the scheme introduced in [131] and [125], which sequentially searches for dominating points by minimizing the rate function on the rare-event set that excludes the half-spaces cut from previous more significant dominating points. In the Gaussian case with piecewise linear rare-event set boundary as in our current example, each iteration amounts to finding the highest-density point on a piecewise-linear-boundary set, which can be conducted using mixed integer programming (see Algorithm 3.1 in Section 3.8). Due to the high dimensionality of the input space and the complexity of the neural network predictor, the number of dominating points in this problem is huge. We implemented this sequential searching algorithm and it took a week to find the first 100 dominating points. Since we stopped the algorithm prematurely, the actual number of dominating points can be much larger. We run IS distributions with different numbers of dominating points (ranging from 1 to 41) and magnitudes of $\sigma$ (ranging from 0.1 to 0.2 ) and report the estimated probabilities and CIs. We use $10^{5}$ samples for IS estimators and $10^{7}$ samples for crude Monte Carlo estimators.

Figures 3.3a and 3.3b show the results. Missing less significant dominating points does not seem to make noticeable differences in this problem. As shown in Figure 3.3a, when we fix the rarity of the problem, the estimate is not sensitive to the number of dominating points. The CI width has an increasing trend as the number of dominating points gets larger, indicating that additional


Figure 3.3: Simulation results for the MNIST experiment. (a) Point estimates and CI widths from IS estimators using different numbers of dominating points. (b) Point estimates from IS estimators using different numbers of dominating points (IS with 1, 20, and 40) and crude Monte Carlo (CMC), with vertical error bars representing their $95 \%$ CIs (the CIs for the IS estimates are extremely narrow).
dominating points can in fact even hurt performances.
In Figure 3.3b, we vary the rarity of the problem and compare the performances of different IS estimators and crude Monte Carlo. Note that estimates using crude Monte Carlo are unavailable for rarer configurations due to its inefficiency. We observe that the estimates from different IS estimators overlap visually in all considered cases, which indicates that the differences among these estimates are negligible. We also note that these estimates are consistent with the crude Monte Carlo estimates (when available), which shows their correctness.

Nonetheless, once again we have an apparent mismatch between theoretical inefficiency and good empirical performances:

Proposition 3.5. Under the problem specification above, the IS estimator that exponentially tilts towards the most significant dominating point, i.e., $X_{\gamma}$ distributed as $N\left(\gamma a_{1}, \gamma I\right)$, is not asymptotically efficient.

We note that in this example, locating all the dominating points for mixture IS construction is computationally very challenging which, connecting to our discussions in Sections 3.1 and 3.2, motivates us to study the validity of IS schemes without using all dominating points. More pre-
cisely, in this example, the total number of dominating points is unknown, finding all of them requires substantial computation time, and finding the less significant dominating points requires more computation efforts than finding the more significant ones. In particular, the computation time is 20 seconds for the most significant dominating point, and increases to several hours for the 90 -th dominating point. This is because obtaining each consecutive dominating point requires iteratively adding new constraints to the large integer program. We should also note that these runtimes are for the considered particular problem instance, and for larger problems the runtimes would be even longer.

### 3.4 Probabilistic Efficiency

Section 3.3 shows that IS estimators that miss some dominating points could perform competitively compared to estimators that consider all of them, thus suggesting a gap between the notion of asymptotic efficiency and empirical performances. In light of this, we propose the concept of probabilistic efficiency as a relaxation of asymptotic efficiency. The key of probabilistic efficiency is to consider the high-probability relative discrepancy of the estimator from the ground truth directly, instead of using the relative error or equivalently the estimation variance. The latter, as can be seen in the arguments in Section 3.2, provides a sufficient, but not necessary, condition on the required sample size. In particular, there is an intrinsic looseness brought by the Markov or Chebyshev inequality (3.2) that converts relative error into the required sample size.

To proceed, we first define the following:

Definition 3.3 (Minimal relative discrepancy). For any estimator $\hat{p}$ of $p$ and any $\varepsilon>0$, the minimal relative discrepancy of $\hat{p}$, at tolerance level $\varepsilon$, is given by

$$
\begin{equation*}
\delta_{\varepsilon}(\hat{p}, p):=\inf \{\delta>0: \tilde{P}(|\hat{p}-p|>\delta p) \leq \varepsilon\} . \tag{3.7}
\end{equation*}
$$

The minimal relative discrepancy $\delta_{\varepsilon}(\hat{p}, p)$ measures the relative accuracy of the estimator $\hat{p}$, in that it gives the smallest relative discrepancy of $\hat{p}$ from $p$ that can be achieved with probability
$1-\varepsilon$. Thus the smaller is $\delta_{\varepsilon}(\hat{p}, p)$, the more accurate is $\hat{p}$. Note that in (3.7), the probability $\tilde{P}$ is the one generating the estimator $\hat{p}$.

We say that $\hat{p}$ is probabilistically efficient if $\delta_{\varepsilon}(\hat{p}, p)$ can be made small in some sense, without needing to use a gigantic amount of computation. More precisely, we propose the following notions:

Definition 3.4 (Probabilistic Efficiency). Suppose that $\left\{\mathcal{A}_{\gamma}\right\}_{\gamma}$ is an indexed family of rare events and $p=P\left(\mathcal{A}_{\gamma}\right) \rightarrow 0$ as $\gamma \rightarrow \infty$. Consider an estimator $\hat{p}$ obtained from $n=n(\gamma)$ independent replications of $Z$. For any $\varepsilon>0$, we define $\delta_{\varepsilon}(\hat{p}, p)$ as in (3.7). Then

1. We call $Z$ strongly probabilistically efficient if we can choose $n$ subexponential in $-\log p$ such that, for any $\varepsilon>0, \lim _{\gamma \rightarrow \infty} \delta_{\varepsilon}(\hat{p}, p)=0 ;$
2. We call $Z$ weakly probabilistically efficient if we can choose $n$ subexponential in $-\log p$ such that, for any $\varepsilon>0, \lim \sup _{\gamma \rightarrow \infty} \delta_{\varepsilon}(\hat{p}, p)<1$.

Note that strong probabilistic efficiency matches the usual notion in statistical estimation. That is, the estimator approaches the target parameter as $\gamma \rightarrow \infty$. In contrast, weak probabilistic efficiency only cares about a correct magnitude. While this may appear less desirable, in rare-event estimation a correct magnitude can be viewed as sufficient as the target quantity is very small, and this weaker notion allows more flexibility in constructing estimators. We also contrast our proposed probabilistic efficiency with a notion named probabilistic bounded relative error proposed in [132], where the IS measure is randomly chosen and efficiency is achieved if the resulting random relative error of the IS estimator is bounded by some constant with high probability, which is conceptually different from our notion.

The following shows that probabilistic efficiency is a relaxation of asymptotic efficiency:

Proposition 3.6. If $Z$ is asymptotically efficient, then $Z$ is strongly probabilistically efficient.

Proof of Proposition 3.6. For any unbiased estimator $Z$, we have that for any $\varepsilon>0$,

$$
\tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{\widetilde{\operatorname{Var}}(Z)}{\varepsilon n p^{2}}} p\right) \leq \varepsilon
$$

and hence by definition,

$$
\delta_{\varepsilon}(\hat{p}, p) \leq \sqrt{\frac{\widetilde{\operatorname{Var}}(Z)}{\varepsilon n p^{2}}} .
$$

If $Z$ is asymptotically efficient, $\frac{\widehat{\operatorname{Var}}(Z)}{p^{2}}$ grows at most subexponentially in $-\log p$, so we could choose $n$ subexponentially growing in $-\log p$ such that $\lim _{\gamma \rightarrow \infty} \delta_{\varepsilon}(\hat{p}, p)=0$ for any $\varepsilon>0$. By definition, $Z$ is strongly probabilistically efficient.

While asymptotic efficiency implies strong probabilistic efficiency, we note that these two notions are not equivalent. In Section 3.3, Propositions 3.3-3.5 show that asymptotic efficiency does not hold for the considered IS estimators in all the presented examples, but Theorem 3.10 in Section 3.6.1 will show that strong probabilistic efficiency actually holds for all of them.

Now we explain how probabilistic efficiency helps us understand the influence of missing some dominating points. Recall the example where $\mathcal{A}_{\gamma}=\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right\}$ and $\mathcal{E}$ comprises two disjoint and faraway pieces $\mathcal{E}_{1}$ and $\mathcal{E}_{2}$. The dominating points are respectively $a_{1}$ and $a_{2}$ (recall Figure 3.1). Denote $p_{j}=P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{j}\right), j=1,2$, and we assume that $p_{2}$ is exponentially smaller than $p_{1}$ (i.e., $p_{2} / p_{1}$ decays exponentially). If we focus on $\mathcal{E}_{1}$ and simply use the exponential tilting towards $a_{1}$ as the IS distribution, then we face the risk of having a sample falling into $\mathcal{E}_{2}$ while the associated likelihood ratio is very high, which leads to asymptotic inefficiency. However, experimentally if we run the simulation with a moderate sample size, then most likely none of the samples fall into $\mathcal{E}_{2}$. Conditional on not hitting $\mathcal{E}_{2}$, we actually get an estimate close to $p_{1}$, which is in turn close to $p$. In other words, even if the resulting IS estimator is not asymptotically efficient, it could still give a good estimate in terms of its distance to $p$, as long as the sample size is not overly big. The latter is precisely the paradigm of probabilistic efficiency.

More concretely, we have the following theorem:

Theorem 3.3 (Achieving strong probabilistic efficiency). Suppose that $\left\{\mathcal{A}_{\gamma}\right\}_{\gamma}$ is an indexed family of rare events and $p=P\left(\mathcal{A}_{\gamma}\right) \rightarrow 0$ as $\gamma \rightarrow \infty$. We write $\mathcal{A}_{\gamma}=\mathcal{A}_{\gamma}^{1} \cup \mathcal{A}_{\gamma}^{2}$ where $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$ are two disjoint events. Denote $p_{j}=P\left(\mathcal{A}_{\gamma}^{j}\right), j=1,2$. Assume that

1. $\frac{p_{1}}{p} \rightarrow 1$ as $\gamma \rightarrow \infty$;
2. We have an asymptotically efficient IS estimator for $p_{1}$ obtained from $Z_{1}=I_{\mathcal{A}_{\gamma}^{1}} \frac{d P}{d \tilde{P}}$ under $\tilde{P}$. This implies that there exists $n=n(\gamma)$ growing subexponentially in $-\log p$ such that $\frac{\widehat{\operatorname{Var}}\left(Z_{1}\right)}{n p_{1}^{2}} \rightarrow 0$ as $\gamma \rightarrow \infty$;
3. $\tilde{p}_{2}:=\tilde{P}\left(\mathcal{A}_{\gamma}^{2}\right)$ satisfies that $n \tilde{p}_{2} \rightarrow 0$ as $\gamma \rightarrow \infty$.

Let $\hat{p}$ be the sample mean of $n$ independent replications of $Z=I_{\mathcal{A}_{\gamma}} \frac{d P}{d \tilde{P}}$ under $\tilde{P}$. For any $\varepsilon>0$, define $\delta_{\varepsilon}(\hat{p}, p)$ as in (3.7). Then we have

$$
\delta_{\varepsilon}(\hat{p}, p) \leq \sqrt{\frac{\widetilde{\operatorname{Var}}\left(Z_{1}\right)}{n p_{1}^{2}\left(\varepsilon-n \tilde{p}_{2}\right)}}+\frac{p_{2}}{p} \rightarrow 0 \text { as } \gamma \rightarrow \infty .
$$

Hence $Z$ is strongly probabilistically efficient.

Proof of Theorem 3.3. Suppose that we sample $\omega_{1}, \ldots, \omega_{n}$ under $\tilde{P}$. Let

$$
\hat{p}=\frac{1}{n} \sum_{i=1}^{n} I_{\mathcal{A}_{\gamma}}\left(\omega_{i}\right) \frac{d P}{d \tilde{P}}\left(\omega_{i}\right)
$$

and

$$
\hat{p}_{j}=\frac{1}{n} \sum_{i=1}^{n} I_{\mathcal{A}_{\gamma}^{j}}\left(\omega_{i}\right) \frac{d P}{d \tilde{P}}\left(\omega_{i}\right), j=1,2 .
$$

Clearly $\hat{p}=\hat{p}_{1}+\hat{p}_{2}$. For simplicity, denote

$$
\begin{align*}
& \delta:=\sqrt{\frac{\widetilde{\operatorname{Var}}\left(Z_{1}\right)}{n p_{1}^{2}\left(\varepsilon-n \tilde{p}_{2}\right)}}+\frac{p_{2}}{p},  \tag{3.8}\\
& \tilde{\delta}:=\frac{p_{2}}{p} . \tag{3.9}
\end{align*}
$$

Then we have that

$$
\begin{aligned}
\tilde{P}(|\hat{p}-p|>\delta p) & \leq \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>(\delta-\tilde{\delta}) p \text { or }\left|\hat{p}_{2}-p_{2}\right|>\tilde{\delta} p\right) \\
& \leq \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>(\delta-\tilde{\delta}) p\right)+\tilde{P}\left(\left|\hat{p}_{2}-p_{2}\right|>\tilde{\delta} p\right) \\
& \leq \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>(\delta-\tilde{\delta}) p_{1}\right)+\tilde{P}\left(\hat{p}_{2}>0\right) \\
& \leq \frac{\widetilde{\operatorname{Var}}\left(Z_{1}\right)}{n(\delta-\tilde{\delta})^{2} p_{1}^{2}}+n \tilde{p}_{2} \\
& =\varepsilon
\end{aligned}
$$

where the second inequality follows from a union bound, the third inequality follows from $p_{1} \leq p$ and that $\left|\hat{p}_{2}-p_{2}\right|>\tilde{\delta} p$ implies $\hat{p}_{2}>0$, the fourth inequality follows from Chebyshev's inequality in the first term and a union bound in the second term, and the last equality follows from the definitions in (3.8) and (3.9). Thus $\delta_{\varepsilon}(\hat{p}, p) \leq \delta$. Finally, $\delta \rightarrow 0$ by a direct use of the assumptions.

Similarly, if we relax the assumption that $p_{1} / p \rightarrow 1$ as $\gamma \rightarrow \infty$, we get sufficient conditions for weak probabilistic efficiency:

Theorem 3.4 (Achieving weak probabilistic efficiency). Suppose that $\left\{\mathcal{A}_{\gamma}\right\}_{\gamma}$ is an indexed family of rare events and $p=P\left(\mathcal{A}_{\gamma}\right) \rightarrow 0$ as $\gamma \rightarrow \infty$. We write $\mathcal{A}_{\gamma}=\mathcal{A}_{\gamma}^{1} \cup \mathcal{A}_{\gamma}^{2}$ where $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$ are two disjoint events. Denote $p_{j}=P\left(\mathcal{A}_{\gamma}^{j}\right), j=1,2$. Assume that

1. $\liminf _{\gamma \rightarrow \infty} \frac{p_{1}}{p}=c$ where $0<c \leq 1$;
2. We have an asymptotically efficient IS estimator for $p_{1}$ obtained from $Z_{1}=I_{\mathcal{A}_{\gamma}^{1}} \frac{d P}{d P}$ under $\tilde{P}$. This implies that there exists $n=n(\gamma)$ growing subexponentially in $-\log p$ such that $\frac{\widehat{\operatorname{Var}}\left(Z_{1}\right)}{n p_{1}^{2}} \rightarrow 0$ as $\gamma \rightarrow \infty$;
3. $\tilde{p}_{2}:=\tilde{P}\left(\mathcal{A}_{\gamma}^{2}\right)$ satisfies that $n \tilde{p}_{2} \rightarrow 0$ as $\gamma \rightarrow \infty$.

Let $\hat{p}$ be the sample mean of $n$ independent replications of $Z=I_{\mathcal{A}_{\gamma}} \frac{d P}{d \tilde{P}}$ under $\tilde{P}$. For any $\varepsilon>0$,
define $\delta_{\varepsilon}(\hat{p}, p)$ as in (3.7). Then we have

$$
\limsup _{\gamma \rightarrow \infty} \delta_{\varepsilon}(\hat{p}, p) \leq \limsup _{\gamma \rightarrow \infty}\left(\sqrt{\frac{\widetilde{\operatorname{Var}}\left(Z_{1}\right)}{n p_{1}^{2}\left(\varepsilon-n \tilde{p}_{2}\right)}}+\frac{p_{2}}{p}\right)=1-c<1 .
$$

Hence $Z$ is weakly probabilistically efficient.

Proof of Theorem 3.4. Following the proof of Theorem 3.3, we still get

$$
\delta_{\varepsilon}(\hat{p}, p) \leq \delta:=\sqrt{\frac{\widetilde{\operatorname{Var}}\left(Z_{1}\right)}{n p_{1}^{2}\left(\varepsilon-n \tilde{p}_{2}\right)}}+\frac{p_{2}}{p} .
$$

Under the conditions of Theorem 3.4, now $\lim \sup _{\gamma \rightarrow \infty} \delta=1-c<1$. By the definition, $Z$ is weakly probabilistically efficient.

We note that in Theorems 3.3 and 3.4, $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$ could be very general events. In particular, when we use dominating points to decompose the rare-event set, $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$ are not necessarily each governed by only one dominating point but could be more as long as the assumptions hold. Moreover, there can be multiple ways to split $\mathcal{A}_{\gamma}$ into $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$, and as long as one of these ways validates the assumptions in Theorem 3.3 or 3.4 then probabilistic efficiency is guaranteed. This provides flexibility in using Theorems 3.3 and 3.4; Sections 3.9.1 and 3.6.1 will demonstrate this in some specific examples.

According to the theorems, supposing that we have found some dominating points while the remaining ones are known to be less significant and "far from" the current ones, we could simply use the current mixture IS distribution instead of keep searching. The remaining question is how we could detect that the remaining dominating points are negligible, i.e. the assumptions of the theorem are satisfied. Besides, probabilistic efficiency only implies that the point estimate is reliable in some sense. This raises questions on inference such as the construction of valid CIs. In the next section, we will make these discussions precise and show our answers under minimal assumptions in the standard Gartner-Ellis regime.

### 3.5 Probabilistically Efficient Estimation in the Gartner-Ellis Regime

We study probabilistically efficient IS in the widely used Gartner-Ellis regime introduced in Section 3.2. Our key result is that probabilistic efficiency can be readily achieved by using only the most significant dominating points, under essentially no more assumptions than what is needed to derive the Gartner-Ellis large deviations asymptotic.

We first consider the case where there is only one most significant dominating point, which is a common scenario (e.g., in all the examples in Section 3.3):

Theorem 3.5 (Using the most significant dominating point is probabilistically efficient). Consider the problem of estimating $p=P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)$. Suppose that Assumptions 3.1 and 3.2 hold. Suppose also that the dominating set $A$ has finite cardinality with a unique most significant dominating point a, i.e., $I(a)<I(\tilde{a})$ for all other $\tilde{a} \in A$. Then the IS distribution $\tilde{P}$ given by the exponential tilting towards a, i.e.,

$$
\begin{equation*}
\frac{d \tilde{P}}{d P}=e^{s_{a}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a}\right)} \tag{3.10}
\end{equation*}
$$

is strongly probabilistically efficient.

Theorem 3.5 stipulates that we only need the most significant dominating point in constructing an efficient IS. This result, which is in sharp contrast to the established IS recipe that suggests using all dominating points, explains the good empirical performance of the "poor" estimators in Section 3.3. Moreover, the proposal in Theorem 3.5 is in closer line with the Gartner-Ellis asymptotic theory, in that the use of the most significant dominating point, which is also the minimizer of the rate function (recall Theorem 3.2), governs both the large deviations asymptotic and the construction of efficient IS. Lastly, regarding the assumptions needed, the only additional condition beyond the standard Gartner-Ellis assumptions (i.e., Assumptions 3.1 and 3.2) is the finite cardinality of the dominating set. In fact, if we have multiple most significant points, we have a natural generalization:

Theorem 3.6 (Mixing most significant dominating points). Consider the problem of estimating
$p=P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)$. Suppose that Assumptions 3.1 and 3.2 hold. Suppose also that the dominating set $A$ has finite cardinality with $k$ most significant dominating points $a_{1}, \ldots, a_{k}$, i.e., $I\left(a_{1}\right)=\cdots=$ $I\left(a_{k}\right)<I\left(a^{\prime}\right)$ for all $a^{\prime} \in A \backslash\left\{a_{1}, \ldots, a_{k}\right\}$. Then the IS distribution $\tilde{P}$ given by the mixture of exponential tiltings towards $a_{1}, \ldots, a_{k}$, i.e.,

$$
\begin{equation*}
\frac{d \tilde{P}}{d P}=\sum_{i=1}^{k} \alpha_{i} e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)} \tag{3.11}
\end{equation*}
$$

is strongly probabilistically efficient for any fixed $\alpha_{i}$ 's such that $\sum_{i=1}^{k} \alpha_{i}=1, \alpha_{i}>0$, $\forall i$.

That is, we use mixture to account for all the most significant dominating points when there are multiple of them. The proofs of Theorems 3.5 and 3.6 amount to verifying the assumptions in Theorem 3.3 using the Gartner-Ellis conditions. In particular, Conditions 1 and 2 in Theorem 3.3 can be routinely verified, while Condition 3 is checked by showing that $\tilde{p}_{2}$ is in fact exponentially decaying in $\gamma$, which requires an application of the Gartner-Ellis theorem under the IS distribution. The verification of Condition 3 especially reveals a key phenomenon that, under the exponential tilting to the most significant dominating point(s), the probability of an IS sample hitting onto the "backyards" of other dominating points is exponentially small, which in turn fulfills the notion of probabilistic efficiency.

Proofs of Theorems 3.5 and 3.6. We focus on Theorem 3.6 since Theorem 3.5 is a special case therein. It suffices to verify all the assumptions in Theorem 3.3. For $\mathcal{A}_{\gamma}=\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right\}$, by Theorem 3.1, $p=P\left(\mathcal{A}_{\gamma}\right)$ satisfies that $\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log p=-I(\mathcal{E})<0$, so $p \rightarrow 0$ as $\gamma \rightarrow \infty$. If $k=|A|$, then the IS estimator from (3.11) already uses all the dominating points and thus is asymptotically efficient by Proposition 3.1. Hence, from now on, we assume that $k<|A|$. For convenience, we denote $a_{k+1}$ as a next most significant point other than $a_{1}, \ldots, a_{k}$, i.e., $I\left(a_{k+1}\right)>$ $I\left(a_{1}\right)=\cdots=I\left(a_{k}\right)$ and $I\left(a_{k+1}\right) \leq I\left(a^{\prime}\right)$ for all $a^{\prime} \in A \backslash\left\{a_{1}, \ldots, a_{k}\right\}$. We split $\mathcal{E}$ into $\mathcal{E}_{1}=$ $\mathcal{E} \cap \bigcup_{i=1}^{k}\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$ and $\mathcal{E}_{2}=\mathcal{E} \backslash \mathcal{E}_{1}$, and define $\mathcal{A}_{j}=\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{j}\right\}, p_{j}=P\left(\mathcal{A}_{j}\right)$ for $j=1,2$.

First, by Theorem 3.1, we have that $\lim \sup _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log p_{2} \leq-I\left(\overline{\mathcal{E}_{2}}\right)$. Since $\overline{\mathcal{E}_{2}} \subset \bigcup_{j=k+1}^{|A|}\{x:$
$\left.s_{a_{j}}^{\top}\left(x-a_{j}\right) \geq 0\right\}$, we know that $I\left(\overline{\mathcal{E}_{2}}\right) \geq I\left(\bigcup_{j=k+1}^{|A|}\left\{x: s_{a_{j}}^{\top}\left(x-a_{j}\right) \geq 0\right\}\right)=I\left(a_{k+1}\right)$, and hence $\lim \sup _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log p_{2} \leq-I\left(a_{k+1}\right)<-I(\mathcal{E})$. As a result, $p_{2} / p \rightarrow 0$ as $\gamma \rightarrow \infty$. This verifies Assumption 1 in Theorem 3.3.

Second, by the definition, $\left\{a_{1}, \ldots, a_{k}\right\}$ is a dominating set for $\mathcal{E}_{1}$, so the IS estimator $Z_{1}=$ $I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}(\omega)$ is asymptotically efficient by Proposition 3.1. This verifies Assumption 2 in Theorem 3.3.

Third, we would prove that $\tilde{p}_{2}$ decays exponentially in $\gamma$ (hence also exponentially in $-\log p$ ), and hence $n \tilde{p}_{2} \rightarrow 0$ for subexponentially growing $n$ which verifies Assumption 3 in Theorem 3.3. Indeed, we have

$$
\tilde{p}_{2}=\sum_{i=1}^{k} \alpha_{i} \tilde{P}_{i}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{2}\right)
$$

where $\frac{d \tilde{P}_{i}}{d P}(\omega)=e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)}$. Denote $\tilde{E}_{i}, \tilde{\mu}_{\gamma, i}$ and $\tilde{I}_{i}$ as the corresponding expectation, scaled logarithmic moment generating function and rate function under $\tilde{P}_{i}$. Then, under $\tilde{P}_{i}$, we have

$$
\tilde{\mu}_{\gamma, i}(x)=\frac{1}{\gamma} \log \tilde{E}_{i}\left(e^{x^{\top} X_{\gamma}}\right)=\frac{1}{\gamma} \log E\left(e^{x^{\top} X_{\gamma}+s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)}\right)=\mu_{\gamma}\left(x+s_{a_{i}}\right)-\mu_{\gamma}\left(s_{a_{i}}\right)
$$

and thus $\tilde{\mu}_{i}(x)=\lim _{\gamma \rightarrow \infty} \tilde{\mu}_{\gamma, i}(x)=\mu\left(x+s_{a_{i}}\right)-\mu\left(s_{a_{i}}\right)$. Then the rate function is

$$
\begin{aligned}
\tilde{I}_{i}(y) & =\sup _{x \in \mathbb{R}^{d}}\left\{x^{\top} y-\tilde{\mu}_{i}(x)\right\} \\
& =\sup _{x \in \mathbb{R}^{d}}\left\{x^{\top} y-\mu\left(x+s_{a_{i}}\right)+\mu\left(s_{a_{i}}\right)\right\} \\
& =\sup _{x \in \mathbb{R}^{d}}\left\{\left(x+s_{a_{i}}\right)^{\top} y-\mu\left(x+s_{a_{i}}\right)\right\}-s_{a_{i}}^{\top} y+\mu\left(s_{a_{i}}\right) \\
& =I(y)-s_{a_{i}}^{\top} y+\mu\left(s_{a_{i}}\right) .
\end{aligned}
$$

For any $y \in \overline{\mathcal{E}_{2}}$, we have that $I(y) \geq I\left(a_{k+1}\right)$ and that $s_{a_{i}}^{\top}\left(y-a_{i}\right) \leq 0$ for $i=1, \ldots, k$, and thus $\tilde{I}_{i}(y)=I(y)-s_{a_{i}}^{\top} y+\mu\left(s_{a_{i}}\right) \geq I\left(a_{k+1}\right)-s_{a_{i}}^{\top} a_{i}+\mu\left(s_{a_{i}}\right)=I\left(a_{k+1}\right)-I\left(a_{i}\right)$. Therefore, $\tilde{I}_{i}\left(\overline{\mathcal{E}_{2}}\right) \geq I\left(a_{k+1}\right)-I\left(a_{i}\right)=I\left(a_{k+1}\right)-I\left(a_{1}\right)>0$. From the above derivations, Assumption 3.1 still
holds for $\tilde{P}_{i}$. By Theorem 3.1,

$$
\limsup _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \tilde{P}_{i}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{2}\right) \leq-\tilde{I}_{i}\left(\overline{\mathcal{E}_{2}}\right) \leq-\left(I\left(a_{k+1}\right)-I\left(a_{1}\right)\right)<0
$$

Overall, we have $\tilde{p}_{2}$ decays exponentially in $\gamma$.
Now we have verified all the assumptions in Theorem 3.3.

We comment that if we know any one of the most significant points, say $a$, among several such points, satisfies $p_{1} / p \rightarrow c$ for some $0<c \leq 1$, where $p_{1}=P\left(\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}\left(\frac{1}{\gamma} X_{\gamma}-a\right) \geq 0\right\}\right)$ is the rare-event probability "contributed" from $a$, then using the IS that exponentially tilts only to $a$, i.e., (3.10), is weakly probabilistically efficient. This can be shown by a similar argument to the proofs of Theorems 3.5 and 3.6 above. Such an approach is in contrast to using the IS mixture in (3.11) suggested by Theorem 3.6 that achieves strong, instead of only weak, probabilistic efficiency. Nonetheless, knowing $p_{1} / p \rightarrow c$ typically requires information on the multiplicative factor in front of the exponential decay dictated by the large deviations rate function, which in turn requires derivation of exact asymptotic that is only known for a relatively small number of problems.

Next, besides point estimates, we investigate inference using probabilistically efficient IS estimators, in particular how to construct (asymptotically) valid CIs. First, we consider the interval

$$
\begin{equation*}
I_{1}=\left[\hat{p}-\left(\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}\right), \hat{p}+\left(\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}\right)\right] \tag{3.12}
\end{equation*}
$$

where $\hat{V}$ is the sample variance and $M_{\gamma}=\max _{i=1, \ldots, k}\left\{\frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right\}$ is deterministic. The following theorem provides an asymptotic coverage guarantee for this CI.

Theorem 3.7 (Constructing confidence intervals with probabilistically efficient estimators). Under the same setting as Theorem 3.6, suppose we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)}=$ $I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}, i=1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$ 's. If $n$ is subexponentially growing in $-\log p$ (or $\gamma$ ) as $\gamma \rightarrow \infty$, then, for any
$0<\alpha<1$,

$$
\liminf _{\gamma \rightarrow \infty} \tilde{P}\left(p \in \mathcal{I}_{1}\right) \geq 1-\alpha
$$

where $M_{\gamma}=\max _{i=1, \ldots, k}\left\{\frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right\}$.
We provide some intuition in constructing interval (3.12). Roughly speaking, our assumptions in Theorem 3.7 guarantee that under our IS, the rare-event region controlled by the less significant dominating points, namely $\mathcal{E}_{2}$ in Theorem 3.3, is not hit with high probability. In this case, $\hat{p}$ and $\hat{V}$ can also be seen as the sample mean and sample variance of the IS estimator for $p_{1}$ instead of $p$, and thus properly utilizing them could result in a CI that covers $p_{1}$ with high confidence. As $p_{1}$ is close to $p$, such a CI also covers $p$ with high confidence. In other words, constructing a CI for $p$ reduces to constructing a CI for $p_{1}$, where the latter is more tangible to attain as the involved variance does not blow up. The execution of this intuition requires a concentration inequality for the sample mean that is an empirical adaption of Bennett's inequality [133]. This concentration inequality leverages the boundedness property of the random variables (where $M_{\gamma}$ is the upper bound) instead of the normal approximation property, which leads to the additional term of order $1 / n$ in (3.12) in addition to the $1 / \sqrt{n}$ term that appears in a typical normality interval.

In Theorem 3.7, note that even if we neglect the higher-order term (in terms of $n) \frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}$, the CI half-width is $\sqrt{2 \log (4 / \alpha)}$ times $\sqrt{\frac{\hat{V}}{n}}$, which is more conservative than the Central Limit Theorem (CLT) based interval

$$
\begin{equation*}
I_{2}=\left[\hat{p}-z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}, \hat{p}+z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right] \tag{3.13}
\end{equation*}
$$

where $z_{1-\alpha / 2}$ is the $(1-\alpha / 2)$-quantile of the standard normal distribution. For instance, when $\alpha=0.05$, we have $\sqrt{2 \log (4 / \alpha)} \approx 2.96$, while $z_{1-\alpha / 2} \approx 1.96$. Our next theorem shows that, under stronger conditions, the CLT-based CI (3.13) is also asymptotically valid.

Theorem 3.8 (Constructing tight confidence intervals with probabilistically efficient estimators). Under the same setting as Theorem 3.6, suppose we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from P . Let $Z^{(i)}=$
$I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}$ and $Z_{1}^{(i)}=I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}, i=1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$ 's. In this case, we could choose $n$ subexponentially growing in $-\log p($ or $\gamma)$ such that $\frac{M_{\gamma}^{2}}{n \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)} \rightarrow 0$ and $\frac{\tilde{E}^{2}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{n \widehat{\operatorname{Var}}{ }^{3}\left(Z_{1}^{(1)}\right)} \rightarrow 0$ as $\gamma \rightarrow \infty$ where $M_{\gamma}$ is as defined in Theorem 3.7. Then, for any $0<\alpha<1$,

$$
\liminf _{\gamma \rightarrow \infty} \tilde{P}\left(p \in \mathcal{I}_{2}\right) \geq 1-\alpha
$$

Theorem 3.8 tightens the interval in Theorem 3.7 to using the CLT-based critical value $z_{1-\alpha / 2}$ with a more careful choice of sample size $n$.

Finally, we prove that if we use all the dominating points in the mixture, so that the estimator satisfies the classical notion of asymptotic efficiency, then, under conditions similar to Theorem 3.8, the CLT-based interval possesses an even stronger guarantee that the asymptotic coverage probability is exactly $1-\alpha$.

Theorem 3.9 (Asymptotically exact confidence intervals with asymptotically efficient estimators). Consider the problem of estimating $p=P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)$. Suppose that Assumptions 3.1 and 3.2 hold, and the dominating set is finite. The IS estimator is $Z=I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}$ under $\tilde{P}$ given by (3.3). We sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)}=I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}, i=1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$ 's. In this case, we could choose $n$ at least subexponentially growing in $-\log p($ or $\gamma)$ such that $\frac{M_{\gamma}^{2}}{n \widetilde{\operatorname{Var}}\left(Z^{(1)}\right)} \rightarrow 0$ and $\frac{\tilde{E}^{2}\left|Z^{(1)}-p\right|^{3}}{n \widetilde{V a r}{ }^{3}\left(Z^{(1)}\right)} \rightarrow 0$ as $\gamma \rightarrow \infty$ where $M_{\gamma}=\max _{i=1, \ldots, r}\left\{\frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right\}$. Then, for any $0<\alpha<1$,

$$
\lim _{\gamma \rightarrow \infty} \tilde{P}\left(p \in \mathcal{I}_{2}\right)=1-\alpha
$$

We make several remarks regarding the properties of the CLT-based CI $I_{2}$ in Theorems 3.8 and 3.9. First, it appears that probabilistically efficient samples sacrifice some looseness in terms of

CI coverage compared to asymptotically efficient samples, as the guarantee is valid in Theorem 3.8 but exact in Theorem 3.9. Second, in Theorem 3.8, like Theorem 3.7, the sample size $n$ is required to be not overly big, manifested by the subexponential growth requirement. This is in contrast to Theorem 3.9 that does not impose any upper bound on $n$, where it suffices that $n$ is at least subexponentially growing, i.e., $n$ could grow exponentially or even faster. This ties to the key idea of probabilistic efficiency that, when the sample size is not overly big, there is a negligible chance of any sample hitting the rare-event region not corresponding to the most significant points. Thus the CI constructed from a probabilistically efficient estimator, much like the point estimate, is valid only when the sample size is not overly big, while asymptotically efficient estimators do not impose such a restriction. Lastly, we see the requirements on $n$ given by $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)} \rightarrow 0$ and $\frac{\tilde{E}^{2}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{n \widetilde{\operatorname{Var}}{ }^{3}\left(Z_{1}^{(1)}\right)} \rightarrow 0$ in Theorem 3.8 and similar ones in Theorem 3.9. While these conditions can be difficult to verify in practice, we should note that they are lower bound requirements, and imposed not only for CIs constructed from probabilistically efficient estimators, but also for classical asymptotically efficient estimators as well (to our best knowledge, conditions on the adequacy of sample size to attain CI coverage guarantees for these classical estimators is not known in the literature). In the next section, we will investigate the performances of all these CIs with reasonable sample sizes.

Lastly, to close this section, we briefly note that Algorithm 3.1 in Section 3.8 shows generally how to identify and compute dominating points, sequentially starting from the most significant one. Moreover, Section 3.10 studies parallel results to this section for an alternative asymptotic regime to Gartner-Ellis that could be suitable for some situations involving highly complex systems.

### 3.6 Further Numerical Experiments and Discussions

We have shown several examples in Section 3.3 where IS estimators using only one or a small number of dominating points perform competitively compared with asymptotically efficient IS estimators that use all dominating points. In fact, we have shown in each example in Sections 3.3.1, 3.3.2 and 3.3.3 that the simple estimator using the most significant dominating point is not
asymptotically efficient. In this section, we argue that they are all probabilistically efficient, which is a direct consequence of Theorem 3.5. We then numerically assess the validity of the conditions in Theorem 3.3, which forms the underlying basis in justifying probabilistic efficiency. Finally, we test the confidence intervals constructed using our probabilistically efficient estimators discussed in Section 3.5 and compare with intervals constructed from asymptotically efficient estimators.

### 3.6.1 Verifying Conditions for Probabilistic Efficiency

We first state the strong probabilistic efficiency of all the proposed estimators that use only the most significant dominating points in Section 3.3:

Theorem 3.10. Under the problem specifications in Sections 3.3.1, 3.3.2 and 3.3.3, the IS estimators that use only the most significant dominating points, namely $\hat{\alpha}(m)$ in Section 3.3.1, $X_{\gamma}$ distributed as $N\left(\gamma a_{1}, \gamma \Sigma\right)$ in Section 3.3.2 and $N\left(\gamma a_{1}, \gamma I\right)$ in Section 3.3.3, are all strongly probabilistically efficient.

Next, we validate the underpinning mechanism of how probabilistic efficiency arises in these examples. Note that the main basis of the strong probabilistic efficiency of these estimators, which follows from Theorem 3.5, is Theorem 3.3. In particular, Theorem 3.3 states three conditions that allow one to conclude strong probabilistic efficiency. Among them, the second condition is a property about asymptotic efficiency for an estimator that applies to a more restrictive rare-event set, which has been well-established in the asymptotic efficiency literature (basically, by mixing the exponential tiltings towards all the dominating points associated with the more restrictive rare-event set). Conditions 1 and 3 are more delicate. In the setting with a unique most significant dominating point, say $a$, the former requires a small proportion of the "contribution" from the less significant dominating points other than $a$ over the total rare-event probability, i.e., $p_{2} / p \rightarrow 0$ where $p_{2}=P\left(\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$. The latter requires a small probability of sampling any points in the rare-event set that does not belong to the backyard of $a$, i.e., $\tilde{P}$ (some of the $n$ samples hits $\left.\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right) \rightarrow 0$ or, as a sufficient condition, $n \tilde{p}_{2} \rightarrow 0$ where $\tilde{p}_{2}=\tilde{P}\left(\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$. Our next goal is to assess the smallness and decreasing
trends (as rarity grows) of $p_{2} / p$ and $\tilde{P}$ (some of the $n$ samples hits $\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ ) that drive Theorem 3.3. In the following, we present the aforementioned investigation for the i.i.d. sum large deviations example in Section 3.3.1 and the MNIST example in Section 3.3.3, and leave the random walk example in Section 3.3.2 to Section 3.9.1. We also illustrate a variation of this latter example that uses the weak probabilistic efficiency in Section 3.9.2.

## Large Deviations of an I.I.D. Sum.

For the experiment in Section 3.3.1, we use the probabilistically efficient estimator $\hat{\alpha}(m)$. Correspondingly, we have $p_{1}=P\left(S_{m} \geq a m\right)$ and $p_{2}=P\left(S_{m} \leq-a m\right)$. Table 3.3 shows these values as $m$ varies, which we approximate respectively by using estimators $\hat{\beta}_{1}(m)=\exp \left(-s_{a} S_{m}+\right.$ $\left.m \mu\left(s_{a}\right)\right) I_{\left\{S_{m} \geq a m\right\}}$ and $\hat{\beta}_{2}(m)=\exp \left(-s_{-a} S_{m}^{\prime}+m \mu\left(s_{-a}\right)\right) I_{\left\{S_{m}^{\prime} \leq-a m\right\}}$, with $s_{a}, s_{-a}$ defined in Section 3.3.1, generated by the same IS samples used in $\hat{\beta}(m)$. From Table 3.3, we observe that the estimate of $p_{2} / p$ is 0.008 with $m=10$ and decreases to $5.00 \times 10^{-19}$ as $m=100$. This shows that $p_{2} / p$ is small and approaches 0 as $m$ increases, which matches Condition 1 in Theorem 3.3 ( $m$ is the rarity parameter here).

Next, we examine $\tilde{p}_{2}=\tilde{P}\left(S_{m} \leq-a m\right)$. We generate $10^{7}$ samples from the strongly probabilistically efficient IS distribution. We observe that none of the samples fall into $\left\{S_{m} \leq-a m\right\}$, which indicates that $\tilde{p}_{2}$ is extremely small so that $\tilde{P}$ (some of the $n$ samples hits $\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ ), with $n=10^{4}$ in our experiment here, is close to zero. This matches Condition 3 of Theorem 3.3.

Table 3.3: Estimates of $p_{1}$ and $p_{2}$ for the i.i.d sum example in Section 3.3.1 with $10^{4}$ samples.

| $m$ | 10 | 30 | 50 | 100 |
| :--- | :--- | :--- | :--- | :--- |
| $p_{1}$ | $8.33 \times 10^{-3}$ | $1.59 \times 10^{-5}$ | $3.59 \times 10^{-8}$ | $1.33 \times 10^{-14}$ |
| $p_{1} / p$ | 0.992 | $\approx 1$ | $\approx 1$ | $\approx 1$ |
| $p_{2}$ | $6.56 \times 10^{-5}$ | $3.15 \times 10^{-11}$ | $1.85 \times 10^{-17}$ | $6.68 \times 10^{-33}$ |
| $p_{2} / p$ | 0.008 | $1.98 \times 10^{-6}$ | $5.16 \times 10^{-10}$ | $5.00 \times 10^{-19}$ |
| $p$ | $8.40 \times 10^{-3}$ | $1.59 \times 10^{-5}$ | $3.59 \times 10^{-8}$ | $1.33 \times 10^{-14}$ |

## MNIST Example.

The experiment in Section 3.3.3 uses a probabilistically efficient estimator based on the exponential tilting towards the most significant dominating point only, i.e., $X_{\gamma}$ distributed as $N(\gamma a, \gamma I)$ with $a=a_{1}$. We define $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$, and the corresponding probabilities $p_{1}=P\left(\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$ and $p_{2}=P\left(\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$ which are shown in Table 3.4. Note that in this MNIST example, the total number of dominating points is large and unknown. Thus we only present the contribution of the first 10 dominating points, i.e. $a_{1}, a_{2}, \ldots, a_{10}$ in Table 3.4, denoted by $p_{a_{2}}=P\left(\mathcal{A}_{\gamma} \cap\left\{s_{a_{2}}^{\top}\left(x-a_{2}\right) \geq 0\right\} \backslash\left\{s_{a_{1}}^{\top}\left(x-a_{1}\right) \geq 0\right\}\right), \ldots$, $p_{a_{10}}=P\left(\mathcal{A}_{\gamma} \cap\left\{s_{a_{10}}^{\top}\left(x-a_{10}\right) \geq 0\right\} \backslash\left(\cup_{j=1}^{9}\left\{s_{a_{j}}^{\top}\left(x-a_{j}\right) \geq 0\right\}\right)\right)$. We estimate each of the probabilities $p_{1}, p_{a_{2}}, \ldots, p_{a_{10}}$ using the IS estimator with the corresponding dominating point, i.e. the IS distribution is exponentially tilted using dominating points $a_{1}, \ldots, a_{10}$ respectively. We borrow the values of $p$ from Table 3.9 where each estimate is computed using crude Monte Carlo, and we estimate $p_{2}$ through $p_{2}=p-p_{1}$. We observe that the ratio $p_{2} / p$ decreases from 0.2154 to 0.1931 as we decrease the value of $\sigma$ from 0.2 to 0.17 , i.e., the problem becomes rarer. We also observe that some individual relative contribution slightly increases in this experiment. However, these increases do not affect the decreasing trend of the total relative contribution of the less significant dominating points. For example, $p_{a_{2}} / p$ and $p_{a_{3}} / p$ both increase slightly as $\sigma$ decreases (from 0.1074 and 0.0035 with $\sigma=0.2$ to 0.1313 and 0.0042 with $\sigma=0.17$ respectively), but the relative contribution of the rest of the less significant dominating points (excluding the first 10) is $0.0812,0.0658,0.0633$, and 0.0362 for $\sigma=0.2,0.19,0.18,0.17$ respectively, which vanishes fast as $\sigma$ decreases.

Table 3.5 presents the estimates of probabilities $\tilde{p}_{1}=\tilde{P}\left(\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$ and $\tilde{p}_{2}=$ $\tilde{P}\left(\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$ under the probabilistically efficient IS distribution. The probabilities $\tilde{p}_{a_{2}}, \ldots, \tilde{p}_{a_{10}}$, defined by $\tilde{p}_{a_{i}}=\tilde{P}\left(\mathcal{A}_{\gamma} \cap\left\{s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\} \backslash\left(\cup_{j=1}^{i-1}\left\{s_{a_{j}}^{\top}\left(x-a_{j}\right) \geq 0\right\}\right)\right)$, for $i=$ $2, \ldots, 10$, are also shown to illustrate the contributions of the dominating points $a_{2}, \ldots, a_{10}$ for $\tilde{p}_{2}$. We find that $\tilde{p}_{2}$ decreases from 0.0090 to 0.0085 as $\sigma$ decreases from 0.2 to 0.17 , i.e., the problem becomes rarer. From the individual contribution, we observe that all the probabilities

Table 3.4: Estimates of $p_{1}, p_{2}$ and the contributions of the less significant dominating points for the MNIST example in Section 3.3.3 with $10^{4}$ samples, where we use $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$.

| $\sigma$ | 0.2 | 0.19 | 0.18 | 0.17 |
| :--- | :--- | :--- | :--- | :--- |
| $p_{1}$ | $9.18 \times 10^{-6}$ | $3.34 \times 10^{-6}$ | $1.02 \times 10^{-6}$ | $2.54 \times 10^{-7}$ |
| $p_{1} / p$ | 0.7846 | 0.7920 | 0.7892 | 0.8069 |
| $p_{2}$ | $2.52 \times 10^{-6}$ | $8.77 \times 10^{-7}$ | $2.72 \times 10^{-7}$ | $6.09 \times 10^{-8}$ |
| $p_{2} / p$ | 0.2154 | 0.2080 | 0.2108 | 0.1931 |
| $p$ | $1.17 \times 10^{-5}$ | $4.22 \times 10^{-6}$ | $1.29 \times 10^{-6}$ | $3.15 \times 10^{-7}$ |
| $p_{a_{2}}$ | $1.26 \times 10^{-6}$ | $4.95 \times 10^{-7}$ | $1.61 \times 10^{-7}$ | $4.14 \times 10^{-8}$ |
| $p_{a_{2}} / p$ | 0.1074 | 0.1173 | 0.1245 | 0.1313 |
| $p_{a_{3}}$ | $4.12 \times 10^{-8}$ | $1.22 \times 10^{-8}$ | $2.93 \times 10^{-9}$ | $1.32 \times 10^{-9}$ |
| $p_{a_{3}} / p$ | 0.0035 | 0.0029 | 0.0023 | 0.0042 |
| $p_{a_{4}}$ | $1.24 \times 10^{-7}$ | $4.45 \times 10^{-8}$ | $1.26 \times 10^{-8}$ | $3.22 \times 10^{-9}$ |
| $p_{a_{4}} / p$ | 0.0106 | 0.0105 | 0.0098 | 0.0102 |
| $p_{a_{5}}$ | $2.48 \times 10^{-8}$ | $9.09 \times 10^{-9}$ | $2.81 \times 10^{-9}$ | $7.01 \times 10^{-10}$ |
| $p_{a_{5}} / p$ | 0.0021 | 0.0022 | 0.0022 | 0.0022 |
| $p_{a_{6}}$ | $3.60 \times 10^{-8}$ | $1.31 \times 10^{-8}$ | $3.28 \times 10^{-9}$ | $8.11 \times 10^{-10}$ |
| $p_{a_{6}} / p$ | 0.0031 | 0.0031 | 0.0025 | 0.0026 |
| $p_{a_{7}}$ | 0 | 0 | 0 | 0 |
| $p_{a_{8}}$ | $8.80 \times 10^{-8}$ | $2.62 \times 10^{-8}$ | $8.03 \times 10^{-9}$ | $2.02 \times 10^{-9}$ |
| $p_{a_{8}} / p$ | 0.0075 | 0.0062 | 0.0062 | 0.0064 |
| $p_{a_{9}}$ | 0 | 0 | 0 | 0 |
| $p_{a_{10}}$ | 0 | 0 | 0 | 0 |

$\tilde{p}_{a_{3}}, \ldots, \tilde{p}_{a_{10}}$ decrease rapidly, except $\tilde{p}_{a_{2}}$ that slightly increases as $\sigma$ decreases. We use the value of $\tilde{p}_{2}$ to estimate probability $\tilde{p}_{h i t}=\tilde{P}$ (some of the $n$ samples hits $\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ ) through $1-\left(1-\tilde{p}_{2}\right)^{n}$. The last row in Table 3.5 presents the results with $n=10^{4}$ (the sample size we use in Section 3.3). We observe that there are samples falling into $\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ with approximately probability 1 and hence this result cannot explain the good performance of our probabilistically efficient estimator.

To this end, we verify the conditions in Theorem 3.3 using an alternative construction of $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash \mathcal{A}_{\gamma}^{1}$. Here, in our previous construction, we have chosen the $\mathcal{A}_{\gamma}^{1}$ to be the halfspace cut by a dominating point and it turns out that the corresponding $\tilde{p}_{2}$ is not small and thus the condition of Theorem 3.3 appears to fail. However, as discussed right after Theorem 3.4, our main theorems allow more flexibility in choosing our $\mathcal{A}_{\gamma}^{1}$, and as long as we find a suitable way to

Table 3.5: Estimates of $\tilde{p}_{1}, \tilde{p}_{2}$ and the contributions of the less significant dominating points for the MNIST example under the probabilistically efficient IS in Section 3.3.3 with $10^{4}$ samples.

| $\sigma$ | 0.2 | 0.19 | 0.18 | 0.17 |
| :--- | :--- | :--- | :--- | :--- |
| $\tilde{p}_{1}$ | 0.4728 | 0.4745 | 0.4760 | 0.4775 |
| $\tilde{p}_{2}$ | 0.0090 | 0.0087 | 0.0086 | 0.0085 |
| $\tilde{p}_{a_{2}}$ | 0.0069 | 0.007 | 0.0072 | 0.0072 |
| $\tilde{p}_{a_{3}}$ | 0.0003 | 0.0003 | 0.0003 | 0.0002 |
| $\tilde{p}_{a_{4}}$ | 0.0006 | 0.0006 | 0.0006 | 0.0006 |
| $\tilde{p}_{a_{5}}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{6}}$ | 0.0003 | 0.0003 | 0.0001 | 0.0001 |
| $\tilde{p}_{a_{7}}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{8}}$ | 0.0006 | 0.0003 | 0.0003 | 0.0003 |
| $\tilde{p}_{a_{9}}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{10}}$ | 0 | 0 | 0 | 0 |
| $\tilde{p}_{h i t}$ | $\approx 1$ | $\approx 1$ | $\approx 1$ | $\approx 1$ |

construct $\mathcal{A}_{\gamma}^{1}$ to satisfy the needed conditions, Theorem 3.3 can be used to explain our estimator's good performance.

Here is how we can construct a suitable alternative $\mathcal{A}_{\gamma}^{1}$ for Theorem 3.3. From the proofs of Propositions 3.4 and 3.5, we know that our probabilistically efficient estimator is not asymptotically efficient if and only if $\min _{x \in \mathcal{A}_{\gamma}}\left(x+a_{1}\right)^{\top}\left(x+a_{1}\right)<4 a_{1}^{\top} a_{1}$. We split the rare-event set $\mathcal{A}_{\gamma}$ into two parts, namely $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{\left(x+a_{1}\right)^{\top}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} a_{1}\right\}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash$ $\left\{\left(x+a_{1}\right)^{\top}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} a_{1}\right\}$, and our probabilistically efficient estimator is asymptotically efficient for estimating the probabilities of $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\} \subseteq \mathcal{A}_{\gamma}^{1}$. We define $p_{1}=P\left(\mathcal{A}_{\gamma}^{1}\right), p_{2}=P\left(\mathcal{A}_{\gamma}^{2}\right), \tilde{p}_{1}=\tilde{P}\left(\mathcal{A}_{\gamma}^{1}\right)$, and $\tilde{p}_{2}=\tilde{P}\left(\mathcal{A}_{\gamma}^{2}\right)$ for our newly constructed $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$. We first show $p_{2} / p \rightarrow 0$ and $n \tilde{p}_{2} \rightarrow 0$. We note that $\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\} \subseteq \mathcal{A}_{\gamma}^{1}$ because $\left\{s_{a}^{\top}(x-a) \geq 0\right\} \subseteq\left\{\left(x+a_{1}\right)^{\top}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} a_{1}\right\}$. Hence we also have $\mathcal{A}_{\gamma}^{2} \subseteq A_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$, which leads to $p_{2} \leq p_{2}^{\text {old }}$ and $\tilde{p}_{2} \leq \tilde{p}_{2}^{\text {old }}$ where $p_{2}^{\text {old }}$ and $\tilde{p}_{2}^{\text {old }}$ refer to the $p_{2}$ and $\tilde{p}_{2}$ evaluated using our old construction of $\mathcal{A}_{\gamma}^{2, \text { old }}=\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$. By $p_{2}^{\text {old }} / p \rightarrow 0$ and $n \tilde{p}_{2}^{\text {old }} \rightarrow 0$ as $\gamma \rightarrow \infty$ from Theorem 3.5 we have $p_{2} / p \rightarrow 0$ and $n \tilde{p}_{2} \rightarrow 0$. That is, our current new construction $\mathcal{A}_{\gamma}^{1}$ for Theorem 3.3 would satisfy the conditions therein, and we would like to numerically verify especially Conditions 1 and 3.

We now check the numerical values of $p_{2}$ and $\tilde{p}_{2}$ to verify these conditions. We use the mixture
of all 100 dominating points as the IS distribution for estimating $p_{2}$. We generate $10^{6}$ samples for $\sigma$ varying from 0.17 to 0.2 and find no samples falling into $\mathcal{A}_{\gamma}^{2}$, which indicates that $p_{2}$ (and hence $\left.p_{2} / p\right)$ is extremely small in all cases. We generate $10^{6}$ samples from the probabilistically efficient IS distribution to estimate $\tilde{p}_{2}$ and observe no samples hitting $\mathcal{A}_{\gamma}^{2}$ for the same range of $\sigma$. In this case, $\tilde{p}_{h i t}=\tilde{P}\left(\right.$ some of the $n$ samples hits $\left.\mathcal{A}_{\gamma}^{2}\right)$ with $n=10^{4}$ would be close to zero due to the extremely small values of $\tilde{p}_{2}$. These results match Conditions 1 and 3 of Theorem 3.3 and hence explain the good performance of our probabilistically efficient estimator in the experiment.

### 3.6.2 Illustration of Confidence Intervals

We investigate the performances of the CIs proposed in Section 3.5. In particular, we construct CIs (3.12) and (3.13) from probabilistically efficient estimators, namely the IS schemes using only the most significant point in Sections 3.3.1 and 3.3.3 (see Section 3.9.1 for the example in Section 3.3.2). For convenience, we call interval (3.12) the "loose CI" and interval (3.13) the "tight CI", since the latter has a shorter length and matches the CLT-based interval. For comparison, we also construct CI (3.13) from asymptotically efficient estimators. In particular, for the settings in Sections 3.3.1 and 3.3.2, these estimators are built from mixtures of exponential tiltings towards all the dominating points. For the setting in Section 3.3.3, computing all dominating points requires insurmountable resources (as discussed therein), and so we use the mixture of 100 dominating points as a proxy of an asymptotically efficient estimator (100 is the total number of dominating points we discover using one-week's computation).

In the experiments, we compare the coverage rates of all three intervals described above. These coverage rates are obtained from a large number of experimental repetitions. Since the ground truths of these problems are unknown, we run a gigantic amount of simulation runs using either asymptotically efficient estimators or crude Monte Carlo to obtain highly accurate estimates, which serve as the "truths" when estimating the coverage of the CIs. The accuracy of these estimates is indicated by the very small widths of the associated CIs relative to the estimated ground-truth values in all examples. The exact number of simulation runs used in our ISs, number of experimental
repetitions, and number of runs to approximate the ground truths are specified in the discussion of each example below.

## Large Deviations of an I.I.D. Sum.

For the experiment in Section 3.3.1, we use the asymptotically efficient estimator $\hat{\beta}(m)$ to obtain highly accurate estimates for all values of $m$ as the ground truths. These estimates are presented in Table 3.6. Our probabilistically efficient estimator $\hat{\alpha}(\mathrm{m})$ is computed using $10^{4}$ independently generated samples. From this, we apply CIs (3.12) and (3.13). We also construct CIs (3.13) using asymptotically efficient estimator $\hat{\beta}(m)$ (which is used to approximate the ground truth) with $10^{4}$ independently generated samples. We approximate the coverage rates using $10^{5}$ experimental iterations. Moreover, we compute the average CI width for each type of CIs. The experiment results are shown in Table 3.7.

From Table 3.7, we observe that the coverage rates of tight CIs by our probabilistically efficient estimator are close to $95 \%$ in three out of the four cases, but is $3 \%$ below $95 \%$ in one case ( $m=$ 10). On the other hand, the loose CIs are valid but perform conservatively with more than $99 \%$ coverage rates and wider average widths in all cases. The tight CIs by asymptotically efficient estimators provide valid coverage in all four cases. In the problems with rarer probabilities (i.e. $m=30,50,100$ ), the tight CIs by probabilistically efficient estimators perform similarly as the CIs by asymptotically efficient estimators in terms of both CI width and coverage. This shows the competitiveness of CIs using probabilistically efficient estimators for rarer problems.

Finally, we empirically verify the conditions in Theorem 3.8 needed to ensure the correctness of the tight confidence interval. In particular, the conditions require the sample size $n$ to grow subexponentially with a growth rate such that $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}\right)} \rightarrow 0$ and $\frac{\tilde{E}^{2}\left|Z_{1}-p_{1}\right|^{3}}{n \widehat{\operatorname{Var}}{ }^{3}\left(Z_{1}^{(1)}\right)} \rightarrow 0$. Here, we focus on checking the smallness of $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}\right)}$ and $\frac{\tilde{E}^{2}\left|Z_{1}-p_{1}\right|^{3}}{n \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}$ as an empirical validation. Table 3.8 presents the estimates of these quantities based on $10^{4}$ samples. We observe that all estimates are close to 0 , which explains the good performance of the tight confidence interval.

We note that our above empirical validation of the growth rate type conditions in Theorem 3.8
by using the smallness of the depicted quantities could be viewed as subjective (after all, there is a question of how small is regarded as small enough). However, we point out that even in the classical asymptotically optimal estimators, these conditions are still needed to ensure the validity of the corresponding CIs (see Theorem 3.9). From our best understanding, these conditions are usually ignored in the literature which focuses on asymptotic validity for large enough $n$ rather than studying the required growth rate as we have made more precise in our theorems.

Table 3.6: Highly accurate point estimates (and 95\% CI) using asymptotically efficient estimators for the problem in Section 3.3.1. The estimates are computed with $10^{7}$ samples for $m=10,30,50$ and $5 \times 10^{6}$ samples for $m=100$.

| $m$ | 10 | 30 | 50 | 100 |
| :--- | :--- | :--- | :--- | :--- |
| $\hat{p}$ | $8.85( \pm 0.0084) \times 10^{-3}$ | $1.58( \pm 0.0021) \times 10^{-5}$ | $3.76( \pm 0.0056) \times 10^{-8}$ | $1.34( \pm 0.0034) \times 10^{-14}$ |

Table 3.7: Coverage rates and average CI widths of the loose confidence intervals ("Loose CI by PE"), the tight confidence intervals ("Tight CI by PE") for probabilistically efficient estimators, and the tight confidence intervals for asymptotically efficient estimators ('Tight CI by AE") in the experiments of Section 3.3.1. These measures are estimated using $10^{5}$ macro replications.

|  | $m$ | 10 | 30 | 50 | 100 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Loose CI by PE | Coverage Rate | 0.998 | 0.999 | 0.9994 | 0.999 |
|  | Average Width | $9.12 \times 10^{-4}$ | $2.31 \times 10^{-6}$ | $6.37 \times 10^{-9}$ | $2.81 \times 10^{-15}$ |
| Tight CI by PE | Coverage Rate | 0.921 | 0.951 | 0.950 | 0.950 |
|  | Average Width | $5.30 \times 10^{-4}$ | $1.31 \times 10^{-6}$ | $3.54 \times 10^{-9}$ | $1.52 \times 10^{-15}$ |
| Tight CI by AE | Coverage Rate | 0.950 | 0.960 | 0.949 | 0.950 |
|  | Average Width | $5.30 \times 10^{-4}$ | $1.30 \times 10^{-6}$ | $3.55 \times 10^{-9}$ | $1.52 \times 10^{-15}$ |

Table 3.8: Verification of the conditions needed to ensure the correctness of the tight confidence intervals for the probabilistically efficient estimators in the example in Section 3.3.1. The estimates are computed with $10^{4}$ samples.

| $m$ | 10 | 30 | 50 | 100 |
| :--- | :--- | :--- | :--- | :--- |
| $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}\right)}$ | 0.00171 | 0.00243 | 0.00296 | 0.00412 |
| $\frac{\tilde{E}^{2}\left\|Z_{1}-p_{1}\right\|^{3}}{\left.n \widehat{\operatorname{Var}^{3}} Z_{1}^{(1)}\right)}$ | 0.00040 | 0.00073 | 0.00094 | 0.00149 |

## MNIST Example.

For the experiment in Section 3.3.3, we use $10^{10}$ runs of crude Monte Carlo to approximate the ground truths, which are shown in Table 3.9. Note that the estimate for $\sigma=0.17$ is relatively less accurate than other estimates, revealed by the CI width in the magnitude of around 0.1 of the probability estimate. We obtain our probabilistically efficient estimator by generating $10^{4}$ independent samples and construct CIs (3.12) and (3.13) based on this estimator. Since locating all dominating points to construct asymptotically efficient estimator is computationally infeasible in this example, we use IS estimators that mix the most significant 100 dominating points (the number of dominating points obtained from our sequential mixed integer programming procedure in Algorithm 3.2) as a proxy. We construct CI (3.13) from this estimator using $10^{4}$ samples. We use $10^{5}$ experimental repetitions to estimate the coverage rates and average widths of the CIs from probabilistically efficient estimators and $10^{3}$ repetitions for the CIs from IS estimators using 100 dominating points (we use repetition size $10^{3}$ instead of $10^{5}$ because of the long computational time caused by a large number of mixtures in the IS distribution). The results are presented in Table 3.10.

From Table 3.10, we observe that in three out of the four cases, the tight CIs constructed from probabilistically efficient estimators provide coverage rates that are slightly below 95\%. Similar to the previous random walk overshoot problem, the coverage rates are closer to $95 \%$ for rarer problems (e.g., the coverage is $94.9 \%$ for $\sigma=0.17$ ). The under-coverage is alleviated when we use more than one dominating point, as shown in the row of "Tight CI by AE" (where we use 100 dominating points). On the other hand, the loose CIs have higher than nominal coverage rates in all cases, but are conservative since the rates are around $97.5 \%-99.5 \%$. Again, we observe the validity of the CIs with probabilistically efficient estimators as the rare-event probability decreases, which validates our analysis.

To validate the tight confidence interval's correctness, we again empirically verify the conditions outlined in Theorem 3.8. Based on $10^{4}$ samples, we estimate the values of $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}\right)}$ and
$\frac{\tilde{E}^{2}\left|Z_{1}-p_{1}\right|^{3}}{n \widetilde{V a r^{3}}\left(Z_{1}^{(1)}\right)}$ which are shown in Table 3.11. We observe that all the estimates are close to 0 , which supports a good performance of the tight confidence interval in the rarer setting.

Table 3.9: Highly accurate point estimates (and 95\% CI) using crude Monte Carlo simulation with $10^{10}$ samples
for the problem in Section 3.3.3.

| $\sigma$ | 0.17 | 0.18 | 0.19 | 0.2 |
| :--- | :--- | :--- | :--- | :--- |
| $\hat{p}$ | $3.15( \pm 0.15) \times 10^{-7}$ | $1.29( \pm 0.031) \times 10^{-6}$ | $4.22( \pm 0.056) \times 10^{-6}$ | $1.17( \pm 0.0094) \times 10^{-5}$ |

Table 3.10: Coverage rates and average CI widths of the loose confidence intervals ("Loose CI by PE"), the tight confidence intervals ("Tight CI by PE") for probabilistically efficient estimators, and the tight confidence intervals for asymptotically efficient estimators ('"Tight CI by AE") in the experiments of Section 3.3.3. These measures for PE estimators are estimated using $10^{5}$ macro replications, while those for AE estimators are estimated using $10^{3}$ macro replications.

|  | $\sigma$ | 0.17 | 0.18 | 0.19 | 0.2 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Loose CI by PE | Coverage Rate | 0.996 | 0.978 | 0.980 | 0.977 |
|  | Average Width | $6.32 \times 10^{-8}$ | $2.61 \times 10^{-7}$ | $1.13 \times 10^{-6}$ | $3.20 \times 10^{-6}$ |
| Tight CI by PE | Coverage Rate | 0.949 | 0.874 | 0.885 | 0.877 |
|  | Average Width | $4.19 \times 10^{-8}$ | $1.73 \times 10^{-7}$ | $7.50 \times 10^{-7}$ | $2.12 \times 10^{-6}$ |
| Tight CI by AE | Coverage Rate | 0.958 | 0.933 | 0.945 | 0.951 |
|  | Average Width | $4.77 \times 10^{-8}$ | $1.96 \times 10^{-7}$ | $6.13 \times 10^{-7}$ | $1.90 \times 10^{-6}$ |

Table 3.11: Verification of the conditions needed to ensure the correctness of the tight confidence intervals for the probabilistically efficient estimators in the example in Section 3.3.3. The estimates are computed with $10^{4}$ samples.

| $\sigma$ | 0.17 | 0.18 | 0.19 | 0.2 |
| :--- | :--- | :--- | :--- | :--- |
| $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}\right)}$ | 0.00231 | 0.00203 | 0.00187 | 0.00151 |
| $\frac{\tilde{E}^{2}\left\|Z_{1}-p_{1}\right\|^{3}}{n \widehat{\operatorname{ar}}{ }^{3}\left(Z_{1}^{(1)}\right)}$ | 0.00252 | 0.00255 | 0.00322 | 0.00735 |

## Summary of Experimental Observations on Confidence Interval Construction.

From the CI construction for the two examples in Section 3.3 investigated above, we draw several conclusions: 1) Tight CIs by probabilistically efficient estimators appear to have close to the nominal coverage rate when the problem is rare enough; 2) Loose CIs by probabilistically efficient estimators tend to over-cover, and also have correspondingly larger widths than other methods;
3) Tight CIs by asymptotically efficient estimators appear to give more accurate coverage rates for a larger range of rarity levels than the tight CIs by probabilistically efficient estimators, even though they could still under-cover in some cases; 4) When tight CIs by probabilistically efficient and asymptotically efficient estimators both have accurate coverage rates, their widths appear to be comparable. Overall, it appears that tight CIs by asymptotically efficient estimators are more robust with respect to the rarity level of the problem, which is also in line with the comparison between Theorems 3.8 and 3.9 (recall the discussion right after Theorem 3.9). Nonetheless, recall that one motivation of us proposing the notion of probabilistic efficiency is that asymptotically efficient estimators, which require using more dominating points in their mixtures, could be computationally challenging to construct.

### 3.7 Future Work

We conclude the chapter with further discussions on the potential risk of the current framework and some future directions.

### 3.7.1 Developing Diagnosis for Finite-Sample Under-Estimation

Similar to the established notion of asymptotic efficiency in the rare-event simulation literature, our probabilistic efficiency framework is asymptotic. For a fixed rarity parameter $\gamma$ and given simulation size $n$ in practice, more work needs to be investigated to judge whether the obtained estimate is reliable or not. In particular, a risk of missing dominating points is finite-sample underestimation. To be more specific, we look at the following example.

Example 3.1. Suppose that our goal is to estimate $p=P\left(\frac{1}{\gamma} X_{\gamma} \in(-\infty,-k] \cup[1, \infty)\right)$ where $X_{\gamma} \sim N(0, \gamma)$ and $k>1$. In this case, the most significant dominating point is 1 , so we consider choosing $X_{\gamma} \sim N(\gamma, \gamma)$ as the IS distribution. That is, the IS estimator is $Z=I\left(\frac{1}{\gamma} X_{\gamma} \in(-\infty,-k] \cup\right.$ $[1, \infty)) e^{-X_{\gamma}+\gamma / 2}$ with $X_{\gamma} \sim N(\gamma, \gamma)$. We generate independent samples $Z^{(1)}, \ldots, Z^{(n)}$ and use $\hat{p}=\frac{1}{n} \sum_{i=1}^{n} Z_{i}$ to estimate $p$. When $1<k<3$, the relative error of $Z$ grows exponentially in $\gamma$ (or $-\log p$ ), so $Z$ is not asymptotically efficient (see the proof of Proposition 3.2). On the other
hand, from our previous derivations, $Z$ is strongly probabilistically efficient, so we can still get a reliable estimate when $\gamma$ is sufficiently large. However, in the finite-sample case where $k$ is close to 1 and $\gamma$ is not large, if $n$ is chosen as a moderate size, then it could happen that $\{X \leq-k \gamma\}$ is not hit and hence we would get an estimate close to $p_{1}=P\left(\frac{1}{\gamma} X_{\gamma} \geq 1\right)$, but $p_{2}=P\left(\frac{1}{\gamma} X_{\gamma} \leq-k\right)$ is not negligible compared to $p_{1}$. In other words, we under-estimate $p$. As a specific example, let $k=1.01$ and $\gamma=16$. Then $\tilde{p}_{2}=\tilde{P}\left(\frac{1}{16} X_{16} \leq-1.01\right)=\bar{\Phi}(8.04) \approx 4.44 \times 10^{-16}$. In this case, for a moderate $n$ like 1000, with probability almost 1 the set $\left\{\frac{1}{16} X_{16} \leq-1.01\right\}$ is never hit. However, $p_{2} / p_{1}=\bar{\Phi}(4.04) / \bar{\Phi}(4) \approx 0.84$. This means with probability almost 1 , we under-estimate $p$ by about $0.84 /(1+0.84) \approx 46 \%$.

Example 3.1 shows that under finite parameter value and finite sample, it could be hard to tell whether we can safely drop less significant dominating points. A good aspect about the conclusion in this example, however, is that the under-estimation is arguably acceptable in relative term (i.e., the estimate is still in the same magnitude as the ground truth), pointing to an estimation resembling weak probabilistic efficiency. We should emphasize that a similar concern applies to asymptotically efficient estimators as well. That is, it is difficult to guarantee whether the sample size $n$ is large enough to give a reliable estimate for a given setup and rarity parameter value. Nevertheless, increasing $n$ would improve the performance of asymptotically efficient estimators, but for probabilistic efficiency, we do not have the luxury of increasing $n$ since our framework requires $n$ to be moderate in size. This points to more need of developing diagnostic methods to detect under-estimation due to finite-sample effects in the future.

### 3.7.2 Further Developing Theory of IS with Missed Dominating Points

Despite the presence of under-estimation risks described above, we maintain our motivation of probabilistic efficiency as a theory to allow one to use few dominating points in problems where finding all of them is infeasible. In fact, what we have focused on in this chapter is only one theory where dropping dominating points is valid, among other possibilities. To support this, we revisit Example 3.1:

Example 3.2 (Example 3.1 continued). Consider the problem setting in Example 3.1. When $k \geq 3$, the relative error of $Z$ grows only polynomially in $\gamma(o r-\log p)$, and hence $Z$ is asymptotically efficient (see the proof of Proposition 3.2). The intuitive explanation is that $\tilde{p}_{2}=\tilde{P}\left(\frac{1}{\gamma} X_{\gamma} \leq-k\right)$ is extremely small which mitigates the blow-up of the likelihood ratio. More rigorously, when $\frac{1}{\gamma} X_{\gamma} \leq-k$ we have $Z \geq e^{\left(k+\frac{1}{2}\right) \gamma}$, but $\tilde{P}\left(\frac{1}{\gamma} X_{\gamma} \leq-k\right)=\bar{\Phi}((k+1) \sqrt{\gamma})=\Theta\left(\frac{1}{\sqrt{\gamma}} e^{-\frac{(k+1)^{2}}{2} \gamma}\right)$. Overall $\tilde{E}\left(Z^{2}\right)=\Theta\left(\frac{1}{\sqrt{\gamma}} e^{-\gamma}\right)$, and hence $\tilde{E}\left(Z^{2}\right) / p^{2}=\Theta(\sqrt{\gamma})$ which grows polynomially in $-\log p=\Theta(\gamma)$.

Example 3.2 shows that sometimes the missed dominating points are so rare that missing them does not even harm the asymptotic efficiency. Here, explaining the validity of IS with missed dominating points does not require probabilistic efficiency, but instead an alternate analysis of asymptotic efficiency that is tighter than the standard approach in the literature. Nonetheless, this phenomenon does not apply to our proposed estimators in Sections 3.3 and 3.6, as we have mathematically verified the asymptotic inefficiency in our considered estimator in each example.

More generally, we conclude our chapter with Figure 3.4, which shows the relations among different efficiency criteria and IS with missed dominating points. In this chapter, we have built sufficient conditions to achieve strong probabilistic efficiency with missed dominating points. Example 3.2, on the other hand, gives an example in achieving asymptotic efficiency with missed dominating points. Our immediate future endeavor is to fill in the regions in Figure 3.4 that are not covered by the current work, including the scenario depicted by Example 3.2, relaxing the current sufficient conditions of probabilistic efficiency to allow $\tilde{p}_{2}$ to be less tiny, achieving weak instead of strong probabilistic efficiency with missed dominating points, and moreover, to understand the "complementary" regions where missing dominating points would be guaranteed to violate the efficiency notions. Our longer-term goals include the creation of a unified theory that encompasses both the classical variance-based criteria and our probabilistic efficiency, using potentially new variability measures of suitably defined limiting quantities, and revisiting the interpretation of the variance reduction notion in IS and other competing techniques. In summary, this work serves as a first step in a new line of analysis aiming to relax existing variance-based efficiency criteria in


Figure 3.4: Relationships among different efficiency criteria and IS estimators missing dominating points.
rare-event simulation to be applicable to larger-scale and more complex problems.

### 3.8 Supplementary A: Algorithms

Algorithm 3.1 shows a procedure to obtain dominating sets. We briefly explain the idea here. First we minimize $I(x)$ over $x \in \mathcal{E}$ and get the first dominating point $a_{1}$. If $\mathcal{E} \subset\left\{x \in \mathbb{R}^{d}\right.$ : $\left.s_{a_{1}}^{\top}\left(x-a_{1}\right) \geq 0\right\}$, then we know that $\left\{a_{1}\right\}$ is a dominating set and hence we could stop. Otherwise, we minimize $I(x)$ over $x \in \mathcal{E} \backslash\left\{x \in \mathbb{R}^{d}: s_{a_{1}}^{\top}\left(x-a_{1}\right) \geq 0\right\}$ to get the second dominating point $a_{2}$. Then we check whether $\mathcal{E} \in \bigcup_{i=1}^{2}\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$. By repeating this process, we would finally get a dominating set $A=\left\{a_{1}, \ldots, a_{r}\right\}$ with $I\left(a_{1}\right) \leq \cdots \leq I\left(a_{r}\right)$.

### 3.9 Supplementary B: Additional Numerical Experiments

### 3.9.1 Overshoot Probability of a Random Walk.

We verify the conditions for probabilistic efficiency for the experiment in Section 3.3.2. Then we illustrate the proposed confidence intervals using the example.

In Section 3.3.2, we consider the most significant dominating point $a=a_{1}$ and our probabilistically efficient estimator is the exponential tilting towards the most significant dominating point only, i.e., $X_{\gamma}$ distributed as $N\left(\gamma a_{1}, \gamma \Sigma\right)$. We define $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ and

```
Algorithm 3.1: Sequentially find all the dominating points.
    Input: Rarity parameter \(\gamma\), rare-event set \(\mathcal{E} \subset \mathbb{R}^{d}\), rate function \(I(y)\), function \(a \mapsto s_{a}\).
    Output: Dominating set \(A\).
    Start with \(A=\emptyset, r=0\);
    While \(\left\{x \in \mathbb{R}^{d}: x \in \mathcal{E}, s_{a_{i}}^{\top}\left(x-a_{i}\right)<0, \forall i=1, \ldots, r\right\} \neq \emptyset\) do
        Find a dominating point \(a_{r+1}\) by solving the optimization problem
```

$$
\begin{array}{rl}
a_{r+1}=\arg \min _{x} & I(x) \\
\text { s.t. } & x \in \mathcal{E}, \\
& s_{a_{i}}^{\top}\left(x-a_{i}\right)<0, \forall i=1, \ldots, r
\end{array}
$$

and update $A \leftarrow A \cup\left\{a_{r+1}\right\}, r \leftarrow r+1$.
4 End
$\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$, and the corresponding probabilities $p_{1}=P\left(\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$ and $p_{2}=P\left(\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$. We compute $p_{1}, p_{2}$, and also the contribution of each of the nine less significant dominating points in $p_{2}$. More precisely, we define $\mathcal{B}_{\gamma}^{2}=\mathcal{A}_{\gamma} \cap\left\{s_{a_{2}}^{\top}\left(x-a_{2}\right) \geq\right.$ $0\} \backslash\left\{s_{a_{1}}^{\top}\left(x-a_{1}\right) \geq 0\right\}, \mathcal{B}_{\gamma}^{3}=\mathcal{A}_{\gamma} \cap\left\{s_{a_{3}}^{\top}\left(x-a_{3}\right) \geq 0\right\} \backslash\left(\left\{s_{a_{1}}^{\top}\left(x-a_{1}\right) \geq 0\right\} \cup\left\{s_{a_{2}}^{\top}\left(x-a_{2}\right) \geq 0\right\}\right), \ldots$, $\mathcal{B}_{\gamma}^{10}=\mathcal{A}_{\gamma} \cap\left\{s_{a_{10}}^{\top}\left(x-a_{10}\right) \geq 0\right\} \backslash\left(\cup_{j=1}^{9}\left\{s_{a_{j}}^{\top}\left(x-a_{j}\right) \geq 0\right\}\right)$, and use $p_{a_{2}}=P\left(\mathcal{B}_{\gamma}^{2}\right), . ., p_{a_{10}}=P\left(\mathcal{B}_{\gamma}^{10}\right)$ to denote the contribution of dominating points $a_{2}, \ldots, a_{10}$ (with decreasing significance). For each probability $p_{1}, p_{a_{2}}, \ldots, p_{a_{10}}$, we construct an IS estimator using the "corresponding" dominating points $a_{1}, \ldots, a_{10}$, e.g., for $p_{a_{2}}$ the IS distribution is mean shifted to $a_{2}$. Then we estimate $p_{2}$ through $p_{2}=p_{a_{2}}+\cdots+p_{a_{10}}$ and $p$ through $p=p_{1}+p_{2}$. Table 3.12 presents the results estimated using independently generated $10^{4}$ samples from the corresponding IS distributions. We observe that $p_{2} / p$ has larger values than those in the previous experiment, in that $p_{2} / p \approx 0.25$ at $\sigma=0.3$ and $p_{2} / p \approx 0.15$ at $\sigma=0.2$. Nonetheless, $p_{2} / p$ 's value decreases rapidly as $\sigma$ decreases, i.e., the problem becomes rarer, which suggests the trend $p_{2} / p \rightarrow 0$ in Condition 1 of Theorem 3.3. Additionally, we observe from the values of $p_{a_{2}} / p, \ldots, p_{a_{10}} / p$ that the contribution of each less significant dominating point vanishes rapidly with decreasing $\sigma$.

Next, we present the probabilities $\tilde{p}_{1}=\tilde{P}\left(\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}\right)$ and $\tilde{p}_{2}=\tilde{P}\left(\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq\right.\right.$ $0\}$ ) under the probabilistically efficient IS distribution. For $\tilde{p}_{2}$, we also present the contributions

Table 3.12: Estimates of $p_{1}, p_{2}$ and the contributions of the less significant dominating points for the random walk example in Section 3.3.2 with $10^{4}$ samples, where we use $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq\right.$ $0\}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$.

| $\sigma$ | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $p_{1}$ | $2.42 \times 10^{-4}$ | $9.07 \times 10^{-5}$ | $2.82 \times 10^{-5}$ | $6.12 \times 10^{-6}$ | $9.79 \times 10^{-7}$ | $7.92 \times 10^{-8}$ |
| $p_{1} / p$ | 0.7434 | 0.7512 | 0.7776 | 0.7960 | 0.8311 | 0.8549 |
| $p_{2}$ | $8.36 \times 10^{-5}$ | $3.00 \times 10^{-5}$ | $8.05 \times 10^{-6}$ | $1.57 \times 10^{-6}$ | $1.99 \times 10^{-7}$ | $1.34 \times 10^{-8}$ |
| $p_{2} / p$ | 0.2566 | 0.2488 | 0.2224 | 0.2040 | 0.1689 | 0.1451 |
| $p$ | $3.26 \times 10^{-4}$ | $1.21 \times 10^{-4}$ | $3.62 \times 10^{-5}$ | $7.66 \times 10^{-6}$ | $1.18 \times 10^{-6}$ | $9.26 \times 10^{-8}$ |
| $p_{a_{2}}$ | $5.29 \times 10^{-5}$ | $1.95 \times 10^{-5}$ | $5.50 \times 10^{-6}$ | $1.14 \times 10^{-6}$ | $1.56 \times 10^{-7}$ | $1.11 \times 10^{-8}$ |
| $p_{a_{2}} / p$ | 0.1625 | 0.1611 | 0.1519 | 0.1489 | 0.1323 | 0.1194 |
| $p_{a_{3}}$ | $2.06 \times 10^{-5}$ | $7.46 \times 10^{-6}$ | $1.88 \times 10^{-6}$ | $3.32 \times 10^{-7}$ | $3.57 \times 10^{-8}$ | $2.06 \times 10^{-9}$ |
| $p_{a_{3}} / p$ | 0.0632 | 0.0618 | 0.0519 | 0.0432 | 0.0303 | 0.0223 |
| $p_{a_{4}}$ | $7.35 \times 10^{-6}$ | $2.48 \times 10^{-6}$ | $5.59 \times 10^{-7}$ | $7.85 \times 10^{-8}$ | $6.69 \times 10^{-9}$ | $2.96 \times 10^{-10}$ |
| $p_{a_{4}} / p$ | 0.0226 | 0.0205 | 0.0154 | 0.0102 | 0.0057 | 0.0032 |
| $p_{a_{5}}$ | $2.23 \times 10^{-6}$ | $5.67 \times 10^{-7}$ | $1.02 \times 10^{-7}$ | $1.19 \times 10^{-8}$ | $6.86 \times 10^{-10}$ | $1.95 \times 10^{-11}$ |
| $p_{a_{5}} / p$ | 0.0069 | 0.0047 | 0.0028 | 0.0015 | $5.82 \times 10^{-4}$ | $2.11 \times 10^{-4}$ |
| $p_{a_{6}}$ | $4.18 \times 10^{-7}$ | $8.07 \times 10^{-8}$ | $1.13 \times 10^{-8}$ | $1.00 \times 10^{-9}$ | $3.53 \times 10^{-11}$ | $4.58 \times 10^{-13}$ |
| $p_{a_{6}} / p$ | 0.0013 | $6.68 \times 10^{-4}$ | $3.11 \times 10^{-4}$ | $1.30 \times 10^{-4}$ | $3.00 \times 10^{-5}$ | $4.94 \times 10^{-6}$ |
| $p_{a_{7}}$ | $3.51 \times 10^{-8}$ | $5.00 \times 10^{-9}$ | $4.07 \times 10^{-10}$ | $1.84 \times 10^{-11}$ | $3.32 \times 10^{-13}$ | $1.75 \times 10^{-15}$ |
| $p_{a_{7}} / p$ | $1.08 \times 10^{-4}$ | $4.14 \times 10^{-5}$ | $1.12 \times 10^{-5}$ | $2.39 \times 10^{-6}$ | $2.82 \times 10^{-7}$ | $1.89 \times 10^{-8}$ |
| $p_{a_{8}}$ | $5.43 \times 10^{-10}$ | $3.95 \times 10^{-11}$ | $1.43 \times 10^{-12}$ | $2.81 \times 10^{-14}$ | $1.58 \times 10^{-16}$ | $1.98 \times 10^{-19}$ |
| $p_{a_{8}} / p$ | $1.67 \times 10^{-6}$ | $3.28 \times 10^{-7}$ | $3.94 \times 10^{-8}$ | $3.66 \times 10^{-9}$ | $1.34 \times 10^{-10}$ | $2.14 \times 10^{-12}$ |
| $p_{a_{9}}$ | $1.32 \times 10^{-13}$ | $2.93 \times 10^{-15}$ | $2.50 \times 10^{-17}$ | $7.30 \times 10^{-20}$ | $3.43 \times 10^{-23}$ | $1.46 \times 10^{-27}$ |
| $p_{a_{9}} / p$ | $4.05 \times 10^{-10}$ | $2.43 \times 10^{-11}$ | $6.91 \times 10^{-13}$ | $9.50 \times 10^{-15}$ | $2.91 \times 10^{-17}$ | $1.57 \times 10^{-20}$ |
| $p_{a_{10}}$ | $2.35 \times 10^{-24}$ | $1.37 \times 10^{-27}$ | $1.36 \times 10^{-31}$ | $1.25 \times 10^{-36}$ | $3.95 \times 10^{-43}$ | $1.37 \times 10^{-51}$ |
| $p_{a_{10}} / p$ | $7.21 \times 10^{-21}$ | $1.13 \times 10^{-23}$ | $3.75 \times 10^{-27}$ | $1.63 \times 10^{-31}$ | $3.36 \times 10^{-37}$ | $1.48 \times 10^{-44}$ |

from the dominating points $a_{2}, \ldots, a_{10}$, denoted as $\tilde{p}_{a_{2}}, \ldots, \tilde{p}_{a_{10}}$ with

$$
\tilde{p}_{a_{i}}=\tilde{P}\left(\mathcal{A}_{\gamma} \cap\left\{s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\} \backslash\left(\cup_{j=1}^{i-1}\left\{s_{a_{j}}^{\top}\left(x-a_{j}\right) \geq 0\right\}\right)\right)
$$

for $i=2, \ldots, 10$. The probabilities are estimated using the proportion of samples falling into the corresponding sets based on $10^{4}$ samples drawn from the probabilistically efficient IS distribution. The results are presented in Table 3.13. We observe that $\tilde{p}_{2}$ generally decreases from 0.0129 at $\sigma=0.3$ to 0.0039 with $\sigma=0.2$. Furthermore, the decreasing trends also appear in each individual contribution from the less significant dominating points, where most of the probabilities (e.g., $\tilde{p}_{a_{4}}, \ldots, \tilde{p}_{a_{10}}$ ) already vanish when $\sigma=0.2$. Based on the value of $\tilde{p}_{2}$, we estimate the
probability $\tilde{P}$ (some of the $n$ samples hits $\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ ) through $1-\left(1-\tilde{p}_{2}\right)^{n}$. We denote this probability as $\tilde{p}_{h i t}$ and present the results with $n=10^{4}$ (the sample size we use in Section 3.3) in the last row of Table 3.13. We observe that there are samples falling into $\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ with approximately probability 1 when we use $n=10^{4}$ samples. This close-to- 1 probability, unfortunately, is quite different from what our Condition 3 in Theorem 3.3 would entail and cannot explain the good performance of our probabilistically efficient estimator.

Table 3.13: Estimates of $\tilde{p}_{1}, \tilde{p}_{2}$ and the contributions of the less significant dominating points under probabilistically efficient IS for the random walk example in Section 3.3.2 with $10^{4}$ samples, where we use $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{s_{a}^{\top}(x-a) \geq 0\right\}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash\left\{s_{a}^{\top}(x-a) \geq 0\right\}$.

| $\sigma$ | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $\tilde{p}_{1}$ | 0.5057 | 0.4916 | 0.496 | 0.5053 | 0.4969 | 0.4985 |
| $\tilde{p}_{2}$ | 0.0129 | 0.0141 | 0.0086 | 0.0076 | 0.0057 | 0.0039 |
| $\tilde{p}_{a_{2}}$ | 0.0111 | 0.0111 | 0.007 | 0.0062 | 0.0048 | 0.0034 |
| $\tilde{p}_{a_{3}}$ | 0.0011 | 0.0024 | 0.0013 | 0.0013 | 0.0008 | 0.0005 |
| $\tilde{p}_{a_{4}}$ | 0.0006 | 0.0006 | 0.0002 | 0.0001 | 0.0001 | 0 |
| $\tilde{p}_{a_{5}}$ | 0.0001 | 0 | 0.0001 | 0 | 0 | 0 |
| $\tilde{p}_{a_{6}}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{7}}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{8}}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{9}}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\tilde{p}_{a_{10}}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\tilde{p}_{h i t}$ | $\approx 1$ | $\approx 1$ | $\approx 1$ | $\approx 1$ | $\approx 1$ | $\approx 1$ |

To explain the good performance of our approach, we consider an alternative construction of $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash \mathcal{A}_{\gamma}^{1}$ to explain our performance. From the proofs of Propositions 3.4 and 3.5 , we know that our probabilistically efficient estimator is not asymptotically efficient if and only if $\min _{x \in \mathcal{A}_{\gamma}}\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right)<4 a_{1}^{\top} \Sigma^{-1} a_{1}$. In other words, if we split the rareevent set $\mathcal{A}_{\gamma}$ into two parts, say $\mathcal{A}_{\gamma}^{1}=\left\{\mathcal{A}_{\gamma} \cap\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}\right\}$ and $\mathcal{A}_{\gamma}^{2}=$ $\left\{\mathcal{A}_{\gamma} \backslash\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}\right\}$, then our probabilistically efficient estimator is asymptotically efficient for estimating $P\left(\mathcal{A}_{\gamma}^{1}\right)$ because $\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} \Sigma^{-1} a_{1}$ for all $x \in \mathcal{A}_{\gamma}^{1}$. This implies that, with these choices of $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$, we satisfy Condition 2 in Theorem 3.3 (since the IS estimator using the most significant dominating point is asymptotically efficient for estimating $P\left(\mathcal{F}_{\gamma}^{1}\right)$ ).

Next we check Conditions 1 and 3 in Theorem 3.3. We define $p_{1}=P\left(\mathcal{A}_{\gamma}^{1}\right), p_{2}=P\left(\mathcal{A}_{\gamma}^{2}\right)$, $\tilde{p}_{1}=\tilde{P}\left(\mathcal{A}_{\gamma}^{1}\right)$, and $\tilde{p}_{2}=\tilde{P}\left(\mathcal{A}_{\gamma}^{2}\right)$ for our newly constructed $\mathcal{A}_{\gamma}^{1}$ and $\mathcal{A}_{\gamma}^{2}$. The use of our newly constructed $\mathcal{A}_{\gamma}^{1}$ for Theorem 3.3 can be theoretically shown to satisfy Conditions 1 and 3 therein like in Section 3.6.1. To estimate $p_{2}$, we construct an IS estimator that mixes the exponential tiltings towards the dominating points for $\mathcal{H}_{i} \cap\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \leq 4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}$ with $i=$ $1, \ldots, 10$. To estimate $\tilde{p}_{1}$ and $\tilde{p}_{2}$, we directly generate samples from the probabilistically efficient IS distribution. We define $\tilde{p}_{\text {hit }}=\tilde{P}\left(\right.$ some of the $n$ samples hits $\left.\mathcal{A}_{\gamma}^{2}\right)$ and estimate $\tilde{p}_{\text {hit }}$ through $1-\left(1-\tilde{p}_{2}\right)^{n}$ with $n=10^{4}$. The results are presented in Table 3.14. We observe that the values of $p_{2} / p$ are now extremely small (smaller than $10^{-8}$ in all cases). Furthermore, the values of $\tilde{p}_{2}$ are also small, which lead to $\tilde{p}_{h i t}<0.01$ in all cases when $\sigma$ varies from 0.2 to 0.3 . These results now justify Conditions 1 and 3 of Theorem 3.3 and explain the good performance of our probabilistically efficient estimator in the experiment.

Table 3.14: Estimates of $p_{2}$ with $10^{4}$ samples, $\tilde{p}_{1}$ and $\tilde{p}_{2}$ with $10^{7}$ samples, and $\tilde{p}_{\text {hit }}$ with $n=10^{4}$ for random walk example in Section 3.3.2, where we use $\mathcal{A}_{\gamma}^{1}=\left\{\mathcal{A}_{\gamma} \cap\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \geq\right.\right.$ $\left.\left.4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}\right\}$ and $\mathcal{A}_{\gamma}^{2}=\left\{\mathcal{A}_{\gamma} \backslash\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}\right\}$.

| $\sigma$ | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $p_{2}$ | $3.85 \times 10^{-13}$ | $2.58 \times 10^{-12}$ | $9.77 \times 10^{-15}$ | $8.58 \times 10^{-18}$ | $2.32 \times 10^{-22}$ | $1.90 \times 10^{-30}$ |
| $p_{2} / p$ | $1.19 \times 10^{-9}$ | $2.15 \times 10^{-8}$ | $2.72 \times 10^{-10}$ | $1.08 \times 10^{-12}$ | $2.00 \times 10^{-16}$ | $1.89 \times 10^{-23}$ |
| $\tilde{p}_{1}$ | 0.5149 | 0.5121 | 0.5098 | 0.5078 | 0.5059 | 0.5038 |
| $\tilde{p}_{2}$ | $3 \times 10^{-7}$ | $5 \times 10^{-7}$ | $7 \times 10^{-7}$ | $1 \times 10^{-7}$ | $1 \times 10^{-7}$ | $4 \times 10^{-7}$ |
| $\tilde{p}_{\text {hit }}$ | 0.003 | 0.005 | 0.007 | 0.001 | 0.001 | 0.004 |

To estimate the coverage of the constructed CIs, we use the asymptotically efficient estimator that mixes all dominating points to approximate the ground truths presented in Table 3.15. Our probabilistically efficient estimator is computed using $10^{4}$ independently generated samples. We construct CIs (3.12) and (3.13) based on this estimator. For comparison we also construct CI (3.13) from asymptotically efficient estimator using $10^{4}$ samples independently generated from the ones used to approximate the ground truth. We use $10^{5}$ experimental repetitions to estimate the coverage rates of all CIs. The coverage rates and average widths are presented in Table 3.16.

From Table 3.16, the loose CIs perform conservatively with near to $99 \%$ coverage rates in all
cases. On the other hand, the tight CIs constructed from our probabilistically efficient estimators have coverage rates below $95 \%$ in most of the cases. Moreover, as $\sigma$ decreases (the probability become rarer), the coverage rates first drop from around 0.93 (with $\sigma=0.3$ ) to around 0.89 (with $\sigma=0.2$ ), then they improve as $\sigma$ further decreases and reaches around $95 \%$ when $\sigma=0.1$. The tight CIs by asymptotically efficient estimators have more stable coverage rates than the CIs by probabilistically efficient estimators, but also suffer under-coverage in several cases (e.g., 0.86 with $\sigma=0.28$ ). We also observe that the tight CIs by the probabilistically efficient estimators have better average widths than the CIs by the asymptotically efficient estimators with smaller $\sigma$ (e.g., $\sigma=0.1,0.12$ ). The results show the validity of the CIs with probabilistically efficient estimators as $\sigma \rightarrow 0$, but also that the coverage rate may not always monotonically improve as the problem becomes rarer.

Table 3.15: Higly accurate point estimates (and 95\% CI) using asymptotically efficient estimators for the problem in Section 3.3.2. The estimates are computed with $10^{7}$ samples.

| $\sigma$ | $\hat{p}$ |
| :--- | :--- |
| 0.1 | $(5.57 \pm 0.039) \times 10^{-26}$ |
| 0.12 | $(1.30 \pm 0.008) \times 10^{-18}$ |
| 0.14 | $(3.82 \pm 0.020) \times 10^{-14}$ |
| 0.16 | $(3.18 \pm 0.015) \times 10^{-11}$ |
| 0.18 | $(3.32 \pm 0.014) \times 10^{-9}$ |
| 0.2 | $9.51( \pm 0.035) \times 10^{-8}$ |
| 0.22 | $7.93( \pm 0.004) \times 10^{-6}$ |
| 0.24 | $1.55( \pm 0.010) \times 10^{-5}$ |
| 0.26 | $3.59( \pm 0.024) \times 10^{-6}$ |
| 0.28 | $1.20( \pm 0.003) \times 10^{-4}$ |
| 0.3 | $3.22( \pm 0.009) \times 10^{-4}$ |

### 3.9.2 Two-sided Overshoot Probability of a Random Walk.

So far we have considered examples on strongly probabilistically efficient estimators. Here, we consider an additional example where we use a weakly probabilistically efficient estimator. We follow the problem setting in Section 3.3.2, where we consider the overshoot probability of the

Table 3.16: Coverage rates and average CI widths of the loose confidence intervals ("Loose CI by PE"), the tight confidence intervals ("Tight CI by PE") for probabilistically efficient estimators, and the tight confidence intervals for asymptotically efficient estimators ("Tight CI by AE") in the experiments of Section 3.3.2.

|  | Loose CI by PE |  | Tight CI by PE |  | Tight CI by AE |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $\sigma$ | Coverage | Width | Coverage | Width | Coverage | Width |
| 0.1 | 0.9997 | $1.49 \times 10^{-26}$ | 0.949 | $7.88 \times 10^{-27}$ | 0.926 | $2.41 \times 10^{-26}$ |
| 0.12 | 0.999 | $3.58 \times 10^{-19}$ | 0.938 | $1.99 \times 10^{-19}$ | 0.917 | $5.08 \times 10^{-19}$ |
| 0.14 | 0.997 | $1.07 \times 10^{-14}$ | 0.914 | $6.15 \times 10^{-15}$ | 0.951 | $1.32 \times 10^{-14}$ |
| 0.16 | 0.994 | $9.37 \times 10^{-12}$ | 0.897 | $5.54 \times 10^{-12}$ | 0.908 | $9.67 \times 10^{-12}$ |
| 0.18 | 0.990 | $9.89 \times 10^{-10}$ | 0.892 | $5.95 \times 10^{-10}$ | 0.951 | $8.72 \times 10^{-10}$ |
| 0.2 | 0.988 | $2.86 \times 10^{-8}$ | 0.895 | $1.75 \times 10^{-8}$ | 0.964 | $2.24 \times 10^{-8}$ |
| 0.22 | 0.988 | $3.15 \times 10^{-7}$ | 0.901 | $1.93 \times 10^{-7}$ | 0.936 | $2.45 \times 10^{-7}$ |
| 0.24 | 0.990 | $2.20 \times 10^{-6}$ | 0.913 | $1.35 \times 10^{-6}$ | 0.959 | $1.55 \times 10^{-6}$ |
| 0.26 | 0.991 | $9.67 \times 10^{-6}$ | 0.918 | $6.00 \times 10^{-6}$ | 0.932 | $6.49 \times 10^{-6}$ |
| 0.28 | 0.992 | $2.99 \times 10^{-5}$ | 0.924 | $1.86 \times 10^{-5}$ | 0.861 | $2.01 \times 10^{-5}$ |
| 0.3 | 0.993 | $7.71 \times 10^{-5}$ | 0.927 | $4.80 \times 10^{-5}$ | 0.956 | $4.90 \times 10^{-5}$ |

finite-horizon maximum of a random walk. However, we modify the probability of interest as

$$
\begin{equation*}
p=P\left(\max _{m=1, \ldots, d}\left|S_{m}\right| \geq a\right), \tag{3.14}
\end{equation*}
$$

where we replace $S_{m}=\sum_{i=1}^{m} Y_{i}$ by its absolute value. The rest of the settings are the same as in Section 3.3.2, i.e., we have $Y_{i}$ 's are Gaussian distributed with mean 0, standard deviation $\sigma$, pairwise correlation -0.02 , and rarity parameter $\gamma=1 / \sigma^{2} \rightarrow \infty$. The target rare event is $\left\{\frac{1}{\gamma} X_{\gamma} \in\left(\bigcup_{m=1}^{d} \mathcal{H}_{m}^{+}\right) \cup\left(\bigcup_{m=1}^{d} \mathcal{H}_{m}^{-}\right)\right\}$where $X_{\gamma}=\gamma\left(Y_{1}, \ldots, Y_{d}\right)^{\top}, \mathcal{H}_{m}^{+}=\left\{x \in \mathbb{R}^{d}: \sum_{i=1}^{m} x_{i} \geq a\right\}$, and $\mathcal{H}_{m}^{-}=\left\{x \in \mathbb{R}^{d}: \sum_{i=1}^{m} x_{i} \leq-a\right\}$ with $x_{i}$ denoting the $i$ th element in $x$. In this case, the rate function is still $I(y)=\frac{1}{2} y^{\top} \Sigma^{-1} y$ and there are two most significant dominating points $a_{1}=\frac{a \Sigma e_{d}}{e_{d}^{\top} \Sigma e_{d}}$ and $-a_{1}=-\frac{a \Sigma e_{d}}{e_{d}^{\top} \Sigma e_{d}}$ where $e_{d} \in \mathbb{R}^{d}$ denotes the vector with 1 in all $d$ elements.

For this experiment, we first introduce an asymptotically efficient estimator. From Proposition 3.1, we know that the IS estimator using dominating points $a_{1}, \ldots, a_{d},-a_{1}, \ldots,-a_{d}$ with $a_{1}, \ldots, a_{d}$ defined in Section 3.3.2 is asymptotically efficient. Next, we show the IS estimator using the dominating point $a_{1}$ is a weakly probabilistically efficient estimator and is not asymptotically efficient:

Theorem 3.11. Under the problem specifications in Section 3.3.2 and rare-event probability defined in (3.14), the IS estimator that use only one of the most significant dominating points, namely $a_{1}$ in Section 3.3.2, and $X_{\gamma}$ distributed as $N\left(\gamma a_{1}, \gamma \Sigma\right)$ in Section 3.3.2, is weakly probabilistically efficient but not asymptotically efficient.

Compared to the one-sided overshoot example in Section 3.3.2, here the rare-event set has two most significant dominating points $a_{1}$ and $-a_{1}$. Because we only use the first one instead of mixing both of the most significant dominating points in our IS, we only have $p_{1} / p \rightarrow 1 / 2$ instead of 0 and thus weak probabilistic efficiency instead of strong probabilistic efficiency holds as guided by Theorem 3.4.

To empirically verify Theorem 3.4, let us consider a partition of the rare event set

$$
\mathcal{A}_{\gamma}=\left\{\left(\bigcup_{m=1}^{d} \mathcal{H}_{m}^{+}\right) \bigcup\left(\bigcup_{m=1}^{d} \mathcal{H}_{m}^{-}\right)\right\},
$$

where we have $\mathcal{A}_{\gamma}^{1}=\mathcal{A}_{\gamma} \cap\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}\left(x+a_{1}\right) \geq 4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}$ and $\mathcal{A}_{\gamma}^{2}=\mathcal{A}_{\gamma} \backslash\left\{\left(x+a_{1}\right)^{\top} \Sigma^{-1}(x+\right.$ $\left.\left.a_{1}\right) \geq 4 a_{1}^{\top} \Sigma^{-1} a_{1}\right\}$. We define $p_{1}=P\left(\mathcal{A}_{\gamma}^{1}\right), p_{2}=P\left(\mathcal{A}_{\gamma}^{2}\right), \tilde{p}_{1}=\tilde{P}\left(\mathcal{A}_{\gamma}^{1}\right)$, and $\tilde{p}_{2}=\tilde{P}\left(\mathcal{A}_{\gamma}^{2}\right)$, where $\tilde{P}$ is the IS distribution exponentially tilted using the dominating point $a_{1}$. In our experiments, we set $d=10$, fix $a_{1},-a_{1}$ and vary $\sigma$ for different rarity levels. For each case, we generate $10^{4}$ samples from IS distributions using the above asymptotically efficient estimator and weakly probabilistically efficient estimator. The results are presented in Table 3.17. We observe that although our weakly probabilistically efficient estimator underestimates the rare-event probability in all considered cases, the estimates have relatively tight CIs and provide a good estimation on the magnitude of the rare-event probability, i.e., the estimates are around 0.5 of the estimates given by the asymptotically efficient estimator.

In Table 3.18, we investigate the numerical values of $p_{1}, p_{2}, \tilde{p}_{1}$ and $\tilde{p}_{2}$ and check the values of $p_{1} / p$ and $\tilde{p}_{h i t}=\tilde{P}$ (some of the $n$ samples hits $\mathcal{A}_{\gamma}^{2}$ ) with $n=10^{4}$. We estimate $p_{1}$ and $p_{2}$ using the asymptotically efficient estimator with $10^{7}$ independently generated samples. For the estimation of $\tilde{p}_{1}$ and $\tilde{p}_{2}$, we generate $10^{7}$ samples using the weakly probabilistically efficient IS distribution.

Table 3.17: Point estimates (and 95\% CIs) from the asymptotically efficient estimator and the weakly probabilistically efficient estimator for the two-sided overshoot probability. "AE" denotes the asymptotically efficient estimator and "PE" denotes the weakly probabilistically efficient estimator.

| $\sigma$ | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| AE | $6.15( \pm 0.47) \times 10^{-4}$ | $2.52( \pm 0.20) \times 10^{-4}$ | $7.53( \pm 0.68) \times 10^{-5}$ | $1.56( \pm 0.15) \times 10^{-5}$ | $2.26( \pm 0.24) \times 10^{-6}$ | $1.93( \pm 0.24) \times 10^{-7}$ |
| PE | $3.18( \pm 0.18) \times 10^{-4}$ | $1.19( \pm 0.08) \times 10^{-4}$ | $3.57( \pm 0.29) \times 10^{-5}$ | $8.00( \pm 0.64) \times 10^{-6}$ | $1.17( \pm 0.11) \times 10^{-6}$ | $9.43( \pm 0.63) \times 10^{-8}$ |
| PE/AE | 0.516 | 0.470 | 0.474 | 0.512 | 0.517 | 0.488 |

We observe that in all cases the values of $p_{1} / p$ are very close to $1 / 2$. On the other hand, the probabilities of falling into $\mathcal{A}_{\gamma}^{2}$ are all below $10^{-6}$, which lead to $\tilde{p}_{\text {hit }}$ valued smaller than 0.01 . These results verify the conditions in Theorem 3.4 that explain the weak probabilistic efficiency of the IS estimator.

Table 3.18: Estimates of $p_{1}, p_{2}, \tilde{p}_{1}$ and $\tilde{p}_{2}$ with $10^{7}$ samples, and $\tilde{p}_{\text {hit }}$ with $n=10^{4}$ for two-sided overshoot probability.

| $\sigma$ | 0.3 | 0.28 | 0.26 | 0.24 | 0.22 | 0.2 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $p_{1}$ | $3.23 \times 10^{-4}$ | $1.20 \times 10^{-4}$ | $3.58 \times 10^{-5}$ | $7.91 \times 10^{-6}$ | $1.17 \times 10^{-6}$ | $9.48 \times 10^{-8}$ |
| $p_{1} / p$ | 0.501 | 0.500 | 0.500 | 0.499 | 0.501 | 0.499 |
| $p_{2}$ | $3.22 \times 10^{-4}$ | $1.20 \times 10^{-4}$ | $3.59 \times 10^{-5}$ | $7.94 \times 10^{-6}$ | $1.16 \times 10^{-6}$ | $9.51 \times 10^{-8}$ |
| $p$ | $6.45 \times 10^{-4}$ | $2.40 \times 10^{-4}$ | $7.17 \times 10^{-5}$ | $1.59 \times 10^{-5}$ | $2.33 \times 10^{-6}$ | $1.90 \times 10^{-7}$ |
| $\tilde{p}_{1}$ | 0.515 | 0.512 | 0.510 | 0.507 | 0.506 | 0.504 |
| $\tilde{p}_{2}$ | $2 \times 10^{-7}$ | $6 \times 10^{-7}$ | $5 \times 10^{-7}$ | $6 \times 10^{-7}$ | $4 \times 10^{-7}$ | $6 \times 10^{-7}$ |
| $\tilde{p}_{\text {hit }}$ | 0.002 | 0.006 | 0.005 | 0.006 | 0.004 | 0.006 |

### 3.10 Supplementary C: A New Alternative Asymptotic Regime and Theoretical Guarantees

In this section, we investigate a new regime that $\mathcal{A}_{\gamma}=\{X \in \mathcal{E}\}$ and $\mathcal{E}=\left\{x \in \mathbb{R}^{d}: g(x) \geq \gamma\right\}$ where $X \in \mathbb{R}^{d}$ and $g: \mathbb{R}^{d} \rightarrow \mathbb{R}$ is a function. We propose this setting as it arises as a generic representation of recent problems in intelligent system safety testing [125]. There, $g$ could be highly complicated and leads to a gigantic number of dominating points, which in turn motivates the consideration of dropping most of them and our notion of probabilistic efficiency. We note that technically this setting is slightly different from the classical Gartner-Ellis regime in terms of the position of the scaling parameter $\gamma$ (see, e.g., [53]), but conceptually similar.

In Section 3.10.1, we adapt the notions of rate function and dominating points to this new
regime. Then we state the assumptions under which we could build on Theorem 3.3 to obtain reliable point estimates and CIs as in Section 3.5. We note that this new regime is harder to analyze as the rare-event set $\mathcal{E}$ can change in a complicated way with $\gamma$. For instance, unlike in the Gartner-Ellis regime, now the number of dominating points can change with $\gamma$. Thus, inevitably our assumptions are relatively restrictive and need to be verified case by case. In Section 3.10.2, we consider the special (but important) case where $X$ follows a Gaussian distribution and $g$ is piecewise linear, in particular propose a simple stopping strategy to determine whether it is safe to stop searching for the remaining dominating points. Throughout this section, we write $\alpha(\gamma) \sim$ $\beta(\gamma)$ if $\alpha(\gamma) / \beta(\gamma)$ is subexponential in $-\log p(\gamma)$.

### 3.10.1 Guarantees for General Input Distribution

We define $\mu(x)=\log E e^{x^{\top} X}, x \in \mathbb{R}^{d}$ as the cumulant generating function of $X$ and $I(y)=$ $\sup _{x \in \mathbb{R}^{d}}\left\{x^{\top} y-\mu(x)\right\}, y \in \mathbb{R}^{d}$ as its Legendre transform. For any set $\mathcal{E} \subset \mathbb{R}^{d}$, we denote $I(\mathcal{E})=$ $\inf _{y \in \mathcal{E}} I(y)$. Parallel to Assumptions 3.1 and 3.2, we make the following two assumptions.

Assumption 3.3. $\mu(x)$ satisfies the following conditions:

1. $0 \in \mathcal{D}(\mu)^{\circ}$;
2. $\mu$ is essentially smooth, i.e., $\mathcal{D}(\mu)^{\circ}$ is non-empty, $\mu$ is differentiable everywhere in $\mathcal{D}(\mu)^{\circ}$ and $\mu$ is steep.

Assumption 3.4. For any $\gamma, \mathcal{E}=\mathcal{E}(\gamma) \subset \mathbb{R}^{d}$ is a Borel set such that $\overline{\mathcal{E}}=\overline{\mathcal{E}^{\circ}, \mathcal{E}^{\circ} \cap \mathcal{D}(I)^{\circ} \neq \emptyset \text { and }}$ $I(\mathcal{E})>0$.

The concepts of dominating set and dominating points are also similar, but now they change with $\gamma$ :

Definition 3.5 (Dominating Set (New Regime)). Suppose that Assumptions 3.3 and 3.4 hold. We call $A=A(\gamma) \subset \partial \mathcal{E}$ a dominating set for $\mathcal{E}$ if

1. For each $a \in A$, there exists a unique $s_{a} \in \mathbb{R}^{d}$ such that $\nabla \mu\left(s_{a}\right)=a$;
2. $\mathcal{E} \subset \bigcup_{a \in A}\left\{x \in \mathbb{R}^{d}: s_{a}^{\top}(x-a) \geq 0\right\}$;
3. $A \backslash\{a\}$ does not satisfy the first two conditions for any $a \in A$.

We call any point in A a dominating point. For two dominating points a and $a^{\prime}$, we say a is more significant than $a^{\prime}$ if $I(a)<I\left(a^{\prime}\right)$.

Suppose that $A=\left\{a_{1}, \ldots, a_{r}\right\}$ is a dominating set. Then the corresponding mixture IS distribution is given by $\frac{d \tilde{P}}{d P}(\omega)=\sum_{i=1}^{r} \alpha_{i} e^{s_{a_{i}}^{\top} X-\mu\left(s_{a_{i}}\right)}$ with $\sum_{i=1}^{r} \alpha_{i}=1, \alpha_{i}>0, \forall i$. Like the discussion in Section 3.2, it is routine to derive that such an IS estimator is asymptotically efficient.

Now we consider using a partial list of dominating points. In particular, we can still sequentially fill in the dominating set $A=\left\{a_{1}, \ldots, a_{r}\right\}$ where $a_{i}=\arg \min \left\{I(y): y \in \mathcal{E}, s_{a_{j}}^{\top}\left(y-a_{j}\right)<0, j=\right.$ $1, \ldots, i-1\}$ and hence $I\left(a_{1}\right) \leq \cdots \leq I\left(a_{r}\right)$, and suppose that we have a stopping strategy $k=k(\gamma)$ with $1 \leq k \leq r$ before locating all the dominating points. We choose the mixture IS distribution given by

$$
\begin{equation*}
\frac{d \tilde{P}}{d P}(\omega)=\sum_{i=1}^{k} \frac{1}{k} e^{s_{a_{i}}^{\top} X-\mu\left(s_{a_{i}}\right)} . \tag{3.15}
\end{equation*}
$$

Unlike in (3.11), we no longer assign an individual weight $\alpha_{i}$ to each dominating point $a_{i}$. Instead, we only consider the uniform mixture for simplicity, since now the number of dominating points $r$ and $k$ can both potentially change with $\gamma$.

First of all, we summarize the basic assumptions on the problem setting to ensure that there exists a dominating set with moderate size:

Assumption 3.5. Consider the problem of estimating $p=P(X \in \mathcal{E})$ with $\mathcal{E}=\left\{x \in \mathbb{R}^{d}: g(x) \geq\right.$ $\gamma\}$ where $X \in \mathbb{R}^{d}$ is a random vector and $g$ is a function. Assume that

1. $p \rightarrow 0$ as $\gamma \rightarrow \infty$;
2. Assumption 3.3 holds for the cumulant generating function of $X$ under $P$;
3. Assumption 3.4 holds for $\mathcal{E}$;
4. For any $\gamma$, there exists a dominating set for $\mathcal{E}$, denoted as $A=\left\{a_{1}, \ldots, a_{r}\right\}$ where $a_{i}=$ $\arg \min \left\{I(y): y \in \mathcal{E}, s_{a_{j}}^{\top}\left(y-a_{j}\right)<0, j=1, \ldots, i-1\right\}$ and hence $I\left(a_{1}\right) \leq \cdots \leq I\left(a_{r}\right)$. Besides, $r=r(\gamma)$ grows at most subexponentially in $-\log p$.

Now consider a stopping strategy $k=k(\gamma)$ and the corresponding IS distribution (3.15). Under Assumption 3.5, for $k=k(\gamma)$ with $1 \leq k \leq r$, denote $\mathcal{E}_{1}=\mathcal{E} \cap \bigcup_{i=1}^{k}\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$ and $\mathcal{E}_{2}=\mathcal{E} \backslash \mathcal{E}_{1}$. Corresponding to the settings in Theorem 3.3, we let $\mathcal{A}_{\gamma}^{j}=\left\{X \in \mathcal{E}_{j}\right\}, j=1,2$. We introduce Assumptions 3.6 and 3.7.

Assumption 3.6. If $k=r$, then let $a_{k+1}=\infty \mathbf{1}_{d}$. Assume that $p_{1}:=P\left(\mathcal{F}_{\gamma}^{1}\right)=P\left(X \in \mathcal{E}_{1}\right) \sim e^{-I\left(a_{1}\right)}$ and that $p_{2}:=P\left(\mathcal{A}_{\gamma}^{2}\right)=P\left(X \in \mathcal{E}_{2}\right)$ is upper bounded by $e^{-I\left(a_{k+1}\right)}$ up to subexponential factor in $-\log p$. Besides, $e^{I\left(a_{1}\right)-I\left(a_{k+1}\right)}$ exponentially decays in $-\log p$.

Assumption 3.7. Assume that $\tilde{p}_{2}:=\tilde{P}\left(\mathcal{A}_{\gamma}^{2}\right)=\tilde{P}\left(X \in \mathcal{E}_{2}\right)$ exponentially decays in $-\log p$.

Roughly, Assumption 3.6 implies that $p_{2}$ is exponentially smaller than $p_{1}$ and Assumption 3.7 implies that $\mathcal{E}_{2}$ is hardly hit even under $\tilde{P}$. Applying Theorem 3.3, we have that

Theorem 3.12 (Attaining probabilistic efficiency with a partial list of dominating points (new regime)). Under Assumptions 3.5, 3.6 and 3.7, the IS estimator $Z=I(X \in \mathcal{E}) \frac{d P}{d \tilde{P}}(\omega)$ under $\tilde{P}$ given by (3.15) is strongly probabilistically efficient.

Similar to Section 3.5, we can also construct asymptotically valid CIs with the sample mean and sample variance.

Theorem 3.13 (Constructing confidence intervals with probabilistically efficient estimators (new regime)). Assume that Assumptions 3.5, 3.6 and 3.7 hold. The $I S$ estimator is $Z=I(X \in$ $\mathcal{E}) \frac{d P}{d \tilde{P}}(\omega)$ under $\tilde{P}$ given by (3.15). We sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)}=I\left(X^{(i)} \in\right.$ $\mathcal{E}) \frac{d P}{d \tilde{P}}, Z_{1}^{(i)}=I\left(X^{(i)} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}, i=1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$ 's. In this case, If $n$ is subexponentially growing in $-\log p$ as $\gamma \rightarrow \infty$.

Then, for any $0<\alpha<1$,

$$
\liminf _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p| \leq \sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \geq 1-\alpha .
$$

That is,

$$
\hat{p} \pm\left(\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right)
$$

is an asymptotically valid $(1-\alpha)$-level CI for $p$.

Similar to Section 3.5, the CI in Theorem 3.13 is more conservative than the CLT-based interval $\hat{p} \pm z_{1-\alpha / 2} \sqrt{\frac{\hat{v}}{n}}$. If additionally the following assumption holds, then the CLT-based CI is also asymptotically valid:

Assumption 3.8. Denote $Z_{1}=I\left(X \in \mathcal{E}_{1}\right) \frac{f(X)}{\tilde{f}(X)}$ under $\tilde{P}$. Assume that $\widetilde{\operatorname{Var}}\left(Z_{1}\right) \sim e^{-2 I\left(a_{1}\right)}$.
Similar to Lemma 3.2, Assumption 3.8 serves to control the Berry-Esseen error bound. To be more concrete, with this additional assumption, we have:

Theorem 3.14 (Constructing tight confidence intervals (new regime)). Assume that Assumptions 3.5, 3.6, 3.7 and 3.8 hold. The IS estimator is $Z=I(X \in \mathcal{E}) \frac{d P}{d \tilde{P}}(\omega)$ under $\tilde{P}$ given by (3.15). We sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)}=I\left(X^{(i)} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}, Z_{1}^{(i)}=I\left(X^{(i)} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}, i=$ $1, \ldots, n$. Use $\hat{p}$ and $\hat{V}$ to respectively denote the sample mean and sample variance of $Z^{(i)}$ 's. In this case, we could choose $n$ subexponentially growing in $-\log p$ such that $\frac{k^{2} e^{-2 I\left(a_{1}\right)}}{n \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)} \rightarrow 0$ and $\frac{\tilde{E}^{2}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{n \widetilde{\operatorname{Var}^{3}}\left(Z_{1}^{(1)}\right)} \rightarrow 0$ as $\gamma \rightarrow \infty$. Then, for any $0<\alpha<1$,

$$
\liminf _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p| \leq z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \geq 1-\alpha .
$$

That is,

$$
\hat{p} \pm z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}
$$

is an asymptotically valid $(1-\alpha)$-level CI for $p$.
In the above, Assumptions 3.5, 3.6, 3.7 and 3.8 are technical conditions that need to be analyzed case by case. The next Section 3.10.2 focuses on a specific but important problem setting where we show how to verify all these conditions.

### 3.10.2 Guarantees for Gaussian Input Distribution

Suppose $X \sim N(\lambda, \Sigma)$ under $P$ where $\lambda \in \mathbb{R}^{d}$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. In this case $\mu(x)=\lambda^{\top} x+\frac{1}{2} x^{\top} \Sigma x, s_{a}=\Sigma^{-1}(a-\lambda)$ and $I(x)=\frac{1}{2}(x-\lambda)^{\top} \Sigma^{-1}(x-\lambda)$. Moreover, we suppose that $g$ is a piecewise linear function. Related theoretical guarantees for piecewise linear function of Gaussian input can be found in [125].

We use the following natural stopping strategy. First, fix a constant $C>1$. Then, we sequentially find the dominating points where we stop as long as $\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)>$ $C\left(a_{k}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k}-\lambda\right)$. If there is no such $k$, then we let $k=r$, the number of all dominating points. The IS distribution is chosen as $X \sim \frac{1}{k} \sum_{i=1}^{k} \phi\left(x ; a_{i}, \Sigma\right)$. We summarize these in Algorithm 3.2.

We have the following theorem suggesting that under the above setting, all the assumptions listed in Section 3.10.1 are satisfied.

Theorem 3.15 (Verification of assumptions for Gaussian inputs). Suppose that $X \sim N(\lambda, \Sigma)$ under $P$ where $\lambda \in \mathbb{R}^{d}$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. $\mathcal{E}=\left\{x \in \mathbb{R}^{d}: g(x) \geq \gamma\right\}$ where $g$ is a piecewise linear function. $C>1$ is a constant. Assume that for any $\gamma, P(X \in \mathcal{E})>0$ and $\lambda \notin \overline{\mathcal{E}}$. For any $\gamma$, we use the stopping strategy described above (more precisely Algorithm 3.2) to sequentially find dominating points $a_{1}, a_{2}, \ldots, a_{k}$ in decreasing significance and set up the $I S$ distribution $X \sim \frac{1}{k} \sum_{i=1}^{k} \phi\left(x ; a_{i}, \Sigma\right)$. Denote $\mathcal{E}_{1}=\mathcal{E} \cap \bigcup_{i=1}^{k}\left\{x \in \mathbb{R}^{d}:\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(x-a_{i}\right) \geq 0\right\}$ and $\mathcal{E}_{2}=\mathcal{E} \backslash \mathcal{E}_{1}$. Then Assumptions 3.5, 3.6, 3.7 and 3.8 hold .

Note that in Theorem 3.15, all assumptions including Gaussianity and piecewise linear $g$ are all straightforward to verify. With Theorem 3.15, we thus get the following corollaries from Theorems 3.12, 3.13 and 3.14:

```
Algorithm 3.2: Estimate \(P(g(X) \geq \gamma)\) with Gaussian input \(X\) and piecewise linear func-
tion \(g\).
```

    Input: Piecewise linear function \(g\), rarity parameter \(\gamma\), input distribution \(N(\lambda, \Sigma)\),
        threshold \(C>1\), sample size \(n\).
    Output: IS estimate \(\hat{p}\).
    Start with \(k=0\);
    While \(\left\{x: g(x) \geq \gamma,\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(x-a_{i}\right)<0, \forall i=1, \ldots, k\right\} \neq \emptyset\) do
        Find a dominating point \(a_{k+1}\) by solving the optimization problem
    $$
\begin{aligned}
a_{k+1}=\arg \min _{x} & (x-\lambda)^{\top} \Sigma^{-1}(x-\lambda) \\
\text { s.t. } & g(x) \geq \gamma \\
& \left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(x-a_{i}\right)<0, \forall i=1, \ldots, k .
\end{aligned}
$$

## If $k>0$ and $\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)>C\left(a_{k}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k}-\lambda\right)$ do

## Break

## Else do

Update $k \leftarrow k+1$;
8 End
9 Sample $X_{1}, \ldots, X_{n}$ from the mixture distribution $\frac{1}{k} \sum_{i=1}^{k} \phi\left(x ; a_{i}, \Sigma\right)$.
10 Compute the IS estimate $\hat{p}=\frac{1}{n} \sum_{i=1}^{n} I\left(g\left(X_{i}\right) \geq \gamma\right) L\left(X_{i}\right)$ where the likelihood ratio function is

$$
L(x)=\frac{e^{-\frac{1}{2}(x-\lambda)^{\top} \Sigma^{-1}(x-\lambda)}}{\frac{1}{k} \sum_{i=1}^{k} e^{-\frac{1}{2}\left(x-a_{i}\right)^{\top} \Sigma^{-1}\left(x-a_{i}\right)}} .
$$

Corollary 3.1 (Probabilistic efficiency for Gaussian inputs). Under the settings in Theorem 3.15, the $I S$ estimator $Z=I(X \in \mathcal{E}) \frac{d P}{d \tilde{P}}$ is strongly probabilistically efficient.

Corollary 3.2 (Confidence intervals for Gaussian inputs). Under the settings in Theorem 3.15, we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)}=I\left(X^{(i)} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}$. Use $\hat{p}$ and $\hat{V}$ to denote the sample mean and sample variance of $Z^{(i)}$ 's. We could choose $n$ subexponentially growing in $-\log p$ such that for any $0<\alpha<1$,

$$
\liminf _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p| \leq \sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-\frac{1}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)}}{3(n-1)}\right) \geq 1-\alpha .
$$

Corollary 3.3 (Tight confidence intervals for Gaussian inputs). Under the settings in Theorem
3.15, we sample $X^{(1)}, \ldots, X^{(n)}$ i.i.d. from $\tilde{P}$ and let $Z^{(i)}=I\left(X^{(i)} \in \mathcal{E}\right) \frac{d P}{d \tilde{P}}$. Use $\hat{p}$ and $\hat{V}$ to denote the sample mean and sample variance of $Z^{(i)}$ 's. We could choose $n$ subexponentially growing in $-\log p$ such that for any $0<\alpha<1$,

$$
\liminf _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p| \leq z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \geq 1-\alpha
$$

While the choice of $C$ (as long as it is $>1$ ) does not affect the guarantee on strong probabilistic efficiency or the asymptotic validity of the CIs, it does affect the accuracy of the estimates for a given, finite $\gamma$. In particular, it is often the case that in practice we only need to solve one problem for a fixed $\gamma$ instead of solving a series of problems with varying $\gamma$. In this scenario, as long as $\left(a_{2}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{2}-\lambda\right)>\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$, we can pick a $C>1$ such that we would stop our dominating point search at $k=1$, i.e., use only the first point. Note that this does not imply that for this choice of $C$ we have $k=1$ for any $\gamma$. Thus, the finiteness of the rarity parameter comes into play in a subtle way.

### 3.11 Supplementary D: Proofs

Proof of Theorem 3.2. First, we know that $I\left(a_{i}\right)=s_{a_{i}}^{\top} a_{i}-\mu\left(s_{a_{i}}\right)$ for $i=1, \ldots, r$. For any $x \in \mathbb{R}^{d}$ such that $s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0$, we have $I(x) \geq s_{a_{i}}^{\top} x-\mu\left(s_{a_{i}}\right) \geq s_{a_{i}}^{\top} a_{i}-\mu\left(s_{a_{i}}\right)=I\left(a_{i}\right)$. By Definition 3.2, we have $\mathcal{E} \subset \bigcup_{i=1}^{r}\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$. Thus, for any $x \in \mathcal{E}$, there exists $i$ such that $s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0$, and hence $I(x) \geq I\left(a_{i}\right)$. Therefore, $I(\mathcal{E})=\inf _{x \in \mathcal{E}} I(x) \geq \min _{i=1, \ldots, r} I\left(a_{i}\right)$. On the other hand, $a_{1}, \ldots, a_{r} \in \partial \mathcal{E} \cap \mathcal{D}(I)^{\circ}$ and $I$ is differentiable in $\mathcal{D}(I)^{\circ}$, so $I(\mathcal{E}) \leq \min _{i=1, \ldots, r} I\left(a_{i}\right)$. Overall we have $I(\mathcal{E})=\min _{i=1, \ldots, r} I\left(a_{i}\right)$.

Proof of Proposition 3.1. We could split the rare-event set as $\mathcal{E}=\bigcup_{i=1}^{r} \mathcal{E}_{i}$ where $\mathcal{E}_{i}$ 's are disjoint and $a_{i} \in \overline{\mathcal{E}_{i}} \subset\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$. Note that the hyperplane $\left\{x \in \mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right)=0\right\}$ is tangent to the rate function level set $\left\{x \in \mathbb{R}^{d}: I(x)=I\left(a_{i}\right)\right\}$ at $a_{i}$, so $s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0$ implies that $I(x) \geq I\left(a_{i}\right)$. Now, the likelihood ratio is given by

$$
L\left(X_{\gamma}\right):=\frac{d P}{d \tilde{P}}\left(X_{\gamma}\right)=\frac{1}{\sum_{i=1}^{r} \alpha_{i} e^{s_{a_{i}}^{\tau} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)}}
$$

and it satisfies that for any $i$,

$$
L\left(X_{\gamma}\right) \leq \frac{1}{\alpha_{i} e^{s_{a_{i}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{i}}\right)}}=\frac{1}{\alpha_{i}} e^{-s_{a_{i}}^{\top}\left(X_{\gamma}-\gamma a_{i}\right)-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)} .
$$

Hence we have that

$$
\begin{equation*}
I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{i}\right) L\left(X_{\gamma}\right) \leq \frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)} \tag{3.16}
\end{equation*}
$$

and

$$
I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L\left(X_{\gamma}\right) \leq \max _{i=1, \ldots, r}\left\{\frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right\}
$$

Thus, using Theorem 3.1, we have that

$$
\begin{aligned}
\liminf _{\gamma \rightarrow \infty} \frac{\log \tilde{E}\left(I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L^{2}\left(X_{\gamma}\right)\right)}{\log \tilde{E}\left(I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L\left(X_{\gamma}\right)\right)} & \geq \liminf _{\gamma \rightarrow \infty} \frac{2 \log \left(\max _{i=1, \ldots, r}\left\{\frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right\}\right)}{\log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)} \\
& =\liminf _{\gamma \rightarrow \infty} \frac{2 \max _{i=1, \ldots, r}\left\{\frac{1}{\gamma} \log \left(\frac{1}{\alpha_{i}} e^{-\gamma\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)}\right)\right\}}{\frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)} \\
& =\frac{-2 \min _{i=1, \ldots, r} I\left(a_{i}\right)}{-I(\mathcal{E})}=2,
\end{aligned}
$$

which verifies that the IS estimator is asymptotically efficient.

Proof of Proposition 3.2. We derive the growth rate of the relative error for a more general case. Suppose that we estimate $p=P\left(\frac{1}{\gamma} X_{\gamma} \in(-\infty,-k] \cup[1, \infty)\right)$ where $k>1, X_{\gamma} \sim N(0, \gamma)$ under $P$ and the IS distribution is chosen as $X_{\gamma} \sim N(\gamma, \gamma)$ under $\tilde{P}$. This is the setup of Examples 3.1 and 3.2. There are two dominating points, 1 and $-k$, and 1 is the more significant one. Under the IS distribution, the likelihood ratio is $L=e^{-X_{\gamma}+\gamma / 2}$ and the IS estimator is $Z=I\left(\frac{1}{\gamma} X_{\gamma} \in\right.$
$(-\infty,-k] \cup[1, \infty)) e^{-X_{\gamma}+\gamma / 2}$. Then we have that

$$
\begin{aligned}
\tilde{E}\left(Z^{2}\right) & =\tilde{E}\left(I\left(\frac{1}{\gamma} X_{\gamma} \in(-\infty,-k] \cup[1, \infty)\right) e^{-2 X_{\gamma}+\gamma}\right) \\
& =\int_{x \leq-k \gamma \text { or } x \geq \gamma} e^{-2 x+\gamma} \frac{1}{\sqrt{2 \pi \gamma}} e^{-\frac{(x-\gamma)^{2}}{2 \gamma}} d x \\
& =\int_{x \leq-k \gamma \text { or } x \geq \gamma} e^{\gamma} \frac{1}{\sqrt{2 \pi \gamma}} e^{-\frac{(x+\gamma)^{2}}{2 \gamma}} d x \\
& =\int_{y \leq(1-k) \sqrt{\gamma} \text { or } y \geq 2 \sqrt{\gamma}} e^{\gamma} \frac{1}{\sqrt{2 \pi}} e^{-\frac{y^{2}}{2}} d y \quad(y=(x+\gamma) / \sqrt{\gamma}) \\
& =e^{\gamma}(\bar{\Phi}((k-1) \sqrt{\gamma})+\bar{\Phi}(2 \sqrt{\gamma}))
\end{aligned}
$$

where $\bar{\Phi}$ denotes the tail distribution function of standard normal distribution. It is known that $\bar{\Phi}(x)=\Theta\left(\frac{1}{x} e^{-x^{2} / 2}\right)$ as $x \rightarrow \infty$, and hence $\tilde{E}\left(Z^{2}\right)=\Theta\left(\frac{1}{\sqrt{\gamma}} e^{\left(1-\frac{(k-1)^{2}}{2}\right) \gamma}\right)$ if $1<k<3$ and $\tilde{E}\left(Z^{2}\right)=$ $\Theta\left(\frac{1}{\sqrt{\gamma}} e^{-\gamma}\right)$ if $k \geq 3$. Besides, $p=\bar{\Phi}(\sqrt{\gamma})+\bar{\Phi}(k \sqrt{\gamma})=\Theta\left(\frac{1}{\sqrt{\gamma}} e^{-\gamma / 2}\right)$. Therefore, $\tilde{E}\left(Z^{2}\right) / p^{2}=$ $\Theta\left(\sqrt{\gamma} e^{\left(2-\frac{(k-1)^{2}}{2}\right) \gamma}\right)$ which grows exponentially in $\gamma$ if $1<k<3$ and $\tilde{E}\left(Z^{2}\right) / p^{2}=\Theta(\sqrt{\gamma})$ which grows polynomially in $\gamma$ if $k \geq 3$.

Proofs of Propositions 3.4 and 3.5. We first show a general result on IS estimator that exponentially tilts to the most significant dominating point. Consider a rare-event set $\mathcal{E}$ with $a_{1}$ as the most significant dominating point. We estimate the target rare event $\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right\}$ using the IS estimator with likelihood ratio

$$
L=\frac{1}{e^{s_{a_{1}}^{\top} X_{\gamma}-\gamma \mu_{\gamma}\left(s_{a_{1}}\right)}} .
$$

That is, we use exponential tilting with respect to $a_{1}$ only. Specifically, we consider $X_{\gamma} \sim N(\gamma \lambda, \gamma \Sigma)$.
In this case we have $\mu(x)=\mu_{\gamma}(x)=x^{\top} \lambda+\frac{1}{2} x^{\top} \Sigma x, s_{a}=\Sigma^{-1}(a-\lambda), I(y)=\frac{1}{2}(y-\lambda)^{\top} \Sigma^{-1}(y-\lambda)$.

The second moment of this IS estimator is

$$
\begin{aligned}
& \tilde{E}\left(I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L^{2}(\omega)\right) \\
= & E\left(I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L(\omega)\right) \\
= & \int_{\frac{x}{\gamma} \in \mathcal{E}} e^{-s_{a_{1}}^{\top} x+\gamma \mu_{\gamma}\left(s_{a_{1}}\right)}(2 \pi)^{-\frac{d}{2}}|\gamma \Sigma|^{-1 / 2} e^{-\frac{1}{2 \gamma}(x-\gamma \lambda)^{\top} \Sigma^{-1}(x-\gamma \lambda)} d x \\
= & \int_{\frac{x}{\gamma} \in \mathcal{E}} e^{-\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1} x+\gamma\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1} \lambda+\frac{\gamma}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)}(2 \pi)^{-\frac{d}{2}}|\gamma \Sigma|^{-1 / 2} e^{-\frac{1}{2 \gamma}(x-\gamma \lambda)^{\top} \Sigma^{-1}(x-\gamma \lambda)} d x \\
= & \int_{\frac{x}{\gamma} \in \mathcal{E}}(2 \pi)^{-\frac{d}{2}}|\gamma \Sigma|^{-1 / 2} e^{\gamma\left[-\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(\frac{x}{\gamma}-\lambda\right)+\frac{1}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)-\frac{1}{2}\left(\frac{x}{\gamma}-\lambda\right)^{\top} \Sigma^{-1}\left(\frac{x}{\gamma}-\lambda\right)\right]} d x \\
= & e^{\gamma\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)} \int_{\frac{x}{\gamma} \in \mathcal{E}}(2 \pi)^{-\frac{d}{2}}|\gamma \Sigma|^{-1 / 2} e^{-\frac{\gamma}{2}\left(\frac{x}{\gamma}+a_{1}-2 \lambda\right)^{\top} \Sigma^{-1}\left(\frac{x}{\gamma}+a_{1}-2 \lambda\right)} d x \\
= & e^{\gamma\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)} \bar{P}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)
\end{aligned}
$$

where $\bar{P}$ is the probability measure given by the exponential tilting with respect to $2 \lambda-a_{1}$ and hence $X_{\gamma} \sim N\left(\gamma\left(2 \lambda-a_{1}\right), \gamma \Sigma\right)$ under $\bar{P}$. We correspondingly denote $\bar{I}$ as the rate function under $\bar{P}$. Now, since $X_{\gamma}$ is Gaussian Assumption 3.1 holds. Suppose also that Assumption 3.2 holds. Then by Theorem 3.1 we know that

$$
\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \bar{P}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)=-\bar{I}(\mathcal{E})=-\frac{1}{2} \min _{y \in \mathcal{E}}\left(y+a_{1}-2 \lambda\right)^{\top} \Sigma^{-1}\left(y+a_{1}-2 \lambda\right)
$$

We also know that $\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)=-I\left(a_{1}\right)=-\frac{1}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$. Therefore,

$$
\begin{aligned}
& \lim _{\gamma \rightarrow \infty} \frac{\log \tilde{E}\left(I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L^{2}(\omega)\right)}{\log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)} \\
= & \lim _{\gamma \rightarrow \infty} \frac{\frac{1}{\gamma} \log \tilde{E}\left(I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right) L^{2}(\omega)\right)}{\frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)} \\
= & \lim _{\gamma \rightarrow \infty} \frac{\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)+\frac{1}{\gamma} \log \bar{P}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)}{\frac{1}{\gamma} \log P\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}\right)} \\
= & \frac{\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)-\frac{1}{2} \min _{y \in \mathcal{E}}\left(y+a_{1}-2 \lambda\right)^{\top} \Sigma^{-1}\left(y+a_{1}-2 \lambda\right)}{-\frac{1}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)} .
\end{aligned}
$$

By definition, the IS estimator is not asymptotically efficient if and only if $\min _{y \in \mathcal{E}}\left(y+a_{1}-\right.$ $2 \lambda)^{\top} \Sigma^{-1}\left(y+a_{1}-2 \lambda\right)<4\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$.

In order to check the existence of $\tilde{y}$ such that $\left(\tilde{y}+a_{1}-2 \lambda\right)^{\top} \Sigma^{-1}\left(\tilde{y}+a_{1}-2 \lambda\right)<4\left(a_{1}-\right.$ $\lambda)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$, we can formulate the following optimization

$$
\begin{equation*}
\min _{y \in \mathcal{E}}\left(y+a_{1}-2 \lambda\right)^{\top} \Sigma^{-1}\left(y+a_{1}-2 \lambda\right)-4\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right) \tag{3.17}
\end{equation*}
$$

and check whether the objective is negative for the optimal (or any feasible) solution. Since the objective function is quadratic on the decision vector $y$, the tractability of the optimization problem (3.17) is determined by the feasible region $\mathcal{E}$. In the example in Section 3.3.2, we have $\bigcup_{m=1}^{d} \mathcal{H}_{m}=\bigcup_{m=1}^{d}\left\{x \in \mathbb{R}^{d}: \sum_{i=1}^{m} x_{i} \geq a\right\}$. Since each $\mathcal{H}_{m}$ is a half-space and hence convex, we can independently solve

$$
\min _{y \in \mathcal{H}_{m}}\left(y+a_{1}-2 \lambda\right)^{\top} \Sigma^{-1}\left(y+a_{1}-2 \lambda\right)-4\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)
$$

for $m=1, \ldots, d$. Since the feasible region of the above optimization associated with any $\mathcal{H}_{m}$ is a subset of the original problem (3.17), any solution with negative objective indicates that the IS
estimator with dominating points $a_{1}$ is not asymptotically efficient. In the example in Section 3.3.3, we formulate the problem (3.17) as a mixed integer programming problem following the treatment of the rare-event set formed by machine learning predictors in [125]. Using this approach, we can show that the IS estimators with the most dominating points are not asymptotically efficient for both examples considered in Sections 3.3.2 and 3.3.3.

To prove Theorem 3.7, we need the following lemma:
Lemma 3.1 (Theorem 4 in [133]). Let $Y, Y_{1}, \ldots, Y_{n}$ be i.i.d. random variables with values in $[0,1]$ and let $\delta>0$. Then with probability at least $1-\delta$

$$
E Y-\frac{1}{n} \sum_{i=1}^{n} Y_{i} \leq \sqrt{\frac{2 V_{n}(\mathbf{Y}) \log (2 / \delta)}{n}}+\frac{7 \log (2 / \delta)}{3(n-1)}
$$

where $\mathbf{Y}=\left(Y_{1}, \ldots, Y_{n}\right)$ and $V_{n}(\mathbf{Y})$ is the sample variance of $Y_{i}$ 's.

Now we prove Theorem 3.7:
Proof of Theorem 3.7. Following the notation in Section 3.4, we split $\mathcal{E}$ into $\mathcal{E}_{1}=\mathcal{E} \cap \bigcup_{i=1}^{k}\{x \in$ $\left.\mathbb{R}^{d}: s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$ and $\mathcal{E}_{2}=\mathcal{E} \backslash \mathcal{E}_{1}$, and then we correspondingly define $\mathcal{A}_{\gamma}^{1}=\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}\right\}$ and $\mathcal{A}_{\gamma}^{2}=\left\{\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{2}\right\}$. We denote $Z_{1}^{(i)}=I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}} . \hat{p}_{1}$ and $\hat{V}_{1}$ respectively denote the sample mean and sample variance of $Z_{1}^{(i)}$,s. We also define $N=\sum_{i=1}^{n} I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}_{2}\right)$. Note that conditional
on $N=0$, we have $Z^{(i)}=Z_{1}^{(i)}$. Then

$$
\begin{aligned}
& \tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}\right) \\
= & \tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}, N=0\right) \\
& +\tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}, N>0\right) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}, N=0\right)+\tilde{P}(N>0) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}\right)+\tilde{P}(N>0) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|+p_{2}>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}\right)+n \tilde{p}_{2} .
\end{aligned}
$$

Similar to the derivation of (3.16), we know that $0 \leq Z_{1}^{(i)} \leq M_{\gamma}, \forall i$. By applying Lemma 3.1 with $Y_{i}=Z_{1}^{(i)} / M_{\gamma}$ and $Y_{i}=1-Z_{1}^{(i)} / M_{\gamma}$ respectively, we get that

$$
\tilde{P}\left(p_{1}>\hat{p}_{1}+\sqrt{\frac{2 \hat{V}_{1} \log (4 / \delta)}{n}}+\frac{7 \log (4 / \delta) M_{\gamma}}{3(n-1)}\right) \leq \delta / 2
$$

and

$$
\tilde{P}\left(p_{1}<\hat{p}_{1}-\sqrt{\frac{2 \hat{V}_{1} \log (4 / \delta)}{n}}-\frac{7 \log (4 / \delta) M_{\gamma}}{3(n-1)}\right) \leq \delta / 2
$$

for any $\delta>0$. Thus,

$$
\begin{equation*}
\tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \delta)}{n}}+\frac{7 \log (4 / \delta) M_{\gamma}}{3(n-1)}\right) \leq \delta \tag{3.18}
\end{equation*}
$$

Find $\alpha^{\prime}=\alpha^{\prime}(\gamma)$ such that

$$
\frac{7 \log \left(4 / \alpha^{\prime}\right) M_{\gamma}}{3(n-1)}=\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}-p_{2}
$$

That is,

$$
\alpha^{\prime}=\alpha \exp \left(\frac{3(n-1) p_{2}}{7 M_{\gamma}}\right)
$$

Clearly $\alpha^{\prime}>\alpha$ and $\log \left(4 / \alpha^{\prime}\right)<\log (4 / \alpha)$. From the proof of Theorem 3.6, we have that either (i) $k=r$ and $p_{2}=0$ or (ii) $k<r$ and $\lim \sup _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log p_{2} \leq-I\left(a_{k+1}\right)<-I\left(a_{1}\right)$. Moreover, we have

$$
\frac{1}{\gamma} \log M_{\gamma}=\max _{i=1, \ldots, k}\left\{-\frac{1}{\gamma} \log \alpha_{i}-\left(s_{a_{i}}^{\top} a_{i}-\mu_{\gamma}\left(s_{a_{i}}\right)\right)\right\} \rightarrow-I\left(a_{1}\right) .
$$

Thus, for subexponentially growing $n$, we have $\frac{3(n-1) p_{2}}{7 M_{\gamma}} \rightarrow 0$ and hence $\alpha^{\prime} \rightarrow \alpha$ as $\gamma \rightarrow \infty$. We replace $\delta$ with $\alpha^{\prime}$ in (3.18), and then we get

$$
\tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>\sqrt{\frac{2 \hat{V}_{1} \log \left(4 / \alpha^{\prime}\right)}{n}}+\frac{7 \log \left(4 / \alpha^{\prime}\right) M_{\gamma}}{3(n-1)}\right) \leq \alpha^{\prime}
$$

Hence

$$
\tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}-p_{2}\right) \leq \alpha^{\prime}
$$

Therefore,

$$
\tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) M_{\gamma}}{3(n-1)}\right) \leq \alpha^{\prime}+n \tilde{p}_{2} \rightarrow \alpha \text { as } \gamma \rightarrow \infty .
$$

To prove Theorem 3.8, we first prove a lemma regarding the variance $\widetilde{\operatorname{Var}}\left(Z_{1}\right)$. In particular, we have assumed that $\widetilde{\operatorname{Var}}\left(Z_{1}\right)$ cannot be "too large", while this lemma implies that it cannot be "too small" as well. As we will see, this lemma is used to control the error of the normal approximation in analog to the Berry-Esseen theorem in order to argue the validity of $I_{2}$.

Lemma 3.2. Under the same setting as Theorem 3.6, split $\mathcal{E}$ into $\mathcal{E}_{1}=\mathcal{E} \cap \bigcup_{i=1}^{k}\left\{x \in \mathbb{R}^{d}\right.$ : $\left.s_{a_{i}}^{\top}\left(x-a_{i}\right) \geq 0\right\}$ and $\mathcal{E}_{2}=\mathcal{E} \backslash \mathcal{E}_{1}$. Denote $Z_{1}=I\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}(\omega)$ with $X_{\gamma} \sim \tilde{P}$. Then $\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \widetilde{\operatorname{Var}}\left(Z_{1}\right)=-2 I(\mathcal{E})=-2 I\left(a_{1}\right)$.

Proof of Lemma 3.2. From Proposition 3.1, we know that $\lim _{\sup }^{\gamma \rightarrow \infty}{ }_{\gamma} \frac{1}{\gamma} \log \widetilde{\operatorname{Var}}\left(Z_{1}\right) \leq-2 I(\mathcal{E})$. Thus we only need to show that $\lim _{\inf }^{\gamma \rightarrow \infty} \boldsymbol{~} \frac{1}{\gamma} \log \widetilde{\operatorname{Var}}\left(Z_{1}\right) \geq-2 I(\mathcal{E})$. Indeed, we have that $\widetilde{\operatorname{Var}}\left(Z_{1}\right)=\tilde{E}\left(\left(Z_{1}-p_{1}\right)^{2}\right) \geq \tilde{E}\left(\left(Z_{1}-p_{1}\right)^{2} I\left(\frac{1}{\gamma} X_{\gamma} \notin \mathcal{E}_{1}\right)\right)=p_{1}^{2} \tilde{P}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}^{c}\right) \geq p_{1}^{2} \alpha_{1} \tilde{P}_{1}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}^{c}\right)$ where $\tilde{P}_{1}$ is as defined in the proof of Theorem 3.6, and hence

$$
\liminf _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \widetilde{\operatorname{Var}}\left(Z_{1}\right) \geq \liminf _{\gamma \rightarrow \infty}\left(\frac{2}{\gamma} \log p_{1}+\frac{1}{\gamma} \log \alpha_{1}+\frac{1}{\gamma} \log \tilde{P}_{1}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}^{c}\right)\right) .
$$

First, we know that $\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log p_{1}=-I(\mathcal{E})$. Second, we know that $\frac{1}{\gamma} \log \alpha_{1} \rightarrow 0$. Third, by Theorem 3.1, we know that $\liminf _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \tilde{P}_{1}\left(\frac{1}{\gamma} X_{\gamma} \in \mathcal{E}_{1}^{c}\right) \geq-\tilde{I}_{1}\left(\left(\mathcal{E}_{1}^{c}\right)^{\circ}\right)$ where $\tilde{I}_{1}(y)=$ $I(y)-s_{a_{1}}^{\top} y+\mu\left(s_{a_{1}}\right)$ as in the proof of Theorem 3.6. As $a_{1}$ is a dominating point, we have required that $a_{1} \in \mathcal{D}(I)^{\circ}$, and thus $a_{1} \in \mathcal{D}\left(\tilde{I}_{1}\right)^{\circ}$. We note that $\left(\mathcal{E}_{1}^{c}\right)^{\circ} \supset \bigcap_{i=1}^{k}\left\{x: s_{a_{i}}^{\top}\left(x-a_{i}\right)<0\right\}$. Then $\tilde{I}_{1}\left(\left(\mathcal{E}_{1}^{c}\right)^{\circ}\right) \leq \tilde{I}_{1}\left(\bigcap_{i=1}^{k}\left\{x: s_{a_{i}}^{\top}\left(x-a_{i}\right)<0\right\}\right)$. We note that since $a_{1}$ is the most significant dominating point, we have $s_{a_{i}}^{\top}\left(a_{1}-a_{i}\right) \leq 0$ (otherwise we get $I\left(a_{i}\right)<I\left(a_{1}\right)$, which is a contradiction). Thus $a_{1}$ is on the boundary of $\bigcap_{i=1}^{k}\left\{x: s_{a_{i}}^{\top}\left(x-a_{i}\right)<0\right\}$, and there exists a sequence of points $\left\{y_{n}\right\}_{n=1}^{\infty} \subset$ $\mathcal{D}\left(\tilde{I}_{1}\right)^{\circ} \cap \bigcap_{i=1}^{k}\left\{x: s_{a_{i}}^{\top}\left(x-a_{i}\right)<0\right\}$ such that $y_{n} \rightarrow a_{1}$ as $n \rightarrow \infty$. Therefore, we get $\tilde{I}_{1}\left(\bigcap_{i=1}^{k}\{x:\right.$ $\left.\left.s_{a_{i}}^{\top}\left(x-a_{i}\right)<0\right\}\right) \leq \tilde{I}_{1}\left(a_{1}\right)=0$. Overall, we have proved that $\liminf _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \widetilde{\operatorname{Var}}\left(Z_{1}\right) \geq-2 I(\mathcal{E})$.

We also need a concentration result for the sample variance:
Lemma 3.3 (Theorem 10 in [133]). Let $n \geq 2$ and $\mathbf{Y}=\left(Y_{1}, \ldots, Y_{n}\right)$ be a vector of independent random variables with values in $[0,1]$. Then for $\delta>0$ we have

$$
\begin{aligned}
& P\left(\sqrt{E V_{n}(\mathbf{Y})}>\sqrt{V_{n}(\mathbf{Y})}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}}\right) \leq \delta \\
& P\left(\sqrt{V_{n}(\mathbf{Y})}>\sqrt{E V_{n}(\mathbf{Y})}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}}\right) \leq \delta
\end{aligned}
$$

## With Lemmas 3.2 and 3.3, now we prove Theorem 3.8:

Proof of Theorem 3.8. We denote $Z_{1}^{(i)}=I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}$. $\hat{p}_{1}$ and $\hat{V}_{1}$ respectively denote the sample mean and sample variance of $Z_{1}^{(i)}$,s. We also define $N=\sum_{i=1}^{n} I\left(\frac{1}{\gamma} X^{(i)} \in \mathcal{E}_{2}\right)$. Similar to the proof of Theorem 3.7, we have

$$
\begin{aligned}
& \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right. \\
= & \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}, N=0\right)+\tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}, N>0\right) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p\right|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}_{1}}{n}}\right)+\tilde{P}(N>0) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}_{1}}{n}}-p_{2}\right)+n \tilde{p}_{2} \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}_{1}}{n}}-\sqrt{\frac{2 \log (1 / \delta)}{n(n-1)}} M_{\gamma}\right)-p_{2}\right)+ \\
& \tilde{P}\left(\sqrt{\tilde{E} \hat{V}_{1}}>\sqrt{\hat{V}_{1}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} M_{\gamma}\right)+n \tilde{p}_{2}
\end{aligned}
$$

for any $\delta=\delta(\gamma)>0$, where $M_{\gamma}$ is as defined in Theorem 3.7. We know that $0 \leq Z_{1}^{(i)} \leq M_{\gamma}, \forall i$. By Lemma 3.3 with $Y_{i}=Z_{1}^{(i)} / M_{\gamma}$, we get that

$$
\tilde{P}\left(\sqrt{\tilde{E} \hat{V}_{1}}>\sqrt{\hat{V}_{1}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} M_{\gamma}\right) \leq \delta
$$

By Berry-Esseen theorem, we know that for any $x \in \mathbb{R}$

$$
\left|\tilde{P}\left(\frac{\sqrt{n}\left(\hat{p}_{1}-p_{1}\right)}{\sqrt{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} \leq x\right)-\Phi(x)\right| \leq \frac{C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}}
$$

where $\Phi$ is the CDF of standard normal distribution and $C$ is a universal constant. Let

$$
x=z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}_{1}}{\widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}-\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} M_{\gamma}\right)-p_{2} \sqrt{\frac{n}{\widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} .
$$

Then we get that

$$
\tilde{P}\left(\frac{\sqrt{n}\left|\hat{p}_{1}-p_{1}\right|}{\sqrt{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}>x\right) \leq 2 \Phi(-x)+\frac{2 C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}}
$$

Hence,

$$
\tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq 2 \Phi(-x)+\frac{2 C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}}+\delta+n \tilde{p}_{2} .
$$

First, we have that $\tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3} \leq \max \left(p_{1}^{3},\left(M_{\gamma}-p_{1}\right)^{3}\right)$. From the proof of Theorems 3.6 and 3.7, we know that $-\frac{1}{\gamma} \log p_{1} \rightarrow-I\left(a_{1}\right)$ and $-\frac{1}{\gamma} \log M_{\gamma} \rightarrow-I\left(a_{1}\right)$. Thus $\left.\lim \sup _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \tilde{E} \right\rvert\, Z_{1}^{(1)}-$
 Hence under the assumptions we could choose $n$ as required in the theorem. In particular, we have that

$$
\frac{2 C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}} \rightarrow 0
$$

Now we analyze $x$. We know that $\sqrt{\frac{E \hat{V}_{1}}{\widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}=1$ and

$$
p_{2} \sqrt{\frac{n}{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}=\frac{p_{2}}{e^{-\gamma I\left(a_{1}\right)}} \sqrt{\frac{n e^{-2 \gamma I\left(a_{1}\right)}}{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} \rightarrow 0
$$

since $\frac{p_{2}}{e^{-\gamma I\left(a_{1}\right)}}$ decays exponentially (proof of Theorem 3.6) while $n$ and $\frac{e^{-2 \gamma I\left(a_{1}\right)}}{\overline{\operatorname{Var}}\left(Z_{1}^{11}\right)}$ grow subexponentially (Lemma 3.2) in $\gamma$. Now we consider

$$
\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} M_{\gamma}=\sqrt{\frac{2 \log (1 / \delta) n}{n-1}} \sqrt{\frac{M_{\gamma}^{2}}{n \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} .
$$

Since we assume that $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)} \rightarrow 0$, we could set $\delta$ such that $\delta \rightarrow 0$ and

$$
\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} M_{\gamma} \rightarrow 0 .
$$

In this case, $x \rightarrow z_{1-\alpha / 2}$ and hence $\Phi(-x) \rightarrow \alpha / 2$. Combining all the results, we get that

$$
\limsup _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq \alpha
$$

Proof of Theorem 3.9. First, following the proof of Lemma 3.2, it is easy to get that

$$
\lim _{\gamma \rightarrow \infty} \frac{1}{\gamma} \log \widetilde{\operatorname{Var}}(Z)=-2 I(\mathcal{E})=-2 I\left(a_{1}\right)
$$

Moreover, we have that

$$
\begin{aligned}
\tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq & \leq \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}}{n}}-\sqrt{\frac{2 \log (1 / \delta)}{n(n-1)}} M_{\gamma}\right)\right) \\
& +\tilde{P}\left(\sqrt{\tilde{E} \hat{V}}>\sqrt{\hat{V}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} M_{\gamma}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \geq & \geq \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}}{n}}+\sqrt{\frac{2 \log (1 / \delta)}{n(n-1)}} M_{\gamma}\right)\right) \\
& -\tilde{P}\left(\sqrt{\hat{V}}>\sqrt{\tilde{E} \hat{V}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} M_{\gamma}\right)
\end{aligned}
$$

for any $\delta=\delta(\gamma)>0$. We know that $0 \leq Z^{(i)} \leq M_{\gamma}, \forall i$. By Lemma 3.3, we get that

$$
\tilde{P}\left(\sqrt{\tilde{E} \hat{V}}>\sqrt{\hat{V}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} M_{\gamma}\right) \leq \delta
$$

and

$$
\tilde{P}\left(\sqrt{\hat{V}}>\sqrt{\tilde{E} \hat{V}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} M_{\gamma}\right) \leq \delta
$$

By Berry-Esseen theorem, we know that

$$
\tilde{P}\left(\frac{\sqrt{n}|\hat{p}-p|}{\sqrt{\widetilde{\operatorname{Var}}\left(Z^{(1)}\right)}}>x_{1}\right) \leq 2 \Phi\left(-x_{1}\right)+\frac{2 C \tilde{E}\left|Z^{(1)}-p\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z^{(1)}\right) \sqrt{n}}
$$

and

$$
\tilde{P}\left(\frac{\sqrt{n}|\hat{p}-p|}{\sqrt{\widetilde{\operatorname{Var}}\left(Z^{(1)}\right)}}>x_{2}\right) \geq 2 \Phi\left(-x_{2}\right)-\frac{2 C \tilde{E}\left|Z^{(1)}-p\right|^{3}}{\widetilde{\operatorname{Var}^{3 / 2}\left(Z^{(1)}\right) \sqrt{n}}}
$$

where $\Phi$ is the CDF of standard normal distribution, $C$ is a universal constant, and

$$
\begin{aligned}
& x_{1}=z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}}{\widehat{\operatorname{Var}}\left(Z^{(1)}\right)}}-\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widehat{\operatorname{Var}}\left(Z^{(1)}\right)}} M_{\gamma}\right), \\
& x_{2}=z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}}{\widehat{\operatorname{Var}}\left(Z^{(1)}\right)}}+\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widetilde{\operatorname{Var}}\left(Z^{(1)}\right)}} M_{\gamma}\right) .
\end{aligned}
$$

Combining the above derivations, we get that

$$
\begin{aligned}
& \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq 2 \Phi\left(-x_{1}\right)+\frac{2 C \tilde{E}\left|Z^{(1)}-p\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z^{(1)}\right) \sqrt{n}}+\delta, \\
& \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \geq 2 \Phi\left(-x_{2}\right)-\frac{2 C \tilde{E}\left|Z^{(1)}-p\right|^{3}}{{\widetilde{\operatorname{Var}^{3 / 2}}\left(Z^{(1)}\right) \sqrt{n}}^{n}}-\delta .
\end{aligned}
$$

First, we have that $\tilde{E}\left|Z^{(1)}-p\right|^{3} \leq \max \left(p^{3},\left(M_{\gamma}-p\right)^{3}\right)$. From the proof of the previous theorems, we know that $-\frac{1}{\gamma} \log p \rightarrow-I\left(a_{1}\right)$ and $-\frac{1}{\gamma} \log M_{\gamma} \rightarrow-I\left(a_{1}\right)$. Thus, $\frac{\tilde{E}\left|Z^{(1)}-p\right|^{3}}{\widehat{\operatorname{Var}^{3 / 2}\left(Z^{(1)}\right)} \text { grows at most }}$
subexponentially in $\gamma$, and hence we could choose $n$ as required in the theorem. In particular, we have $\frac{2 C \tilde{E}\left|Z^{(1)}-p\right|^{3}}{\widehat{\operatorname{Var}}^{3 / 2}\left(Z^{(1)}\right) \sqrt{n}} \rightarrow 0$. Next, since we assume that $\frac{M_{\gamma}^{2}}{n \widehat{\operatorname{Var}}\left(Z^{(1)}\right)} \rightarrow 0$, we could set $\delta$ such that $\delta \rightarrow 0$ and

$$
\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widetilde{\operatorname{Var}}\left(Z^{(1)}\right)}} M_{\gamma} \rightarrow 0 .
$$

Hence, $x_{1}, x_{2} \rightarrow z_{1-\alpha / 2}$ as $\gamma \rightarrow \infty$. Overall, we get that

$$
\begin{aligned}
& \limsup _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq \alpha \\
& \liminf _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \geq \alpha
\end{aligned}
$$

Therefore, the theorem is proved.

Proof of Theorem 3.10. The conclusion follows from Theorem 3.5, by checking, for each setting in Sections 3.3.1, 3.3.2 and 3.3.3, Assumptions 3.1 and 3.2 hold and the most significant dominating point is unique. Since we have $\mu(x)=1.5 x+0.5 x^{2}-\log (1+x)$ for the case in Section 3.3.1 and $\mu(x)=\lambda^{\top} x+\frac{1}{2} x^{\top} \Sigma x$ for the cases in Sections 3.3.2 and 3.3.3, we verify the conditions in Assumption 3.1. On the other hand, since all rare-event sets in these examples are closed and contain unique optimal solutions for minimizing the corresponding rate function $I(x)$, we can verify Assumption 3.2 and the uniqueness of most significant dominating point.

Proof of Theorem 3.11. We check the three conditions in Theorem 3.4 to show the probabilistic efficiency of the IS estimator using the dominating point $a_{1}$. By symmetry we have $p_{1} / p \rightarrow 1 / 2$ as $\gamma \rightarrow \infty$. Then, following the argument in Section 3.9.1, the IS estimator using the dominating point $a_{1}$ is asymptotically efficient for $\mathcal{A}_{\gamma}^{1}$. Lastly we have $\tilde{p}_{2} \rightarrow 0$ exponentially fast and hence we have $n \tilde{p}_{2} \rightarrow 0$ with subexponentially growing sample size $n$. We conclude that the IS estimator using the dominating point $a_{1}$ is weakly probabilistically efficient.

On the other hand, based on Proposition 3.4, the IS estimator using the dominating point $a_{1}$ is
not asymptotically efficient for $\bigcup_{m=1}^{d} \mathcal{H}_{m}^{+}$and hence is not asymptotically efficient for

$$
\left(\bigcup_{m=1}^{d} \mathcal{H}_{m}^{+}\right) \bigcup\left(\bigcup_{m=1}^{d} \mathcal{H}_{m}^{-}\right) .
$$

Proof of Theorem 3.12. We only need to verify the conditions in Theorem 3.3. First, from Assumption 3.6 , we get that $p_{2}$ is exponentially smaller than $p_{1}$, and hence $\frac{p_{1}}{p} \rightarrow 1$ as $\gamma \rightarrow \infty$. Second, Let $Z_{1}=I\left(X \in \mathcal{E}_{1}\right) \frac{d P}{d \tilde{P}}$ under $\tilde{P}$. Clearly $\left\{a_{1}, \ldots, a_{k}\right\}$ is a dominating set for $\mathcal{E}_{1}$. We get that $\widetilde{\operatorname{Var}}\left(Z_{1}\right) \leq k^{2} e^{-2 I\left(a_{1}\right)} \leq r^{2} e^{-2 I\left(a_{1}\right)}$. Assumption 3.6 gives that $p_{1} \sim e^{-I\left(a_{1}\right)}$, and hence $Z_{1}$ is an asymptotically efficient estimator for $p_{1}$. Third, Assumption 3.7 implies that $n \tilde{p}_{2} \rightarrow 0$ for any $n$ subexponentially growing in $-\log p$. Therefore, all the conditions in Theorem 3.3 hold.

Proof of Theorem 3.13. We denote $\hat{p}_{1}=\frac{1}{n} \sum_{i=1}^{n} Z_{1}^{(i)}$ and $\mathbf{Z}_{1}=\left(Z_{1}^{(1)}, \ldots, Z_{1}^{(n)}\right)$. We also define $N=\sum_{i=1}^{n} I\left(X^{(i)} \in \mathcal{E}_{2}\right)$. Note that conditional on $N=0$, we have $\mathbf{Z}=\mathbf{Z}_{1}$. Then

$$
\begin{aligned}
& \tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \\
= & \tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}, N=0\right) \\
& +\tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}, N>0\right) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}, N=0\right)+P(N>0) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right)+P(N>0) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|+p_{2}>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right)+n \tilde{p}_{2} .
\end{aligned}
$$

We know that $0 \leq Z_{1}^{(i)} \leq k e^{-I\left(a_{1}\right)}$, $\forall i$. By applying Lemma 3.1 with $Y_{i}=Z_{1}^{(i)} /\left(k e^{-I\left(a_{1}\right)}\right)$ and
$Y_{i}=1-Z_{1}^{(i)} /\left(k e^{-I\left(a_{1}\right)}\right)$ respectively, we get that

$$
\tilde{P}\left(p_{1}>\hat{p}_{1}+\sqrt{\frac{2 \hat{V}_{1} \log (4 / \delta)}{n}}+\frac{7 \log (4 / \delta) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \leq \delta / 2
$$

and

$$
\tilde{P}\left(p_{1}<\hat{p}_{1}-\sqrt{\frac{2 \hat{V}_{1} \log (4 / \delta)}{n}}-\frac{7 \log (4 / \delta) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \leq \delta / 2
$$

for any $\delta>0$. Thus,

$$
\begin{equation*}
\tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \delta)}{n}}+\frac{7 \log (4 / \delta) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \leq \delta \tag{3.19}
\end{equation*}
$$

Find $\alpha^{\prime}=\alpha^{\prime}(\gamma)$ such that

$$
\frac{7 \log \left(4 / \alpha^{\prime}\right) k e^{-I\left(a_{1}\right)}}{3(n-1)}=\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}-p_{2}
$$

That is,

$$
\alpha^{\prime}=\alpha \exp \left(\frac{3(n-1) p_{2}}{7 k e^{-I\left(a_{1}\right)}}\right) .
$$

Clearly $\alpha^{\prime}>\alpha$ and $\log \left(4 / \alpha^{\prime}\right)<\log (4 / \alpha)$. Moreover, we know that $\frac{p_{2}}{e^{-I\left(a_{1}\right)}}$ decays exponentially in $-\log p$ since Assumption 3.6 holds and that $n$ grows subexponentially in $-\log p$, and thus $\alpha^{\prime} \rightarrow \alpha$ as $\gamma \rightarrow \infty$. We replace $\delta$ with $\alpha^{\prime}$ in (3.19), and then we get

$$
\tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>\sqrt{\frac{2 \hat{V}_{1} \log \left(4 / \alpha^{\prime}\right)}{n}}+\frac{7 \log \left(4 / \alpha^{\prime}\right) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \leq \alpha^{\prime}
$$

Hence

$$
\tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>\sqrt{\frac{2 \hat{V}_{1} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}-p_{2}\right) \leq \alpha^{\prime}
$$

Therefore,

$$
\tilde{P}\left(|\hat{p}-p|>\sqrt{\frac{2 \hat{V} \log (4 / \alpha)}{n}}+\frac{7 \log (4 / \alpha) k e^{-I\left(a_{1}\right)}}{3(n-1)}\right) \leq \alpha^{\prime}+n \tilde{p}_{2} \rightarrow \alpha \text { as } \gamma \rightarrow \infty
$$

Proof of Theorem 3.14. We denote $\hat{p}_{1}=\frac{1}{n} \sum_{i=1}^{n} Z_{1}^{(i)}$ and $\mathbf{Z}_{1}=\left(Z_{1}^{(1)}, \ldots, Z_{1}^{(n)}\right)$. We also define $N=\sum_{i=1}^{n} I\left(X^{(i)} \in \mathcal{E}_{2}\right)$. We have

$$
\begin{aligned}
& \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right. \\
= & \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}, N=0\right)+\tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}, N>0\right) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p\right|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}_{1}}{n}}\right)+P(N>0) \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}_{1}}{n}-p_{2}}\right)+n \tilde{p}_{2} \\
\leq & \tilde{P}\left(\left|\hat{p}_{1}-p_{1}\right|>z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}_{1}}{n}}-\sqrt{\frac{2 \log (1 / \delta)}{n(n-1)}} k e^{-I\left(a_{1}\right)}\right)-p_{2}\right) \\
& +\tilde{P}\left(\sqrt{\tilde{E} \hat{V}_{1}}>\sqrt{\hat{V}_{1}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} k e^{-I\left(a_{1}\right)}\right)+n \tilde{p}_{2}
\end{aligned}
$$

for any $\delta=\delta(\gamma)>0$. We know that $0 \leq Z_{1}^{(i)} \leq k e^{-I\left(a_{1}\right)}$, $\forall i$. By Lemma 3.3 with $Y_{i}=$ $Z_{1}^{(i)} /\left(k e^{-I\left(a_{1}\right)}\right)$, we get that

$$
\tilde{P}\left(\sqrt{\tilde{E} \hat{V}_{1}}>\sqrt{\hat{V}_{1}}+\sqrt{\frac{2 \log (1 / \delta)}{n-1}} k e^{-I\left(a_{1}\right)}\right) \leq \delta
$$

By Berry-Esseen theorem, we know that for any $x \in \mathbb{R}$

$$
\left|\tilde{P}\left(\frac{\sqrt{n}\left(\hat{p}_{1}-p_{1}\right)}{\sqrt{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} \leq x\right)-\Phi(x)\right| \leq \frac{C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}}
$$

where $\Phi$ is the CDF of standard normal distribution and $C$ is a universal constant. Let

$$
x=z_{1-\alpha / 2}\left(\sqrt{\frac{\tilde{E} \hat{V}_{1}}{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}-\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} k e^{-I\left(a_{1}\right)}\right)-p_{2} \sqrt{\frac{n}{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} .
$$

Then we get that

$$
\tilde{P}\left(\frac{\sqrt{n}\left|\hat{p}_{1}-p_{1}\right|}{\sqrt{\widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}>x\right) \leq 2 \Phi(-x)+\frac{2 C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}}
$$

Hence,

$$
\tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq 2 \Phi(-x)+\frac{2 C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}}+\delta+n \tilde{p}_{2}
$$

First, we have that $\tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3} \leq \max \left(p_{1}^{3},\left(k e^{-I\left(a_{1}\right)}-p_{1}\right)^{3}\right) \sim e^{-3 I\left(a_{1}\right)}$. Since Assumption 3.8 holds, we get that $\frac{\tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widehat{\operatorname{Var}^{3 / 2}\left(Z_{1}^{(1)}\right)}}$ grows at most subexponentially in $-\log p$. Hence under the assumptions we could choose $n$ as required in the theorem. In particular, we have that

$$
\frac{2 C \tilde{E}\left|Z_{1}^{(1)}-p_{1}\right|^{3}}{\widetilde{\operatorname{Var}}^{3 / 2}\left(Z_{1}^{(1)}\right) \sqrt{n}} \rightarrow 0
$$

Now we analyze $x$. We know that $\sqrt{\frac{E \hat{V}_{1}}{\overline{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}=1$ and

$$
p_{2} \sqrt{\frac{n}{\widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}=\frac{p_{2}}{e^{-I\left(a_{1}\right)}} \sqrt{\frac{n e^{-2 I\left(a_{1}\right)}}{\widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} \rightarrow 0
$$

since $\frac{p_{2}}{e^{-I\left(a_{1}\right)}}$ decays exponentially (Assumption 3.6) while $n$ and $\frac{e^{-2 I\left(a_{1}\right)}}{\widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}$ grow subexponentially
(Assumption 3.8) in $-\log p$. Now we consider

$$
\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} k e^{-I\left(a_{1}\right)}=\sqrt{\frac{2 k^{2} \log (1 / \delta) e^{-2 I\left(a_{1}\right)}}{(n-1) \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}}
$$

Since we assume that $\frac{k^{2} e^{-2 I\left(a_{1}\right)}}{n \widehat{\operatorname{Var}}\left(Z_{1}^{(1)}\right)} \rightarrow 0$, we could set $\delta$ such that $\delta \rightarrow 0$ and

$$
\sqrt{\frac{2 \log (1 / \delta)}{(n-1) \widetilde{\operatorname{Var}}\left(Z_{1}^{(1)}\right)}} k e^{-I\left(a_{1}\right)} \rightarrow 0
$$

In this case, $x \rightarrow z_{1-\alpha / 2}$ and hence $\Phi(-x) \rightarrow \alpha / 2$. Combining all the results, we get that

$$
\limsup _{\gamma \rightarrow \infty} \tilde{P}\left(|\hat{p}-p|>z_{1-\alpha / 2} \sqrt{\frac{\hat{V}}{n}}\right) \leq \alpha
$$

Proof of Theorem 3.15. First of all, it is easy to verify that the cumulant generating function $\mu(x)=$ $\lambda^{\top} x+\frac{1}{2} x^{\top} \Sigma x$ satisfies Assumption 3.3. Moreover, since $g$ is a piecewise linear function, we can express $\mathcal{E}$ as the union of finite closed polyhedrons. If $P(X \in \mathcal{E})>0$ and $\lambda \notin \mathcal{E}$, then $0<$ $\inf _{x \in \mathcal{E}} \frac{1}{2}(x-\lambda)^{\top} \Sigma^{-1}(x-\lambda)<\infty$. From [125], we know that

$$
p=P(g(X) \geq \gamma) \sim e^{-\frac{1}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)}
$$

and for sufficient large $\gamma,\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$ is a quadratic function of $\gamma$ which goes to $\infty$ as $\gamma \rightarrow \infty$. Finally, it is clear that the number of dominating points will not grow exponentially in $-\log p=\Theta\left(\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right)$. Therefore, Assumption 3.5 is satisfied in this problem setting.

Next, we check Assumptions 3.6 and 3.7. Without loss of generality, we may assume that $k<$ $r$, since otherwise $\mathcal{E}_{2}=\emptyset$ and $p_{2}=\tilde{p}_{2}=0$. We know that $a_{k+1}=\arg \min _{x \in \mathcal{E}_{2}}(x-\lambda)^{\top} \Sigma^{-1}(x-\lambda)$, and hence

$$
\mathcal{E}_{2} \subset\left\{x \in \mathbb{R}^{d}:(x-\lambda)^{\top} \Sigma^{-1}(x-\lambda) \geq\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)\right\} .
$$

Denote $X^{\prime}=\Sigma^{-1 / 2}(X-\lambda)$ and then $X^{\prime} \sim N\left(0, I_{d}\right)$ under $P$. We have that

$$
\begin{aligned}
p_{2} & \leq P\left((X-\lambda)^{\top} \Sigma^{-1}(X-\lambda) \geq\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)\right) \\
& =P\left(X^{\prime} T X^{\prime} \geq\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)\right) \\
& \sim e^{-\frac{1}{2}\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)} .
\end{aligned}
$$

We know that $\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)>C\left(a_{k}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k}-\lambda\right) \geq C\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$. Thus, Assumption 3.6 holds. Moreover,

$$
\begin{aligned}
\tilde{p}_{2} & =\frac{1}{k} \sum_{i=1}^{k} P_{X \sim N\left(a_{i}, \Sigma\right)}\left(X \in \mathcal{E}_{2}\right) \\
& \leq \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N\left(a_{i}, \Sigma\right)}\left((X-\lambda)^{\top} \Sigma^{-1}(X-\lambda) \geq\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)\right) .
\end{aligned}
$$

We have that

$$
\begin{aligned}
& \left(X-a_{i}\right)^{\top} \Sigma^{-1}\left(X-a_{i}\right) \\
= & (X-\lambda)^{\top} \Sigma^{-1}(X-\lambda)+\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)-2\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}(X-\lambda) \\
\geq & (X-\lambda)^{\top} \Sigma^{-1}(X-\lambda)+\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)-2 \sqrt{(X-\lambda)^{\top} \Sigma^{-1}(X-\lambda)} \sqrt{\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)} \\
= & \left(\sqrt{(X-\lambda)^{\top} \Sigma^{-1}(X-\lambda)}-\sqrt{\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)}\right)^{2} .
\end{aligned}
$$

Hence, if $(X-\lambda)^{\top} \Sigma^{-1}(X-\lambda) \geq\left(a_{k+1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k+1}-\lambda\right)>C\left(a_{k}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{k}-\lambda\right) \geq C\left(a_{i}-\right.$
$\lambda)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)$, then $\left(X-a_{i}\right)^{\top} \Sigma^{-1}\left(X-a_{i}\right) \geq(\sqrt{C}-1)^{2}\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)$. Thus we get that

$$
\begin{aligned}
\tilde{p}_{2} & \leq \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N\left(a_{i}, \Sigma\right)}\left(\left(X-a_{i}\right)^{\top} \Sigma^{-1}\left(X-a_{i}\right) \geq(\sqrt{C}-1)^{2}\left(a_{i}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{i}-\lambda\right)\right) \\
& \leq \frac{1}{k} \sum_{i=1}^{k} P_{X \sim N\left(a_{i}, \Sigma\right)}\left(\left(X-a_{i}\right)^{\top} \Sigma^{-1}\left(X-a_{i}\right) \geq(\sqrt{C}-1)^{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right) \\
& =P\left(X^{\prime T} X^{\prime} \geq(\sqrt{C}-1)^{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right) \\
& \sim e^{-\frac{1}{2}(\sqrt{C}-1)^{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)} .
\end{aligned}
$$

Hence, $\tilde{p}_{2}$ exponentially decays in $-\log p$. That is, Assumption 3.7 holds.
Finally, we verify Assumption 3.8. We have $\widetilde{\operatorname{Var}}\left(Z_{1}\right)=\tilde{E}\left(Z_{1}-p_{1}\right)^{2} \geq \tilde{E}\left(\left(Z_{1}-p_{1}\right)^{2} I_{\mathcal{E}_{1}^{c}}(X)\right)=$ $p_{1}^{2} \tilde{P}\left(X \notin \mathcal{E}_{1}\right)$. Since $p_{1} \sim e^{-\frac{1}{2}\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)}$, it suffices to justify that $\tilde{P}\left(X \notin \mathcal{E}_{1}\right)$ does not decay exponentially in $-\log p$. Indeed, we have that $\mathcal{E}_{1} \subset\left\{x \in \mathbb{R}^{d}:(x-\lambda)^{\top} \Sigma^{-1}(x-\lambda) \geq\right.$ $\left.\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right\}$ and thus

$$
\begin{aligned}
\tilde{P}\left(X \in \mathcal{E}_{1}\right) & =\frac{1}{k} \sum_{i=1}^{k} P_{X \sim N\left(a_{i}, \Sigma\right)}\left(X \in \mathcal{E}_{1}\right) \\
& \leq \frac{k-1}{k}+\frac{1}{k} P_{X \sim N\left(a_{1}, \Sigma\right)}\left(X \in \mathcal{E}_{1}\right) \\
& \leq \frac{k-1}{k}+\frac{1}{k} P_{X \sim N\left(a_{1}, \Sigma\right)}\left((X-\lambda)^{\top} \Sigma^{-1}(X-\lambda) \geq\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right) \\
& =\frac{k-1}{k}+\frac{1}{k} P\left(\left(X^{\prime}+\Sigma^{-1 / 2}\left(a_{1}-\lambda\right)\right)^{\top}\left(X^{\prime}+\Sigma^{-1 / 2}\left(a_{1}-\lambda\right)\right) \geq\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right) .
\end{aligned}
$$

Then we have that

$$
\begin{aligned}
\tilde{P}\left(X \notin \mathcal{E}_{1}\right) & \geq \frac{1}{k} P\left(\left(X^{\prime}+\Sigma^{-1 / 2}\left(a_{1}-\lambda\right)\right)^{\top}\left(X^{\prime}+\Sigma^{-1 / 2}\left(a_{1}-\lambda\right)\right)<\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)\right) \\
& \geq \frac{1}{r} P\left(X^{\prime} \in B\left(-\Sigma^{-1 / 2}\left(a_{1}-\lambda\right), \sqrt{\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)}\right)\right)
\end{aligned}
$$

where $B(x, R):=\left\{y \in \mathbb{R}^{d}:\|y-x\|<R\right\}$. For sufficiently large $\gamma,\left(a_{1}-\lambda\right)^{\top} \Sigma^{-1}\left(a_{1}-\lambda\right)$ monotonely grows to $\infty$, and hence $\tilde{P}\left(X \notin \mathcal{E}_{1}\right) \geq \frac{c}{r}$ for some constant $c>0$. As a result, Assumption 3.8 holds.

# Chapter 4: On the Error of Naive Rare-Event Monte Carlo Estimator 

### 4.1 Introduction

We consider the problem of estimating a minuscule probability, denoted $p=P(A)$, for some rare event $A$, using data or Monte Carlo samples. This problem, known as rare-event estimation, is of wide interest to communities such as system reliability [43, 40, 42, 44], queueing systems [24, $25,26,27,29,30,31]$, finance and insurance $[34,36,35,134,37,38,39]$, where it is crucial to estimate the likelihood of events which, though unlikely, can cause catastrophic impacts.

There are multiple prominent lines of work addressing this estimation problem, depending on how information is collected. In settings where real-world data are collected, methods based on extreme value theory $[135,136,134,137]$ are often used to extrapolate distributional tails to assist such estimation. Despite the theoretical guarantees and wide applications of these methods, the performance still depends on subjective choices such as the hyperparameters and the approach to fit the tail distribution.

In settings where $A$ is an event described by a simulatable model, Monte Carlo methods can be used, and to speed up computation one often harnesses variance reduction tools such as importance sampling [51, 53, 23], conditional Monte Carlo [54, 63] and multi-level splitting [87, 138, 139]. While variance reduction is greatly beneficial in reducing the number of Monte Carlo samples needed to estimate rare events [22,54,63], it is also widely known that they rely heavily on model assumptions [23, 64]. That is, to guarantee the successful performances of these techniques, one typically needs to analyze the underlying model dynamics carefully to design the Monte Carlo scheme. However, recent applications, such as autonomous vehicle safety evaluation [17, 20, 18, 19,21 ] and robustness evaluation of machine learning predictors [14, 140, 125], lead to rare-event estimation problems with extremely sophisticated structures that hinder the design of efficiency-
guaranteed variance reduction schemes. On the other hand, with the remarkable recent surge of computational infrastructure, in some situations one could afford to run gigantic amount of simulation trials.

Motivated by the limitations of the above techniques and the potential to generate numerous samples, in this chapter we focus on a more basic setting than some of the above literature, but in a sense fundamental. More precisely, we focus on the situation where all we have to estimate $p$ is a set of i.i.d. Bernoulli observations $I(A)$. A natural point estimate of $p$ is the sample proportion $\hat{p}$, i.e., given a set of Bernoulli data $I_{1}, \ldots, I_{n}$ of size $n$, we output $\hat{p}=(1 / n) \sum_{i=1}^{n} I_{i}$. We are interested in understanding the statistical error in using $\hat{p}$, in the situation where $p$ could be very small, importantly with no lower bound on how small it could be. Unlike the estimates given by efficiency-guaranteed variance reduction techniques, it is open, at least to our best knowledge, whether using simple sample proportion can give meaningful guarantee to estimating rare-event probabilities, in relation to the sample size $n$ and the (unknown) magnitude of $p$.

In Section 4.2, we will describe the problem setting and the challenges more concretely, and also summarize our main results. In Section 4.3 and 4.4, we construct and analyze various confidence intervals (CI) under two different settings respectively. In Section 4.5, we conduct numerical experiments to visualize the comparison. In Section 4.6, we conclude this chapter with our findings and recommendations. All the proofs can be found in Section 4.7.

### 4.2 Problem Setting and Main Results

In this chapter we would focus on the construction of CI. That is, using information from the Bernoulli data, or equivalently $\hat{p}$, we would like to construct a simple CI for $p$ that has justifiable statistical guarantees. In answering this, we would also quantify the error between the point estimate $\hat{p}$ and $p$.

First of all, we explain what a good CI is supposed to be. In order to measure the goodness of a CI, we mainly consider the validity and the tightness. Throughout this chapter, we say that [ $\hat{p}_{l}(\alpha), \hat{p}_{u}(\alpha)$ ] is a valid $(1-\alpha)$-level CI if $P\left(\hat{p}_{l}(\alpha) \leq p \leq \hat{p}_{u}(\alpha)\right) \geq 1-\alpha$, i.e., the nominal con-
fidence level is achieved. The validity can be defined similarly for one-sided confidence bounds. A good CI should be valid or approximately valid such that it covers the true probability $p$ with a high confidence. On the other hand, a good CI should not be too wide. In an extreme case, the trivial CI $[0,1]$ always covers $p$, but it does not provide any useful information. The tightness could be quantified by the "half-width", i.e., $\hat{p}_{u}(\alpha)-\hat{p}$ or $\hat{p}-\hat{p}_{l}(\alpha)$. Considering that $p$ is tiny in the rare-event settings, the CI is meaningful only if the half-width is small relative to $p$ and $\hat{p}$.

To understand the challenges, we first examine the use of a standard "textbook" CI, and we focus on the upper confidence bound for now since the lower confidence bound can be argued analogously. More specifically, we use the following as the $(1-\alpha)$-level upper confidence bound

$$
\begin{equation*}
\hat{p}^{C L T}=\hat{p}+z_{1-\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4.1}
\end{equation*}
$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of a standard normal variable. The typical way to justify (4.1) is a Gaussian approximation using the central limit theorem (CLT), which entails that

$$
\begin{equation*}
P\left(p \leq \hat{p}^{C L T}\right) \approx \bar{\Phi}\left(-z_{1-\alpha}\right)=1-\alpha \tag{4.2}
\end{equation*}
$$

where we denote $\bar{\Phi}($ and $\Phi)$ as the tail (and cumulative) distribution function of standard normal.
To delve a little further, note that the approximation error in (4.2) is controlled by the BerryEssen (B-E) Theorem. To simplify the discussion, suppose we are in a more idealized (but unrealistic) case that we know the precise value of the variance of the Bernoulli trial, i.e., $\sigma^{2}=p(1-p)$, so that we use $\hat{p}+z_{1-\alpha} \sigma / \sqrt{n}$. Then the Berry-Essen theorem stipulates that

$$
\begin{equation*}
\left|P\left(p \leq \hat{p}^{C L T}\right)-\Phi\left(z_{1-\alpha}\right)\right| \leq \frac{C \rho}{\sigma^{3} \sqrt{n}} \tag{4.3}
\end{equation*}
$$

where $\rho=E\left|I_{i}-p\right|^{3}=p(1-p)\left(1-2 p+2 p^{2}\right)$, and $C$ is a universal constant $(\approx 0.4748)$. Thus,
the error in (4.3) is bounded by

$$
\begin{equation*}
\frac{C p(1-p)\left(1-2 p+2 p^{2}\right)}{p^{3 / 2}(1-p)^{3 / 2} \sqrt{n}} \leq \frac{C}{\sqrt{n p(1-p)}} \tag{4.4}
\end{equation*}
$$

The issue is that when $p$ is tiny, $n p$ can also be tiny unless $n$ is sufficiently big, but a priori we would not know what $n$ is "sufficient". If we have used the confidence bound given by (4.1) where the variance $\sigma^{2}$ is unknown and estimated by $\hat{p}(1-\hat{p})$, a similar Berry-Esseen bound would ultimately conclude the same issue as revealed by (4.4) [141]. These lead to the following exemplified questions:

Q1. Suppose we have, say, 30 "success" outcomes among $n$ trials, then we may think that $n p \approx 30$, so that from the bound (4.4) the error of $\hat{p}^{C L T}$ appears controlled. As another more extreme case, suppose we only have only 1 success, then we may be led to believe $n p \approx 1$, so that $\hat{p}^{C L T}$ is not trivial but its coverage is likely way off from $1-\alpha$. Are these conclusions on $\hat{p}^{C L T}$ correct? Note that the guess that $n p \approx 30$ or $n p \approx 1$ is itself based on some central limit or concentration argument, which apparently leads to a circular reasoning.

Q2. If we have 1 success among $n$ trials, how do we construct a confidence upper bound that is guaranteed the correct coverage? Correspondingly, would constructing such a confidence bound be easier if we have 30 successes?

Q3. For the valid confidence bounds constructed in Question 2, what is the typical "half-width" (i.e., the difference between the confidence bound and $\hat{p}$ ), relative to the point estimate $\hat{p}$ ? If the relative half-width is too big, then the provided upper bound may not be meaningful. For instance, if $n$ is 0 , then (4.1) is clearly meaningless. What about if $n$ is 1 when we use a valid confidence bound? Note that the use of relative half-width is important since $p$ (and correspondingly $\hat{p}$ ) could be tiny, and a meaningful upper bound should have a similar magnitude as $p($ and $\hat{p})$.

Q4. Following up Question 3, we would also like to understand the relative error of the point
estimate, i.e., $(\hat{p}-p) / p$, and the relative error of the confidence bound, e.g., $\left(\hat{p}^{C L T}-p\right) / p$, the latter related to Question 3.

Q5. Do all the above answers hold if we stop whenever we observe enough successes (e.g., when the number of successes is 30 , or 1 ) in our simulation experiment?

Note that a quick and implementable approach to Question 2 is to utilize the fact that $n \hat{p}$ follows a binomial distribution and extract a finite-sample confidence region using this exact distribution. This is often called the Clopper-Pearson CI or the exact method. Though this is computationally easy, we are interested in simpler mathematical forms that allow us to answer Questions 3 and 4 above. In this regard, Wilson's interval [142] has been studied and shown to give superior empirical performances, even in the case that $p$ is tiny, but we are not aware of any rigorous proof on its validity. In the following sections, we will offer two different ways of constructing CIs for $p$, one using a concentration inequality called Chernoff's inequality, while another one using the Berry-Esseen bound. They both offer answers to Question 2 on top of the exact CI and Wilson's interval. Compared to the exact CI, these two intervals are in a similar form as the CLT interval, which allows us to answer Question 1. They also have the explicit forms that allow us to investigate their half-widths, thereby answering Question 3 and 4. Finally, they could be adapted to the setting in Question 5.

Before going into details, now we provide a roadmap for our main results. We will focus on two settings. The first one is called the "standard" setting, where the sample size $n$ is fixed. The other setting is as described in Question 5 where we fix the number of successes, and we call it as the "targeted stopping" setting. Under each setting, we will review the CLT CI, the Wilson's CI and the exact CI. We will also introduce how to construct the Chernoff CI and the B-E CI respectively via inverting the Chernoff's inequality and the $\mathrm{B}-\mathrm{E}$ theorem.

From (4.1), we clearly see that the half-width of the CLT CI scales in the same magnitude as $\sqrt{\hat{p} / n}=\hat{p} / \sqrt{\hat{n}}$ where $\hat{n}=n \hat{p}$ is the number of positive outcomes. With this transformation, it is easier to see how the half-width scales relative to $\hat{p}$ as asked in Question 3. That is, the relative half-width is of magnitude $1 / \sqrt{\hat{n}}$. Figure 4.1 illustrates the comparisons of these available CIs
in terms of upper and lower bound (the exact CI is not included since it is hard to analyze its magnitude). More concretely, Table 4.1 summarizes the formula, scale, pros and cons of each CI under both settings. For the confidence upper bounds under the standard setting, we have that the B-E bound is valid yet the difference between it and the CLT bound is of magnitude $\hat{p} / \hat{n}$, so the relative difference with respect to $\hat{p}$ is of magnitude $1 / \hat{n}$, which is of higher order in $\hat{n}$. This can be viewed as a price of validity paid to make $\hat{p}^{C L T}$ correct. We will also show that similar arguments hold for the Wilson's bound. The same conclusions could be achieved for both upper and lower bounds under the targeted stopping setting. However, B-E fails to give a non-trivial confidence lower bound under the standard setting, and thus it remains to be an open problem how close is the CLT (or the Wilson's) lower bound to a valid one.

(a) Standard Setting - Upper
(b) Standard Setting Lower
(c) Targeted Stopping Upper
(d) Targeted Stopping Lower

Figure 4.1: Comparisons of confidence upper and lower bounds under the standard and the targeted stopping setting.

### 4.3 Confidence Intervals under the Standard Setting

Under the standard setting, we fix the sample size $n$ and let $\hat{p}=(1 / n) \sum_{i=1}^{n} I_{i}$ where $I_{i}$ 's are i.i.d. Bernoulli (p) random variables.

| Standard Setting ( $I_{1}, \ldots, I_{n} \stackrel{\text { i.i.d. }}{\sim}$ Bernoulli $\left.(p), \hat{p}=\frac{1}{n} \sum_{i=1}^{n} I_{i}, \hat{n}=n \hat{p}\right)$ |  |  |
| :---: | :---: | :---: |
| CLT | $\hat{p} \pm z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} .$ | Scale: $\hat{p}_{u}^{C L T}-\hat{p}=\hat{p}-\hat{p}_{l}^{C L T}=O(\hat{p} / \sqrt{\hat{n}})$; <br> Pros: Easy to compute; <br> Cons: Not reliable when $n p$ is not sufficiently large. |
| Wilson | $\frac{1+\frac{2 n \hat{p}}{z_{1-\alpha / 2}^{2}} \pm \sqrt{1+\frac{4 n \hat{\rho}(1-\hat{\rho})}{z_{1-\alpha / 2}^{2}}}}{2\left(1+\frac{n}{z_{1-\alpha / 2}^{2}}\right)} .$ | Scale: $\left\|\hat{p}_{u}^{\text {Wilson }}-\hat{p}_{u}^{C L T}\right\|=O(\hat{p} / \hat{n}),\left\|\hat{p}_{l}^{\text {Wilson }}-\hat{p}_{l}^{C L T}\right\|=O(\hat{p} / \hat{n})$; <br> Pros: Tight, the coverage probability is usually close to $1-\alpha$; Cons: Not always valid, lacks theoretical error control. |
| Exact | Solutions to $\begin{aligned} & \sum_{k=0}^{\hat{n}}\binom{n}{k} p^{k}(1-p)^{n-k}=\alpha / 2, \\ & \sum_{k=\hat{n}}^{n}\binom{n}{k} p^{k}(1-p)^{n-k}=\alpha / 2 \end{aligned}$ <br> except that $\hat{p}_{u}^{\text {Exact }}=1$ if $\hat{n}=n$ and $\hat{p}_{l}^{\text {Exact }}=0$ if $\hat{n}=0$. | Pros: Always valid, tighter than other valid CI; Cons: Conservative, hard to analyze. |
| Chernoff | $\begin{aligned} & \hat{p}+\frac{\log (2 / \alpha)}{n}+\sqrt{\frac{(\log (2 / \alpha))^{2}}{n^{2}}+\frac{2 \hat{p} \log (2 / \alpha)}{n}}, \\ & \hat{p}+\frac{\log (2 / \alpha)}{2 n}-\sqrt{\frac{(\log (2 / \alpha))^{2}}{4 n^{2}}+\frac{2 \hat{p} \log (2 / \alpha)}{n}} . \end{aligned}$ | Scale: $\hat{p}_{u}^{\text {Chernoff }}-\hat{p}_{u}^{C L T}=O(\hat{p} / \sqrt{\hat{n}}), \hat{p}_{l}^{C L T}-\hat{p}_{l}^{\text {Chernoff }}=O(\hat{p} / \sqrt{\hat{n}})$; Pros: Always valid, helps us understand the relative error of $\hat{p}$; Cons: Extremely conservative. |
| B-E | $\begin{gathered} \left\{0<p \leq \hat{p} \wedge \frac{1}{2}: \Phi\left(\frac{p-\hat{p}}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \frac{\alpha}{2}\right\} \\ \cup\left\{\hat{p} \leq p<\frac{1}{2}: \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \frac{\alpha}{2}\right\} \end{gathered}$ <br> where $C$ is the universal constant in the $\mathrm{B}-\mathrm{E}$ theorem. | Scale: $\left\|\hat{p}_{u}^{B E}-\hat{p}_{u}^{C L T}\right\|=O(\hat{p} / \hat{n})$; <br> Pros: Always valid, helps us understand the error of the CLT upper bound; Cons: Extremely conservative, trivial lower bound. |
| Targeted Stopping Setting $\left(N_{1}, \ldots, N_{n_{0}} \stackrel{\text { i.i.d. }}{\sim}\right.$ Geometric $\left.(p), N=\sum_{i=1}^{n_{0}} N_{i}, \hat{p}=n_{0} / N\right)$ |  |  |
| CLT | $\frac{n_{0}}{N} \pm z_{1-\alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}}$. | Scale: $\hat{p}_{u, n_{0}}^{C L T}-\hat{p}=\hat{p}-\hat{p}_{l, n_{0}}^{C L T}=O\left(\sqrt{n_{0}} / N\right)$; <br> Pros: Easy to compute; <br> Cons: Not always valid. |
| Wilson | $\frac{1+\frac{2 n_{0}}{z_{1-\alpha / 2}} \pm \sqrt{1+\frac{4 n_{0}\left(N-n_{0}\right)}{z_{1-\alpha / 2^{2}}^{N}}}}{2\left(1+\frac{N}{z_{1-\alpha / 2}^{2}}\right)} .$ | Scale: $\left\|\hat{p}_{u, n_{0}}^{\text {Wilson }}-\hat{p}_{u, n_{0}}^{C L T}\right\|=O(1 / N),\left\|\hat{p}_{l, n_{0}}^{\text {Wilson }}-\hat{p}_{l, n_{0}}^{C L T}\right\|=O(1 / N)$; <br> Pros: Tight, the coverage probability is usually close to $1-\alpha$; Cons: Not always valid. |
| Exact | Solutions to $\begin{gathered} \sum_{k=0}^{N-n_{0}-1}\binom{k+n_{0}-1}{n_{0}-1}(1-p)^{k} p^{n_{0}}=1-\alpha / 2, \\ \sum_{k=0}^{N-n_{0}}\binom{k+n_{0}-1}{n_{0}-1}(1-p)^{k} p^{n_{0}}=\alpha / 2 \\ \quad \text { except that } \hat{p}_{u, n_{0}}^{\text {Exact }}=1 \text { if } N=n_{0} . \end{gathered}$ | Pros: Always valid, tighter than other valid CI; Cons: Conservative, hard to analyze. |
| Chernoff | $\left\{0<p<1: p^{n_{0}}(1-p)^{N-n_{0}} \geq \frac{\alpha}{2}\left(\frac{n_{0}}{N}\right)^{n_{0}}\left(1-\frac{n_{0}}{N}\right)^{N-n_{0}}\right\}$. | Pros: Always valid; <br> Cons: Extremely conservative, hard to analyze. |
| B-E | $\begin{aligned} & \left\{0<p \leq \frac{n_{0}}{N} \wedge \frac{1}{2}: \Phi\left(\frac{N p-n_{0}}{\sqrt{n_{0}(1-p)}}\right)+\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq \frac{\alpha}{2}\right\} \\ & \cup\left\{\frac{n_{0}}{N} \leq p<\frac{1}{2}: \Phi\left(\frac{n_{0}-N p}{\sqrt{n_{0}(1-p)}}\right)+\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq \frac{\alpha}{2}\right\} \end{aligned}$ <br> where $C^{\prime}=16 C$ is a universal constant. | Scale: $\hat{p}_{u, n_{0}}^{B E}-\hat{p}_{u, n_{0}}^{C L T}=O(1 / N), \hat{p}_{l, n_{0}}^{C L T}-\hat{p}_{l, n_{0}}^{B E}=O(1 / N)$; <br> Pros: Aways valid, helps us understand the error of $\hat{p}$ and the CLT CI; Cons: Extremely conservative, trivial for small $n_{0}$. |

Table 4.1: Summary of the CIs.

### 4.3.1 Confidence Intervals via Normal Approximation

First, we review two existing CIs which are constructed via normal approximation: the CLT interval and the Wilson's interval. We know that $(\hat{p}-p) / \sqrt{p(1-p) / n} \xrightarrow{d} N(0,1)$, and then normal distribution could serve as an approximation.

## CLT CI

Using $\hat{p}(1-\hat{p})$ to approximate $\sigma^{2}=p(1-p)$, we get the CLT CI:

$$
\begin{aligned}
& \hat{p}_{u}^{C L T}=\hat{p}+z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \\
& \hat{p}_{l}^{C L T}=\hat{p}+z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} .
\end{aligned}
$$

Clearly, $\hat{p}_{u}^{C L T}-\hat{p}=O(\sqrt{\hat{p} / n})=O(\hat{p} / \sqrt{\hat{n}})$ and $\hat{p}-\hat{p}_{l}^{C L T}=O(\hat{p} / \sqrt{\hat{n}})$. As explained in Section 4.2, we express the half-width as $O(\hat{p} / \sqrt{\hat{n}})$ instead of $O(\sqrt{\hat{p} / n})$ in order to understand the magnitude of the relative half-width with respect to $\hat{p}$ more clearly.

## Wilson's CI

Now we do not approximate the variance. Instead, we directly solve

$$
\frac{\alpha}{2} \leq \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right) \leq 1-\frac{\alpha}{2}
$$

and then get the Wilson's CI:

$$
\begin{aligned}
& \hat{p}_{u}^{\text {Wilson }}=\frac{1+\frac{2 n \hat{p}}{z_{1-\alpha / 2}^{2}}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{1-\alpha / 2}^{2}}}}{2\left(1+\frac{n}{z_{1-\alpha / 2}^{2}}\right)} ; \\
& \hat{p}_{l}^{\text {Wilson }}=\frac{1+\frac{2 n \hat{p}}{z_{1-\alpha / 2}^{2}}-\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{1-\alpha / 2}^{2}}}}{2\left(1+\frac{n}{z_{1-\alpha / 2}^{2}}\right)} .
\end{aligned}
$$

It can be easily derived that

## Theorem 4.1.

$$
\begin{aligned}
& \left|\hat{p}_{u}^{\text {Wilson }}-\hat{p}_{u}^{C L T}\right| \leq \frac{z_{1-\alpha / 2}^{2}}{n}+\frac{z_{1-\alpha / 2}^{3}}{2 n^{3 / 2}} \\
& \left|\hat{p}_{l}^{\text {Wilson }}-\hat{p}_{l}^{C L T}\right| \leq \frac{z_{1-\alpha / 2}^{2}}{n}+\frac{z_{1-\alpha / 2}^{3}}{2 n^{3 / 2}}
\end{aligned}
$$

Remark 4.1. Note that $1 / n=\hat{p} / \hat{n}$, so the difference between the Wilson's CI and the CLT CI is of magnitude $O(\hat{p} / \hat{n})$, which is of higher order than $\hat{p} / \sqrt{\hat{n}}$ in $\hat{n}$. Since the half-width of the CLT CI is of magnitude $\hat{p} / \sqrt{\hat{n}}$, we get that the half-width of the Wilson's CI has similar scale to the CLT CI.

In practice, the Wilson's confidence bound has a satisfactory performance, in the sense that it is relatively tight while the coverage probability is usually close to the nominal confidence level.

### 4.3.2 Valid Confidence Intervals

To construct valid CIs, we consider the following set:

$$
\begin{equation*}
\left\{0<p<1: F(\hat{p}) \geq \alpha / 2, F_{-}(\hat{p}) \leq 1-\alpha / 2\right\} \tag{4.5}
\end{equation*}
$$

where $F(x)=P(\hat{p} \leq x)$ and $F_{-}(x)=P(\hat{p}<x)$. Note that $F$ and $F_{-}$depend on $p$. If $F$ were continuous, then we know that $P\left(F(\hat{p}) \geq \alpha / 2, F_{-}(\hat{p}) \leq 1-\alpha / 2\right)=1-\alpha$ since in this case $F(\hat{p})=$ $F_{-}(\hat{p}) \stackrel{d}{=}$ Unif $[0,1]$. Now we argue that $P\left(F(\hat{p}) \geq \alpha / 2, F_{-}(\hat{p}) \leq 1-\alpha / 2\right)>1-\alpha$ in this discrete case. Indeed, for any $\alpha \in(0,1)$, there exist $0 \leq k, l \leq n$ such that $F((k-1) / n)<\alpha / 2 \leq F(k / n)$ and $F_{-}(l / n) \leq 1-\alpha / 2<F_{-}((l+1) / n)$. Then

$$
\begin{aligned}
P\left(F(\hat{p})<\alpha / 2 \text { or } F_{-}(\hat{p})>1-\alpha / 2\right) & \leq P(F(\hat{p})<\alpha / 2)+P\left(F_{-}(\hat{p})>1-\alpha / 2\right) \\
& =P(\hat{p} \leq(k-1) / n)+P(\hat{p} \geq(l+1) / n) \\
& =F((k-1) / n)+1-F_{-}((l+1) / n)<\alpha .
\end{aligned}
$$

Therefore, the set (4.5) is a valid $(1-\alpha)$-level confidence region. From this derivation, we find that due to the discreteness, the probability that this confidence region covers the true value $p$ is strictly larger than the nominal confidence level $1-\alpha$, and hence this confidence region is inevitably conservative.

## Exact CI

The exact CI is obtained by directly solving (4.5). More specifically, $\hat{p}_{u}^{\text {Exact }}$ and $\hat{p}_{l}^{\text {Exact }}$ are respectively the solution to

$$
\begin{aligned}
& \sum_{k=0}^{\hat{n}}\binom{n}{k} p^{k}(1-p)^{n-k}=\alpha / 2 \\
& \sum_{k=\hat{n}}^{n}\binom{n}{k} p^{k}(1-p)^{n-k}=\alpha / 2
\end{aligned}
$$

except that $\hat{p}_{u}^{\text {Exact }}=1$ if $\hat{n}=n$ and $\hat{p}_{l}^{\text {Exact }}=0$ if $\hat{n}=0$. When $0<\hat{n}<n$, the bounds could be expressed explicitly via quantiles of $F$ distribution or Beta distribution, and hence are easy to compute numerically [142]. However, it is hard to analyze the scale of this CI, which motivates us to further relax the confidence region (4.5) to get other valid CIs which are more conservative but easier to analyze.

## Chernoff CI

Now we present our first approach to construct a valid CI for $p$ by utilizing a concentration inequality. By Chernoff's inequality, we have

$$
\begin{aligned}
& P(\hat{p} \leq(1-\delta) p) \leq \exp \left(-\frac{\delta^{2}}{2} n p\right), 0<\delta<1 \\
& P(\hat{p} \geq(1+\delta) p) \leq \exp \left(-\frac{\delta^{2}}{2+\delta} n p\right), \delta>0
\end{aligned}
$$

Replacing $(1-\delta) p$ or $(1+\delta) p$ by $x$, we have

$$
\begin{aligned}
& F(x) \leq \exp \left\{-\left(1-\frac{x}{p}\right)^{2} \frac{n p}{2}\right\}, x \leq p \\
& F_{-}(x) \geq 1-\exp \left\{-\frac{\left(\frac{x}{p}-1\right)^{2}}{1+\frac{x}{p}} n p\right\}, x \geq p
\end{aligned}
$$

Hence $F(\hat{p}) \geq \alpha / 2, F_{-}(\hat{p}) \leq 1-\alpha / 2$ implies that either

$$
p \geq \hat{p} \text { and } \exp \left\{-\left(1-\frac{\hat{p}}{p}\right)^{2} \frac{n p}{2}\right\} \geq \alpha / 2
$$

or

$$
p \leq \hat{p} \text { and } 1-\exp \left\{-\frac{\left(\frac{\hat{p}}{p}-1\right)^{2}}{1+\frac{\hat{p}}{p}} n p\right\} \leq 1-\alpha / 2
$$

Therefore,

$$
\left\{0<p \leq \hat{p}: \exp \left\{-\frac{\left(\frac{\hat{p}}{p}-1\right)^{2}}{1+\frac{\hat{p}}{p}} n p\right\} \geq \alpha / 2\right\} \cup\left\{\hat{p} \leq p<1: \exp \left\{-\left(1-\frac{\hat{p}}{p}\right)^{2} \frac{n p}{2}\right\} \geq \alpha / 2\right\}
$$

is a confidence region for $\hat{p}$ with confidence level at least $1-\alpha$. Simplifying the expression above, we have that

$$
\begin{aligned}
& 0<p \leq \hat{p}, \exp \left\{-\frac{\left(\frac{\hat{p}}{p}-1\right)^{2}}{1+\frac{\hat{p}}{p}} n p\right\} \geq \alpha / 2 \\
\Rightarrow & \hat{p}+\frac{\log (2 / \alpha)}{2 n}-\sqrt{\frac{(\log (2 / \alpha))^{2}}{4 n^{2}}+\frac{2 \hat{p} \log (2 / \alpha)}{n}} \leq p \leq \hat{p}
\end{aligned}
$$

and

$$
\begin{aligned}
& \hat{p} \leq p<1, \exp \left\{-\left(1-\frac{\hat{p}}{p}\right)^{2} \frac{n p}{2}\right\} \geq \alpha / 2 \\
\Rightarrow & \hat{p} \leq p \leq \hat{p}+\frac{\log (2 / \alpha)}{n}+\sqrt{\frac{(\log (2 / \alpha))^{2}}{n^{2}}+\frac{2 \hat{p} \log (2 / \alpha)}{n}} .
\end{aligned}
$$

Hence, by taking the union, we get a valid $(1-\alpha)$-level CI for $p$, for any finite sample $n$. This can be summarized as:

Theorem 4.2. The interval given by

$$
\begin{aligned}
& \hat{p}_{u}^{\text {Chernoff }}=\hat{p}+\frac{\log (2 / \alpha)}{n}+\sqrt{\frac{(\log (2 / \alpha))^{2}}{n^{2}}+\frac{2 \hat{p} \log (2 / \alpha)}{n}}, \\
& \hat{p}_{l}^{\text {Chernoff }}=\hat{p}+\frac{\log (2 / \alpha)}{2 n}-\sqrt{\frac{(\log (2 / \alpha))^{2}}{4 n^{2}}+\frac{2 \hat{p} \log (2 / \alpha)}{n}}
\end{aligned}
$$

is a valid $(1-\alpha)$-level CI for $p$, for any finite sample $n$. That is,

$$
P\left(\hat{p}_{l}^{\text {Chernoff }} \leq p \leq \hat{p}_{u}^{\text {Chernoff }}\right) \geq 1-\alpha
$$

for any $n$.

When $\hat{p}=0$, the interval reduces to $[0,2 \log (1 / \alpha) / n]$ (and in fact we can construct even tighter bounds by using the binomial distribution of $n \hat{p}$ directly in this case). On the other hand, when $\hat{p}>0$, we can re-express using $\hat{n}=n \hat{p}$ to get

$$
\begin{aligned}
& \hat{p}_{u}^{\text {Chernoff }}=\hat{p}\left(1+\frac{\log (2 / \alpha)}{\hat{n}}+\sqrt{\frac{(\log (2 / \alpha))^{2}}{\hat{n}^{2}}+\frac{2 \log (2 / \alpha)}{\hat{n}}}\right), \\
& \hat{p}_{l}^{\text {Chernoff }}=\hat{p}\left(1+\frac{\log (2 / \alpha)}{2 \hat{n}}-\sqrt{\frac{(\log (2 / \alpha))^{2}}{4 \hat{n}^{2}}+\frac{2 \log (2 / \alpha)}{\hat{n}}}\right) .
\end{aligned}
$$

We highlight that in this case, the half-width of the Chernoff CI is of order $O(\hat{p} / \sqrt{\hat{n}})$, which scales in the same order as the CLT CI.

If we check the difference between this interval and the CLT interval, then we will find that it is of the same order as the difference between the CLT interval and $\hat{p}$. The following theorem presents the details of this claim. We will contrast this result with another one presented in the next section momentarily.

## Theorem 4.3.

$$
\begin{aligned}
& \hat{p}_{u}^{\text {Chernoff }}-\hat{p}_{u}^{C L T} \geq\left(\sqrt{2 \log (2 / \alpha)}-z_{1-\alpha / 2}\right) \sqrt{\frac{\hat{p}}{n}}+\frac{\log (2 / \alpha)}{n}, \\
& \hat{p}_{l}^{C L T}-\hat{p}_{l}^{\text {Chernoff }} \geq\left(\sqrt{2 \log (2 / \alpha)}-z_{1-\alpha / 2}\right) \sqrt{\frac{\hat{p}}{n}}-\frac{\log (2 / \alpha)}{2 n} .
\end{aligned}
$$

Note that $\sqrt{2 \log (2 / \alpha)}-z_{1-\alpha / 2}>0$ for $0<\alpha<1$.

Remark 4.2. One may note that as long as $\hat{n} \geq 1$, that is, we have at least one positive observation, then $\sqrt{\hat{p} / n}=\sqrt{\hat{n}} / n \geq 1 / n$. Provided that $\sqrt{2 \log (2 / \alpha)}-z_{1-\alpha / 2}>0, \hat{p}_{u}^{\text {Chernoff }}-\hat{p}_{u}^{C L T}$ (or, $\left.\hat{p}_{l}^{\text {Chernoff }}-\hat{p}_{l}^{C L T}\right)$ is of no higher order than $\hat{p}_{u}^{C L T}-\hat{p}\left(\right.$ or, $\left.\hat{p}_{l}^{C L T}-\hat{p}\right)$.

Finally, we can conclude that $P(|\hat{p}-p| \leq \epsilon \hat{p} \mid \hat{n} \geq c) \geq 1-\alpha$ for some small $\epsilon$ depending on the constant $c$, which is chosen large enough independently. This in turn concludes that $P(|\hat{p}-p| \leq$ $(\epsilon /(1-\epsilon)) p \mid \hat{n} \geq c) \geq 1-\alpha$. In other words, given we observe that $\hat{n}$ is large enough, $\hat{p}$ can be viewed as a conditionally relatively efficient estimator. The caution, however, is that this is under the precondition that we can observe enough positive outcomes.

## Berry-Esseen CI

We develop another CI for $p$ by inverting the Berry-Esseen theorem. Here, we assume that $p$ is known to satisfy $p<\frac{1}{2}$ a priori (which is reasonable if we consider rare event). In this chapter, we use the standard version of the Berry-Esseen theorem, and a potential future investigation is to consider a B-E bound for the studentized statistic [143, 144].

By the Berry-Esseen theorem, we have that

$$
\begin{aligned}
& \left|P\left((\hat{p}-p) \sqrt{\frac{n}{p(1-p)}} \leq x\right)-\Phi(x)\right| \leq \frac{C}{\sqrt{n p(1-p)}} \\
& \left|P\left((p-\hat{p}) \sqrt{\frac{n}{p(1-p)}} \leq x\right)-\Phi(x)\right| \leq \frac{C}{\sqrt{n p(1-p)}}
\end{aligned}
$$

where $C$ is a universal constant. We replace $x$ by $\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}$ in the first inequality and $\frac{p-\hat{p}}{\sqrt{p(1-p) / n}}$ in the second one. Then we get that

$$
\begin{aligned}
\left|F(\hat{p})-\Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right)\right| & \leq \frac{C}{\sqrt{n p(1-p)}} ; \\
\left|1-F_{-}(\hat{p})-\Phi\left(\frac{p-\hat{p}}{\sqrt{p(1-p) / n}}\right)\right| & \leq \frac{C}{\sqrt{n p(1-p)}}
\end{aligned}
$$

Hence $F(\hat{p}) \geq \alpha / 2, F_{-}(\hat{p}) \leq 1-\alpha / 2$ implies that either

$$
p \geq \hat{p} \text { and } \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \alpha / 2
$$

or

$$
p \leq \hat{p} \text { and } \Phi\left(\frac{p-\hat{p}}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \alpha / 2 .
$$

Thus,

$$
\begin{aligned}
& \left\{0<p \leq \hat{p}: \Phi\left(\frac{p-\hat{p}}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \frac{\alpha}{2}\right\} \cup \\
& \left\{\hat{p} \leq p<1: \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \frac{\alpha}{2}\right\}
\end{aligned}
$$

is a valid $(1-\alpha)$-level confidence region for $p$. Since we have assumed that $p<1 / 2$, the above confidence region can be further shrunk. To summarize, we have the following theorem:

Theorem 4.4. Assume that $p<1 / 2$. Then the set

$$
\begin{align*}
& \left\{0<p \leq \hat{p} \wedge \frac{1}{2}: \Phi\left(\frac{p-\hat{p}}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \frac{\alpha}{2}\right\} \cup \\
& \left\{\hat{p} \leq p<\frac{1}{2}: \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \frac{\alpha}{2}\right\} \tag{4.6}
\end{align*}
$$

is a valid $(1-\alpha)$-level confidence region for $p$, for any finite sample $n$.

Unfortunately, we cannot derive a non-trivial confidence lower bound from (4.6) since any $0<p<1 / 2$ such that $C / \sqrt{n p(1-p)} \geq \alpha / 2$ is contained in this confidence region. Now we further relax (4.6) to develop a more explicit upper bound. In fact, (4.6) could be relaxed to

$$
\left\{0<p<1 / 2: \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right)+\frac{C}{\sqrt{n p(1-p)}} \geq \alpha / 2\right\} .
$$

In fact, for any $0 \leq \lambda \leq 1-\frac{4 C}{\sqrt{n} \alpha}$, we have that

$$
0<p<1 / 2, \frac{C}{\sqrt{n p(1-p)}} \geq(1-\lambda) \alpha / 2 \Rightarrow 0<p \leq \frac{1-\sqrt{1-\frac{16 C^{2}}{n(1-\lambda)^{2} \alpha^{2}}}}{2}
$$

and

$$
0<p<1 / 2, \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p) / n}}\right) \geq \lambda \alpha / 2 \Rightarrow 0<p \leq \frac{1+\frac{2 n \hat{p}}{z_{\lambda \alpha / 2}^{2}}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}}{2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right)}
$$

Therefore, we get that

$$
0<p \leq\left(\frac{1-\sqrt{1-\frac{16 C^{2}}{n(1-\lambda)^{2} \alpha^{2}}}}{2}\right) \vee\left(\frac{1+\frac{2 n \hat{p}}{z_{\lambda \alpha / 2}^{2}}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}}{2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right)}\right)
$$

is a $(1-\alpha)$-level CI. For simplicity, we denote the two parts as $U_{1}$ and $U_{2}$ respectively. One may note that $\lambda$ is not necessarily deterministic. Instead, it can be dependent on the data as long as it
stays within the interval $\left[0,1-\frac{4 C}{\sqrt{n} \alpha}\right]$. In fact, we may choose $\lambda$ carefully such that $U_{1} \leq U_{2}$ is guaranteed for sufficiently large $n$. Specifically, the following theorem proposes another valid CI.

Theorem 4.5. Assume that $p<1 / 2$. Let

$$
\lambda=1-\frac{2 \tilde{C}}{\sqrt{n} \alpha}
$$

where

$$
\tilde{C}=\left(\frac{C}{\sqrt{\hat{p}(1-\hat{p})}}\right) \wedge\left(\frac{u \sqrt{n} \alpha}{2}\right) .
$$

Here, $u<1$ is any constant such that $\frac{4 C^{2}}{u^{2} \alpha^{2}}<z_{(1-u) \alpha / 2}^{2}$. In the case that $\hat{p}=0$ or 1 , naturally we set $\tilde{C}=u \sqrt{n} \alpha / 2$. Then there exists $N_{0}$, which does not depend on $p$ and $\hat{p}$, such that for any $n>N_{0}$,

$$
\hat{p}_{u}^{B E}=\frac{1+\frac{2 n \hat{p}}{z_{\lambda \alpha / 2}^{2}}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}}{2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right)}, \hat{p}_{l}^{B E}=0
$$

is a valid $(1-\alpha)$-level CI for $p$. In particular, $N_{0}$ can be chosen as

$$
\left(\frac{4 C}{u \alpha}\right)^{2} \vee \frac{12 z_{(1-u) \alpha / 2}^{2} C^{2}}{z_{(1-u) \alpha / 2}^{2} u^{2} \alpha^{2}-4 C^{2}}
$$

Remark 4.3. Actually, $\hat{p}_{u}^{B E}$ itself is a valid $(1-\alpha / 2)$-level confidence upper bound for $p$. The series of relaxations makes this CI more and more conservative, but we will show that it still has similar scale with $\hat{p}_{u}^{C L T}$ and $\hat{p}_{u}^{\text {Wilson }}$.

Next we will show that $\hat{p}_{u}^{B E}-\hat{p}_{u}^{C L T}$ is bounded by order $1 / n$. In other words, though $\hat{p}_{u}^{C L T}$ has undesirable coverage probability in the rare-event setting, it is not "too far" from a valid upper bound. The following theorem states this result.

Theorem 4.6. Assume that $p<1 / 2 . \hat{p}_{u}^{B E}$ is as defined in Theorem 4.5. Then there is a constant
$C_{0}$ which does not depend on $p$ and $\hat{p}$ such that

$$
\left|\hat{p}_{u}^{B E}-\hat{p}_{u}^{C L T}\right| \leq C_{0} / n .
$$

Note that the bound in Theorem 4.6 can be rephrased as $\left|\hat{p}_{u}^{B E}-\hat{p}_{u}^{C L T}\right| \leq C_{0} \hat{p} / \hat{n}$. In other words, $\hat{p}_{u}^{B E}$ differ from $\hat{p}_{u}^{C L T}$ by a magnitude that is of higher order than the half-width of $\hat{p}_{u}^{C L T}$ (i.e., $\hat{p}_{u}^{C L T}-\hat{p}$ ) in terms of $\hat{n}$, while all quantities scale with $\hat{p}$ in a similar manner. Compared to Theorem 4.3, we see in Theorem 4.6 that $\hat{p}_{u}^{B E}$ is substantially tighter than $\hat{p}_{u}^{\text {Chernoff }}$ when $\hat{n}$ increases, although due to the implicit constant $C_{0}$ it may not be the case for small $\hat{n}$.

### 4.4 Confidence Intervals under Targeted Stopping

Now we consider experiments where we keep sampling until we get $n_{0}$ (say, 30) successes. Our goal is to construct $(1-\alpha)$-level CIs for $p$ using similar methods to Section 4.3 and then carry out the analysis on the difference from the point estimate and the CLT interval. Under this setting, the sample size $N$ is a random variable. More specifically, $N=N_{1}+\cdots+N_{n_{0}}$ where $N_{1}, \cdots, N_{n_{0}}$ are i.i.d. Geometric $(p)$ random variables, or equivalently, $N-n_{0} \sim N B\left(n_{0}, p\right)$. Note that $N \geq n_{0}$.

### 4.4.1 CLT Confidence Interval and Wilson's Confidence Interval

Under the targeted stopping setting, we could still use the CLT CI and the Wilson's CI with $\hat{p}=n_{0} / N$. More specifically, the CLT CI is

$$
\begin{aligned}
& \hat{p}_{u, n_{0}}^{C L T}=\frac{n_{0}}{N}+z_{1-\alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}} ; \\
& \hat{p}_{l, n_{0}}^{C L T}=\frac{n_{0}}{N}-z_{1-\alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}} .
\end{aligned}
$$

The Wilson's CI is

$$
\begin{aligned}
& \hat{p}_{u}^{\text {Wilson }}=\frac{1+\frac{2 n_{0}}{z_{1-\alpha / 2}^{2}}+\sqrt{1+\frac{4 n_{0}\left(N-n_{0}\right)}{z_{1-\alpha / 2}^{2} N}}}{2\left(1+\frac{N}{z_{1-\alpha / 2}^{2}}\right)} ; \\
& \hat{p}_{l}^{\text {Wilson }}=\frac{1+\frac{2 n_{0}}{z_{1-\alpha / 2}^{2}}-\sqrt{1+\frac{4 n_{0}\left(N-n_{0}\right)}{z_{1-\alpha / 2}^{2} N}}}{2\left(1+\frac{N}{z_{1-\alpha / 2}^{2}}\right)} .
\end{aligned}
$$

Clearly, we still have that $\hat{p}_{u, n_{0}}^{C L T}-\hat{p}=\hat{p}-\hat{p}_{l, n_{0}}^{C L T}=O\left(\hat{p} / \sqrt{n_{0}}\right)=O\left(\sqrt{n_{0}} / N\right)$ and that $\mid \hat{p}_{u, n_{0}}^{\text {Wilson }}-$ $\hat{p}_{u, n_{0}}^{C L T}\left|=O(1 / N),\left|\hat{p}_{l, n_{0}}^{\text {Wilson }}-\hat{p}_{l, n_{0}}^{C L T}\right|=O(1 / N)\right.$.

### 4.4.2 Valid Confidence Intervals

Now we define $F_{N}(x)=P(N \leq x)$ and $F_{N-}(x)=P(N<x)$. Similar to Section 4.3, we will argue that the following set is a valid $(1-\alpha)$-level confidence regon for $p$ :

$$
\begin{equation*}
\left\{0<p<1: F_{N}(N) \geq \alpha / 2, F_{N-}(N) \leq 1-\alpha / 2\right\} . \tag{4.7}
\end{equation*}
$$

Indeed, for any $\alpha \in(0,1)$, there exist $1 \leq k, l<\infty$ such that $F_{N}(k-1)<\alpha / 2 \leq F_{N}(k)$ and $F_{N-}(l) \leq 1-\alpha / 2<F_{N-}(l+1)$. Then

$$
\begin{aligned}
P\left(F_{N}(N)<\alpha / 2 \text { or } F_{N-}(N)>1-\alpha / 2\right) & \leq P\left(F_{N}(N)<\alpha / 2\right)+P\left(F_{N-}(N)>1-\alpha / 2\right) \\
& =P(N \leq k-1)+P(N \geq l+1) \\
& =F_{N}(k-1)+1-F_{N-}(l+1)<\alpha .
\end{aligned}
$$

By definition, the set (4.7) is a valid $(1-\alpha)$-level confidence region.

## Exact CI

Similar to the standard setting, we may directly solve (4.7). That is, we set $\hat{p}_{u, n_{0}}^{E x a c t}$ and $\hat{p}_{l, n_{0}}^{\text {Exact }}$ as the solutions of

$$
\begin{gathered}
\sum_{k=0}^{N-n_{0}-1}\binom{k+n_{0}-1}{n_{0}-1}(1-p)^{k} p^{n_{0}}=1-\alpha / 2, \\
\sum_{k=0}^{N-n_{0}}\binom{k+n_{0}-1}{n_{0}-1}(1-p)^{k} p^{n_{0}}=\alpha / 2
\end{gathered}
$$

except that $\hat{p}_{u, n_{0}}^{\text {Exact }}=1$ if $N=n_{0}$. While the interval is easy to compute numerically, it is not easy to analyze. Similar to the standard setting, we will relax the confidence region (4.7) to construct valid CIs respectively via inverting a concentration inequality and inverting the $\mathrm{B}-\mathrm{E}$ theorem.

## Chernoff CI

First, we propose the Chernoff CI similar to the one in the standard setting. By Markov's inequality, we get that

$$
P(N \geq x) \leq e^{-t x} E\left(e^{t N}\right)=e^{-t x}\left(\frac{p e^{t}}{1-(1-p) e^{t}}\right)^{n_{0}}, 0<t<-\log (1-p)
$$

Then for $x>n_{0} / p$,

$$
P(N \geq x) \leq \min _{0<t<-\log (1-p)} e^{-t x}\left(\frac{p e^{t}}{1-(1-p) e^{t}}\right)^{n_{0}}=\frac{(1-p)^{x-n_{0}} x^{x} p^{n_{0}}}{\left(x-n_{0}\right)^{x-n_{0}} n_{0}^{n_{0}}}
$$

Similarly,

$$
P(N \leq x) \leq e^{t x} E\left(e^{-t N}\right)=e^{t x}\left(\frac{p e^{-t}}{1-(1-p) e^{-t}}\right)^{n_{0}}, t>0
$$

and thus for $0<x<n_{0} / p$,

$$
P(N \leq x) \leq \min _{t>0} e^{t x}\left(\frac{p e^{-t}}{1-(1-p) e^{-t}}\right)^{n_{0}}=\frac{(1-p)^{x-n_{0}} x^{x} p^{n_{0}}}{\left(x-n_{0}\right)^{x-n_{0}} n_{0}^{n_{0}}}
$$

Therefore, $F_{N}(N) \geq \alpha / 2, F_{N-}(N) \leq 1-\alpha / 2$ implies that either

$$
N \geq n_{0} / p \text { and } \frac{(1-p)^{N-n_{0}} N^{N} p^{n_{0}}}{\left(N-n_{0}\right)^{N-n_{0}} n_{0}^{n_{0}}} \geq \alpha / 2
$$

or

$$
N \leq n_{0} / p \text { and } \frac{(1-p)^{N-n_{0}} N^{N} p^{n_{0}}}{\left(N-n_{0}\right)^{N-n_{0}} n_{0}^{n_{0}}} \geq \alpha / 2
$$

Finally we get that

$$
\left\{0<p<1: \frac{(1-p)^{N-n_{0}} N^{N} p^{n_{0}}}{\left(N-n_{0}\right)^{N-n_{0}} n_{0}^{n_{0}}} \geq \alpha / 2\right\}
$$

is a valid ( $1-\alpha$ )-level confidence region for $p$ under the targeted stopping setting. After simplification, we summarize our result with the folowing theorem:

Theorem 4.7. Suppose that we keep sampling from Bernoulli( $p$ ) until we get $n_{0}$ successes and the sample size is denoted by $N$. Then

$$
\begin{equation*}
\left\{0<p<1: p^{n_{0}}(1-p)^{N-n_{0}} \geq \frac{\alpha}{2}\left(\frac{n_{0}}{N}\right)^{n_{0}}\left(1-\frac{n_{0}}{N}\right)^{N-n_{0}}\right\} \tag{4.8}
\end{equation*}
$$

is a valid $(1-\alpha)$-level confidence region for $p$.

Remark 4.4. It is easy to verify that $f(p)=p^{n_{0}}(1-p)^{N-n_{0}}-\frac{\alpha}{2}\left(n_{0} / N\right)^{n_{0}}\left(1-n_{0} / N\right)^{N-n_{0}}$ is increasing in $\left[0, n_{0} / N\right]$ and decreasing in $\left[n_{0} / N, 1\right]$. Moreover, we observe that $f(0)=f(1)<0$ and $f\left(n_{0} / N\right)>0$. Thus (4.8) is actually an interval and numerically, we could use the bisection method to compute the bounds.

## Berry-Esseen CI

Now we apply the Berry-Esseen theorem again. We still assume that $p<1 / 2$ is known as a priori. By the theorem, we get that

$$
\begin{align*}
& \left|P\left(\frac{N-n_{0} / p}{\sqrt{n_{0}(1-p) / p^{2}}} \leq x\right)-\Phi(x)\right| \leq \frac{C \rho_{N}}{\sigma_{N}^{3} \sqrt{n_{0}}}  \tag{4.9}\\
& \left|P\left(\frac{n_{0} / p-N}{\sqrt{n_{0}(1-p) / p^{2}}} \leq x\right)-\Phi(x)\right| \leq \frac{C \rho_{N}}{\sigma_{N}^{3} \sqrt{n_{0}}} \tag{4.10}
\end{align*}
$$

where $\sigma_{N}^{2}=E\left(N_{i}-1 / p\right)^{2}=(1-p) / p^{2}$ and $\rho_{N}=E\left|N_{i}-1 / p\right|^{3}$.
We need to deal with $\rho_{N}$ first. In fact, we know that

$$
p^{3} \rho_{N}=p^{3} E\left|N_{i}-\frac{1}{p}\right|^{3}=E\left|p N_{i}-1\right|^{3} \leq 1+3 p E\left(N_{i}\right)+3 p^{2} E\left(N_{i}^{2}\right)+p^{3} E\left(N_{i}^{3}\right)
$$

Since $N_{i} \sim \operatorname{Geometric}(p)$, we know that

$$
E\left(N_{i}\right)=\frac{1}{p}, E\left(N_{i}^{2}\right)=\frac{2-p}{p^{2}}, E\left(N_{i}^{3}\right)=\frac{p^{2}-6 p+6}{p^{3}}
$$

and thus

$$
p^{3} \rho_{N} \leq p^{2}-6 p+6+3(2-p)+3+1=p^{2}-12 p+16 \leq 16 .
$$

Hence,

$$
\frac{C \rho_{N}}{\sigma_{N}^{3} \sqrt{n_{0}}}=\frac{C p^{3} \rho_{N}}{(1-p)^{\frac{3}{2}} \sqrt{n_{0}}} \leq \frac{C^{\prime}}{(1-p)^{\frac{3}{2}} \sqrt{n_{0}}}
$$

where $C^{\prime}=16 C$ is an absolute constant.
By setting $x=\sqrt{\frac{p^{2}}{n_{0}(1-p)}}\left(N-\frac{n_{0}}{p}\right)$ in (4.9) and $x=\sqrt{\frac{p^{2}}{n_{0}(1-p)}}\left(\frac{n_{0}}{p}-N\right)$ in (4.10), we get that

$$
\begin{aligned}
\left|F_{N}(N)-\Phi\left(\frac{N p-n_{0}}{\sqrt{n_{0}(1-p)}}\right)\right| & \leq \frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \\
\left|1-F_{N-}(N)-\Phi\left(\frac{n_{0}-N p}{\sqrt{n_{0}(1-p)}}\right)\right| & \leq \frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}}
\end{aligned}
$$

Hence $F_{N}(N) \geq \alpha / 2, F_{N-}(N) \leq 1-\alpha / 2$ implies that either

$$
p \geq n_{0} / N \text { and } \Phi\left(\frac{n_{0}-N p}{\sqrt{n_{0}(1-p)}}\right)+\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq \alpha / 2
$$

or

$$
p \leq n_{0} / N \text { and } \Phi\left(\frac{N p-n_{0}}{\sqrt{n_{0}(1-p)}}\right)+\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq \alpha / 2 .
$$

Thus, we develop a valid confidence region under this particular setting, which is similar to the one in Section 4.3:

Theorem 4.8. Suppose that we keep sampling from Bernoulli( $p$ ) until we get $n_{0}$ successes and the sample size is denoted by $N$. Assume that $p<1 / 2$. Then

$$
\begin{align*}
& \left\{0<p \leq \frac{n_{0}}{N} \wedge \frac{1}{2}: \Phi\left(\frac{N p-n_{0}}{\sqrt{n_{0}(1-p)}}\right)+\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq \frac{\alpha}{2}\right\} \cup \\
& \left\{\frac{n_{0}}{N} \leq p<\frac{1}{2}: \Phi\left(\frac{n_{0}-N p}{\sqrt{n_{0}(1-p)}}\right)+\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq \frac{\alpha}{2}\right\} \tag{4.11}
\end{align*}
$$

is a valid $(1-\alpha)$-level confidence region for $p$. Here, $C^{\prime}$ is a universal constant. In particular, one may pick $C^{\prime}=16 C$ where $C$ is the constant in the Berry-Esseen theorem.

Remark 4.5. Note that when $n_{0}$ is not large enough, we have $C^{\prime} / \sqrt{n_{0}(1-p)^{3}} \geq \alpha / 2$ anyway. That is to say, this confidence region is not really practical. However, it could still provide an insight on how close are the CLT or the Wilson's intervals to a valid one.

Similar to Section 4.3, the confidence region (4.11) could be further relaxed. However, unlike in the standard setting, now the error term $C^{\prime} / \sqrt{n_{0}(1-p)^{3}}$ could be well controlled for tiny $p$ and as a result, we are able to get a non-trivial lower bound in this case. More concretely, for any $0<\lambda<1$, we have that

$$
0<p<1, \frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \geq(1-\lambda) \alpha / 2 \Rightarrow p \geq 1-\left(\frac{4 C^{\prime}}{n_{0}(1-\lambda)^{2} \alpha^{2}}\right)^{1 / 3}
$$

If we could find $0<\lambda<1$ such that

$$
\left(\frac{4 C^{\prime 2}}{n_{0}(1-\lambda)^{2} \alpha^{2}}\right)^{1 / 3}=\frac{1}{2}
$$

then for any $0<p<1 / 2$, we have that

$$
\frac{C^{\prime}}{\sqrt{n_{0}(1-p)^{3}}} \leq(1-\lambda) \alpha / 2 .
$$

As a result, any $p$ in (4.11) must satisfy that

$$
0<p \leq \frac{n_{0}}{N} \wedge \frac{1}{2}, \Phi\left(\frac{N p-n_{0}}{\sqrt{n_{0}(1-p)}}\right) \geq \lambda \alpha / 2 \text { or } \frac{n_{0}}{N} \leq p<\frac{1}{2}, \Phi\left(\frac{n_{0}-N p}{\sqrt{n_{0}(1-p)}}\right) \geq \lambda \alpha / 2
$$

After simplification, we get that

$$
\begin{aligned}
& 0<p \leq \frac{n_{0}}{N}, \Phi\left(\frac{N p-n_{0}}{\sqrt{n_{0}(1-p)}}\right) \geq \lambda \alpha / 2 \\
\Rightarrow & p \geq \frac{2 N n_{0}-z_{\lambda \alpha / 2}^{2} n_{0}-\sqrt{4 z_{\lambda \alpha / 2}^{2} N n_{0}\left(N-n_{0}\right)+z_{\lambda \alpha / 2}^{4} n_{0}^{2}}}{2 N^{2}}, \\
& \frac{n_{0}}{N} \leq p<1, \Phi\left(\frac{n_{0}-N p}{\sqrt{n_{0}(1-p)}}\right) \geq \lambda \alpha / 2 \\
\Rightarrow & p \leq \frac{2 N n_{0}-z_{\lambda \alpha / 2}^{2} n_{0}+\sqrt{4 z_{\lambda \alpha / 2}^{2} N n_{0}\left(N-n_{0}\right)+z_{\lambda \alpha / 2}^{4} n_{0}^{2}}}{2 N^{2}} .
\end{aligned}
$$

Thus (4.11) could be relaxed into a valid $(1-\alpha)$-level CI, which is defined more rigorously in the following theorem:

Theorem 4.9. Suppose that we keep sampling from Bernoulli( $p$ ) until we get $n_{0}$ successes and the sample size is denoted by $N$. Assume that $p<1 / 2$. Let

$$
\lambda=1-\frac{4 \sqrt{2} C^{\prime}}{\sqrt{n_{0}} \alpha} .
$$

Then for any $n_{0}>32 C^{\prime 2} / \alpha^{2}$, we have that

$$
\begin{aligned}
& \hat{p}_{u, n_{0}}^{B E}=\frac{2 N n_{0}-z_{\lambda \alpha / 2}^{2} n_{0}+\sqrt{4 z_{\lambda \alpha / 2}^{2} N n_{0}\left(N-n_{0}\right)+z_{\lambda \alpha / 2}^{4} n_{0}^{2}}}{2 N^{2}}, \\
& \hat{p}_{l, n_{0}}^{B E}=\frac{2 N n_{0}-z_{\lambda \alpha / 2}^{2} n_{0}-\sqrt{4 z_{\lambda \alpha / 2}^{2} N n_{0}\left(N-n_{0}\right)+z_{\lambda \alpha / 2}^{4} n_{0}^{2}}}{2 N^{2}}
\end{aligned}
$$

is a valid $(1-\alpha)$-level CI for $p$. Here, $C^{\prime}$ is the same as in Theorem 4.8.

Finally, like in the standard setting, we will compare the difference between the B-E CI and the CLT CI.

Theorem 4.10. Assume that $p<1 / 2 . \hat{p}_{u, n_{0}}^{B E}$ and $\hat{p}_{l, n_{0}}^{B E}$ are as defined in Theorem 4.9. Then there is a constant $C_{0}^{\prime}$ which does not depend on $p$ and $N$ such that

$$
\begin{aligned}
& \hat{p}_{u, n_{0}}^{B E}-\hat{p}_{u, n_{0}}^{C L T} \leq C_{0}^{\prime} / N, \\
& \hat{p}_{l, n_{0}}^{C L T}-\hat{p}_{l, n_{0}}^{B E} \leq C_{0}^{\prime} / N .
\end{aligned}
$$

Therefore, under the targeted stopping setting, we could justify that the CLT CI is not too far from a valid one in terms of both upper bound and lower bound.

### 4.5 Numerical Experiments

To visualize the differences among the CIs, we perform some numerical experiments.

### 4.5.1 Experiments under the Standard Setting

The true value is chosen as $p=1 e-6$. For $n=5 / p, 10 / p, 30 / p, 50 / p, 100 / p$, we respectively do 1000 simulations and calculate the CIs with $\alpha=0.05$. Figure 4.2 and Table 4.2 respectively present the mean values and the coverage probabilities of the five CIs covered in this chapter.


| $n p$ | CLT | Wilson | Exact | Chernoff | BE |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 0.858 | 0.969 | 0.986 | 0.999 | 1.0 |
| 10 | 0.923 | 0.959 | 0.971 | 0.999 | 1.0 |
| 30 | 0.931 | 0.943 | 0.951 | 0.994 | 1.0 |
| 50 | 0.950 | 0.948 | 0.953 | 0.996 | 1.0 |
| 100 | 0.945 | 0.945 | 0.951 | 0.995 | 1.0 |

Table 4.2: Coverage probabilities of the CIs.
Figure 4.2: Mean values of the CIs.

As we analyzed, when $n p$ is large, actually the CIs scale similarly, except that B-E fails to give a non-zero lower bound. While the CLT interval is closest to the truth in terms of the mean value of the upper bound, it is not reliable especially when $n p$ is small. For instance, when $n p=5$, its coverage probability is only 0.858 , which is much lower than the nominal confidence level 0.95 . The Wilson's and the exact CIs are quite similar, especially for the upper bound. However, we notice that the Wilson's bound sometimes fails to achieve the nominal confidence level, but the error in the coverage probability is acceptable to some extent. The Chernoff and the B-E CIs are conservative as expected. We would like to point out that though the B-E upper bound seems to be much larger than the Chernoff one, it decays much faster as $n p$ increases, which coincides with our theorems.

### 4.5.2 Experiments under the Targeted Stopping Setting

Now we simulate the targeted stopping setting. We still set $p=1 e-6$. For $n_{0}=5,10,30,50,100$, we respectively do 1000 simulations and calculate the CIs with $\alpha=0.05$. Figure 4.3 and Table 4.3 respectively present the mean values and the coverage probabilities of the CIs. Note that we do not include the B-E CI since it is trivial due to small $n_{0}$ as aforementioned.

The results are basically similar to the standard setting. We conclude that if we could observe enough (say, 30) successes, then the CLT CI is indeed reliable.


Figure 4.3: Mean values of the CIs.

| $n_{0}$ | CLT | Wilson | Exact | Chernoff |
| :---: | :---: | :---: | :---: | :---: |
| 5 | 0.943 | 0.916 | 0.944 | 0.984 |
| 10 | 0.955 | 0.926 | 0.936 | 0.994 |
| 30 | 0.964 | 0.961 | 0.965 | 0.993 |
| 50 | 0.942 | 0.943 | 0.943 | 0.989 |
| 100 | 0.950 | 0.950 | 0.952 | 0.991 |

Table 4.3: Coverage probabilities of the CIs.

### 4.6 Conclusion

In this chapter, we focus on constructing CIs for some rare-event probability with Bernoulli data. We respectively consider two settings: the standard setting where the sample size is fixed, and the targeted stopping setting where the number of successes is fixed. Under each setting, we first review the commonly used CLT, Wilson's and exact CIs. It is known that the CLT and the Wilson's CIs are not necessarily valid in the sense that the actural coverage probability can be lower than the nominal confidence level, and the exact CI is valid yet hard to analyze, which motivates us to derive other valid CIs with more explicit expressions. More specifically, we further relax the exact confidence region via inverting a concentration inequality and inverting the BerryEssen theorem to get the Chernoff and the B-E CIs respectively. Table 4.1 in Section 4.2 is a comprehensive summary of our findings.

Now we could briefly answer the five motivating questions:

A1. From our analysis, the difference of the CLT upper bound from the B-E upper bound is of order $\hat{p} / \hat{n}$, which is of higher order in $\hat{n}$ than the difference from $\hat{p}$. Hence, when $\hat{n}$ is large, the CLT upper bound is not too far from a valid one, and hence it is relatively reliable. However, temporarily we do not have such guarantees for the lower bound.

A2. There are different ways to construct valid CIs. Among them, the exact CI is tight and can be computed easily.

A3. We have analyzed the "half-width" of the Chernoff CI and the B-E upper bound. All of them are of order $O(\hat{p} / \sqrt{\hat{n}})$, which is a reasonable magnitude.

A4. Due to the order of the "half-width" of the Chernoff CI, we can conclude that $\hat{p}$ can be viewed as relatively efficient conditional on $\hat{n}$ is sufficiently large (see Section 4.3 for more details).

A5. In the targeted stopping setting, we have similar results. Moreover, in this case we also have guarantees for the CLT lower bound.

Overall, we recommend the exact CI in the situations where absolute safety is critical while in other general applications, we could suggest using the Wilson's CI given its excellent empirical performance. Moreover, when the number of successes is large, either in the standard setting or the targeted stopping setting, we have justified that the CLT CI and the point estimate are also reliable by comparisons with the valid Chernoff and B-E CIs. Though the latter two intervals are extremely conservative, they provide us useful insights in understanding the errors. The remaining problem is to obtain the guarantees for the CLT lower bound in the standard setting, which will be left for future work.

### 4.7 Supplementary A: Proofs

Proof of Theorem 4.1. We get directly from the formula of $\hat{p}_{u}^{\text {Wilson }}$ that

$$
\begin{aligned}
\left|\hat{p}_{u}^{W i l s o n}-\hat{p}_{u}^{C L T}\right| & =\left|\frac{1-2 \hat{p}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{1-\alpha / 2}^{2}}}-2\left(1+\frac{n}{z_{1-\alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{2\left(1+\frac{n}{z_{1-\alpha / 2}^{2}}\right)}\right| \\
& \leq \frac{|1-2 \hat{p}|+\left|2 z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right|+\left|\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{1-\alpha / 2}^{2}}}-\frac{2 n}{z_{1-\alpha / 2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right|}{2 n / z_{1-\alpha / 2}^{2}} .
\end{aligned}
$$

We have that $|1-2 \hat{p}| \leq 1,\left|2 z_{1-\alpha / 2} \sqrt{\hat{p}(1-\hat{p}) / n}\right| \leq z_{1-\alpha / 2} / \sqrt{n}$ and

$$
\left|\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{1-\alpha / 2}^{2}}}-\frac{2 n}{z_{1-\alpha / 2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right|=\left|\frac{1}{\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{1-\alpha / 2}^{2}}}+\frac{2 n}{z_{1-\alpha / 2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\right| \leq 1,
$$

which concludes the proof for the upper bounds. The proof for the lower bound is almost the same.

Proof of Theorem 4.3. We have that

$$
\begin{aligned}
\hat{p}_{u}^{\text {Chernoff }}-\hat{p}_{u}^{C L T} & =\sqrt{\frac{2 \log (2 / \alpha) \hat{p}}{n}+\frac{(\log (2 / \alpha))^{2}}{n^{2}}}+\frac{\log (2 / \alpha)}{n}-z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
& \geq\left(\sqrt{2 \log (2 / \alpha)}-z_{1-\alpha / 2}\right) \sqrt{\frac{\hat{p}}{n}}+\frac{\log (2 / \alpha)}{n} .
\end{aligned}
$$

Similarly we could prove the inequality for the lower bounds.

Proof of Theorem 4.5. Following our derivations, it suffices to show that for the given $N_{0}$ and any $n>N_{0}$, we have that $0 \leq \lambda \leq 1-\frac{4 C}{\sqrt{n} \alpha}$ and $U_{1} \leq U_{2}$. Obviously, $\frac{2 \tilde{C}}{\sqrt{n} \alpha} \leq u<1$, so $\lambda>0$. On the other side, $\lambda \leq 1-\frac{4 C}{\sqrt{n} \alpha}$ holds since $n \geq\left(\frac{4 C}{u \alpha}\right)^{2}$. Now we prove that $U_{1} \leq U_{2}$ for $n>N_{0}$. Indeed, if $\tilde{C}=C / \sqrt{\hat{p}(1-\hat{p})}$, then $U_{1}=\hat{p} \wedge(1-\hat{p}) \leq \hat{p}$ and we know that $U_{2} \geq \hat{p}$, so $U_{1} \leq U_{2}$. In the other case that $\tilde{C}=u \sqrt{n} \alpha / 2$, we have that

$$
U_{1}=\frac{1-\sqrt{1-\frac{16 C^{2}}{n u^{2} \alpha^{2}}}}{2}=\frac{\frac{16 C^{2}}{n u^{2} \alpha^{2}}}{2\left(1+\sqrt{1-\frac{16 C^{2}}{n u^{2} \alpha^{2}}}\right)} \leq \frac{\frac{16 C^{2}}{n u^{2} \alpha^{2}}}{2\left(1+1-\frac{16 C^{2}}{n u^{2} \alpha^{2}}\right)}=\frac{4 C^{2}}{n u^{2} \alpha^{2}-8 C^{2}}
$$

and

$$
U_{2} \geq \frac{1}{1+\frac{n}{z_{(1-u) \alpha / 2}^{2}}}=\frac{z_{(1-u) \alpha / 2}^{2}}{z_{(1-u) \alpha / 2}^{2}+n}
$$

Since $u$ is chosen such that $\frac{4 C^{2}}{u^{2} \alpha^{2}}<z_{(1-u) \alpha / 2}^{2}$ and

$$
n \geq \frac{12 z_{(1-u) \alpha / 2}^{2} C^{2}}{z_{(1-u) \alpha / 2}^{2} u^{2} \alpha^{2}-4 C^{2}}
$$

we get that

$$
\frac{4 C^{2}}{n u^{2} \alpha^{2}-8 C^{2}} \leq \frac{z_{(1-u) \alpha / 2}^{2}}{z_{(1-u) \alpha / 2}^{2}+n}
$$

and hence $U_{1} \leq U_{2}$. Note that as $u \uparrow 1, \frac{4 C^{2}}{u^{2} \alpha^{2}} \rightarrow \frac{4 C^{2}}{\alpha^{2}}$ while $z_{(1-u) \alpha / 2}^{2} \rightarrow \infty$, and thus such $u$ exists.

Proof of Theorem 4.6. We have that

$$
\hat{p}_{u}^{B E}-\hat{p}_{u}^{C L T}=\frac{1-2 \hat{p}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}-2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right)} .
$$

We first deal with

$$
\begin{aligned}
& \sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}-2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
= & \frac{\left(1-\frac{8 z_{1-\alpha / 2}^{2} \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}\right)-\frac{4 z_{1-\alpha / 2}^{2} \hat{p}(1-\hat{p})}{n}+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}\left(1-\frac{z_{1-\alpha / 2}^{2}}{z_{\lambda \alpha / 2}^{2}}\right)}{\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}+2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} .
\end{aligned}
$$

The denominator satisfies that

$$
\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}+2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \geq\left(\frac{2 \sqrt{n \hat{p}(1-\hat{p})}\left(z_{1-\lambda \alpha / 2}+z_{1-\alpha / 2}\right)}{z_{\lambda \alpha / 2}^{2}}\right) \vee 1
$$

Note that $\left(z_{1-\lambda \alpha / 2}+z_{1-\alpha / 2}\right) / z_{\lambda \alpha / 2}^{2}$ increases with the value of $\lambda$. Since $\lambda \geq 1-u>0$, we can find
a constant $C_{1}$ such that

$$
\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}+2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \geq\left(C_{1} \sqrt{n \hat{p}(1-\hat{p})}\right) \vee 1
$$

Then we deal with the numerator. We know that $z_{\alpha / 2}=\Phi^{-1}(\alpha / 2)$ and $z_{\lambda \alpha / 2}=\Phi^{-1}(\lambda \alpha / 2)$. By Taylor expansion, we have that

$$
\frac{1}{z_{\lambda \alpha / 2}^{2}}=\frac{1}{z_{\alpha / 2}^{2}}-\frac{2 \sqrt{2 \pi}}{z_{\alpha / 2}^{3}} e^{\frac{z_{\alpha / 2}^{2}}{2}}(\lambda-1) \alpha / 2+r(\lambda)
$$

Here, $r(\lambda)$ is continuous in $\lambda$ and $r(\lambda) /(1-\lambda) \rightarrow 0$ as $\lambda \uparrow 1$. We also know that $1-\lambda \leq u$, and thus $|r(\lambda) /(1-\lambda)|=|(\sqrt{n} \alpha r(\lambda)) /(2 \tilde{C})|$ is bounded by a constant. Hence $|\sqrt{n \hat{p}(1-\hat{p})} r(\lambda)|$ is bounded by a constant. We have that

$$
1-\frac{z_{1-\alpha / 2}^{2}}{z_{\lambda \alpha / 2}^{2}}=\frac{2 \sqrt{2 \pi}}{z_{1-\alpha / 2}} e^{\frac{z_{1-\alpha / 2}^{2}}{2}} \frac{\tilde{C}}{\sqrt{n}}-z_{1-\alpha / 2}^{2} r(\lambda)
$$

Thus, the numerator satisfies that

$$
\begin{aligned}
& \left|\left(1-\frac{8 z_{1-\alpha / 2}^{2} \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}\right)-\frac{4 z_{1-\alpha / 2}^{2} \hat{p}(1-\hat{p})}{n}+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}\left(1-\frac{z_{1-\alpha / 2}^{2}}{z_{\lambda \alpha / 2}^{2}}\right)\right| \\
& \leq 1+8 \hat{p}(1-\hat{p})+\frac{4 z_{1-\alpha / 2}^{2} \hat{p}(1-\hat{p})}{n}+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}\left(\frac{2 \sqrt{2 \pi}}{z_{1-\alpha / 2}} e^{\frac{z_{1-\alpha / 2}^{2}}{2}} \frac{\tilde{C}}{\sqrt{n}}-z_{1-\alpha / 2}^{2} r(\lambda)\right) .
\end{aligned}
$$

Clearly, the first three terms divided by the denominator are bounded by some constants. Now we consider the fourth term. Since $|\sqrt{n \hat{p}(1-\hat{p})} r(\lambda)|$ is bounded, we can also get that the fourth term divided by the denominator is bounded by some universal constant.

Therefore, combining the above results, we know that

$$
\left|1-2 \hat{p}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha / 2}^{2}}}-2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) z_{1-\alpha / 2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right| \leq C_{2}
$$

where $C_{2}$ is a positive constant. We also have that

$$
2\left(1+\frac{n}{z_{\lambda \alpha / 2}^{2}}\right) \geq \frac{2 n}{z_{(1-u) \alpha / 2}^{2}}
$$

Hence the error term satisfies that

$$
\left|\frac{1-2 \hat{p}+\sqrt{1+\frac{4 n \hat{p}(1-\hat{p})}{z_{\lambda \alpha}^{2}}}-2\left(1+\frac{n}{z_{\lambda \alpha}^{2}}\right) z_{1-\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{2\left(1+\frac{n}{z_{\lambda \alpha}^{2}}\right)}\right| \leq \frac{C_{0}}{n}
$$

for some constant $C_{0}$. From the above derivations, we find that $C_{0}$ only depends on $\alpha$ and the choice of $u$.

Proof of Theorem 4.9. If $n_{0}>32 C^{2} / \alpha^{2}$, then we get that $0<\lambda<1$. The validness of the CI is justified by the derivations above the theorem.

Proof of Theorem 4.10. We have that

$$
\begin{aligned}
\hat{p}_{u, n_{0}}^{B E}-\hat{p}_{u, n_{0}}^{C L T} & =-\frac{z_{\lambda \alpha / 2}^{2} n_{0}}{2 N^{2}}+\sqrt{\frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}}+\frac{z_{\lambda \alpha / 2}^{4} n_{0}^{2}}{4 N^{4}}}-z_{1-\alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}} \\
& =\frac{\frac{\left(z_{\lambda \alpha / 2}^{2}-z_{1-\alpha / 2}^{2}\right) n_{0}\left(N-n_{0}\right)}{N^{3}}-\frac{n_{0} z_{\lambda \alpha / 2}^{2} z_{1-\alpha / 2}}{N^{2}} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}}}{\sqrt{\frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}}+\frac{z_{\lambda \alpha / 2}^{4} n_{0}^{2}}{4 N^{4}}}+\frac{z_{\lambda / / 2}^{2} n_{0}}{2 N^{2}}+z_{1-\alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}}}
\end{aligned}
$$

First, we have that the denominator

$$
\begin{aligned}
& \sqrt{\frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}}+\frac{z_{\lambda \alpha / 2}^{4} n_{0}^{2}}{4 N^{4}}+\frac{z_{\lambda \alpha / 2}^{2} n_{0}}{2 N^{2}}+z_{1-\alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}}} \\
\geq & \sqrt{\frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}}}=z_{1-\lambda \alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}} .
\end{aligned}
$$

Next, we have that the numerator

$$
\frac{\left(z_{\lambda \alpha / 2}^{2}-z_{1-\alpha / 2}^{2}\right) n_{0}\left(N-n_{0}\right)}{N^{3}}-\frac{n_{0} z_{\lambda \alpha / 2}^{2} z_{1-\alpha / 2}}{N^{2}} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}} \leq\left(1-\frac{z_{1-\alpha / 2}^{2}}{z_{\lambda \alpha / 2}^{2}}\right) \frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}} .
$$

As mentioned in the proof of Theorem 4.6, we have that

$$
\frac{1}{z_{\lambda \alpha / 2}^{2}}=\frac{1}{z_{\alpha / 2}^{2}}-\frac{2 \sqrt{2 \pi}}{z_{\alpha / 2}^{3}} e^{\frac{z_{\alpha / 2}^{2}}{2}}(\lambda-1) \alpha / 2+r(\lambda)
$$

and

$$
1-\frac{z_{1-\alpha / 2}^{2}}{z_{\lambda \alpha / 2}^{2}}=\frac{\sqrt{2 \pi}}{z_{1-\alpha / 2}} e^{\frac{z_{\alpha / 2}^{2}}{2}} \frac{C_{1}^{\prime}}{\sqrt{n_{0}}} \alpha-r(\lambda) z_{\alpha / 2}^{2}
$$

where $r(\lambda)$ is continuous in $\lambda$ and $r(\lambda) /(1-\lambda) \rightarrow 0$ as $\lambda \uparrow 1$. We know that $1-\lambda=C_{1}^{\prime} / \sqrt{n_{0}}$ for some constant $C_{1}^{\prime}>0$ from the choice of $\lambda$. For $n_{0}>32 C^{2} / \alpha^{2}, \lambda$ is bounded away from 0 , and hence $|r(\lambda) /(1-\lambda)|=\left|\sqrt{n_{0}} r(\lambda) / C_{1}^{\prime}\right|$ has a constant upper bound. Hence, there exists a constant $C_{2}^{\prime}>0$ such that

$$
\left(1-\frac{z_{1-\alpha / 2}^{2}}{z_{\lambda \alpha / 2}^{2}}\right) \frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}} \leq \frac{C_{2}^{\prime}}{\sqrt{n_{0}}} \frac{z_{\lambda \alpha / 2}^{2} n_{0}\left(N-n_{0}\right)}{N^{3}} .
$$

Combining the results, we get that

$$
\hat{p}_{u, n_{0}}^{B E}-\hat{p}_{u, n_{0}}^{C L T} \leq \frac{C_{2}^{\prime}}{\sqrt{n_{0}}} z_{1-\lambda \alpha / 2} \sqrt{\frac{n_{0}\left(N-n_{0}\right)}{N^{3}}} \leq \frac{C_{2}^{\prime} z_{1-\lambda \alpha / 2}}{N} \leq \frac{C_{3}^{\prime}}{N}
$$

where $C_{3}^{\prime}>0$ is a constant and we get the last inequality since $\lambda$ has a non-zero lower bound.
Now notice that

$$
\hat{p}_{l, n_{0}}^{C L T}-\hat{p}_{l, n_{0}}^{B E}=\hat{p}_{u, n_{0}}^{B E}-\hat{p}_{u, n_{0}}^{C L T}+\frac{z_{\lambda \alpha / 2}^{2} n_{0}}{N^{2}}
$$

and

$$
\frac{z_{\lambda \alpha / 2}^{2} n_{0}}{N^{2}} \leq \frac{z_{\lambda \alpha / 2}^{2}}{N}=O(1 / N) .
$$

Therefore, we could find a constant $C_{0}^{\prime}$ such that the theorem holds.

# Chapter 5: Rare-Event Simulation without Variance Reduction: An Extreme Value Theory Approach 

### 5.1 Introduction

A major goal of rare-event simulation is to estimate tiny probabilities that are triggered by rare but catastrophic events [22,23, 116]. This problem has been of wide interest to various application areas such as queueing systems [117, 31, 26, 27, 25, 30, 24, 29], communication networks [32], finance $[34,36,35]$ and insurance [38, 37]. In recent years, with the extensive development of machine learning and artificial intelligence, rare-event simulation is also applied to evaluate the robustness of machine learning predictors [14, 125] or quantify the risk of intelligent physical systems [18, 19, 17, 20, 21]. In using Monte Carlo (MC) to estimate rare-event probabilities, a main challenge is that, by its own nature, the target rare events seldom occur in the simulation experiments. Since sufficient hits on the target events are required to achieve meaningfully accurate estimation, this makes crude MC computationally costly as the required simulation size to attain enough accuracy becomes enormous.

To address the inefficiency of crude MC, a range of variance reduction techniques have been developed. These techniques aim to alter the naive sampling procedure of crude MC to improve the error per simulation run, thus attaining adequate accuracy with less number of samples. Among them, importance sampling (IS) $[51,53,23]$ is one of the most popular methods. The idea is to utilize a change of measure to amplify the frequency of target events in the simulation, and then to correct this bias with the likelihood ratio. Conditional Monte Carlo (CMC) [54, 63] uses the conditional probability of the target events on certain information as an unbiased estimator, which is especially useful in the heavy-tailed case where the classical exponential tilting IS technique cannot be applied. Multi-level splitting or subset simulation [87, 138, 139] chooses a sequence
of nested subsets, factorizes the target probability into the product of conditional probabilities and then estimates each conditional probability by generating samples from the corresponding conditional distribution. Despite demonstratably powerful in many problems, in order to attain good performances, the variance reduction techniques described above often rely on tractable problem structures that allow careful algorithmic design. Unfortunately, this requirement could be difficult or even impossible to meet in complex practical applications. Thus, the main goal of this chapter is to study an approach to improve upon the efficiency of crude MC in the absence of variance reduction schemes.

More specifically, we resort to extreme value theory (EVT) [135], which has been a prominent approach in extreme event analysis for real data. There, the main challenge is the scarcity of observations in the tail portion of the data that can directly infer the distributional extreme, and thus one needs to suitably extrapolate information from the "central" part of the distribution. One major approach, which we would borrow here, is the peak-over-threshold (POT) [137] method. This method is based on the Pickands-Balkema-de Haan theorem [145, 146] which states that, under suitable assumptions, the distribution function of the so-called excess loss above a threshold converges to the generalized Pareto distribution (GPD) as this threshold increases. In this sense, GPD is a justified model for tail data fitting. Analogically, we propose applying the POT method on the crude MC simulation data. That is, within an acceptable computational budget, we use the crude MC simulator to generate a simulation dataset, treat it as a real dataset, and fit the tail portion of the data with GPD. The target probability is then estimated with the fitted GPD. Although the POT method is well-established and this idea seems simple, to our best knowledge, POT has not been considered in the rare-event simulation context. As discussed above, the latter literature mainly focuses on developing algorithms to improve the efficiency of crude MC by utilizing tractable problem structures.

We apply this method to several numerical examples and compare the results with crude MC, including examples where variance reduction techniques such as IS cannot be applied easily. We find that, in a suitably wide parameter range, our POT estimator achieves smaller variance than
crude MC. Moreover, with limited simulation samples, while crude MC often outputs a trivial estimate of 0, POT can output an estimate of a roughly correct magnitude.

The rest of this chapter is organized as follows. Section 5.2 introduces relevant backgrounds about the challenge of crude MC and the theory of POT. Section 5.3 demonstrates the numerical results of several experiments. Finally, Section 5.4 concludes and discusses future work.

### 5.2 Background

### 5.2.1 Challenge of Crude MC

We consider a simulation model that outputs a random vector $X \in \mathbb{R}^{d}$ under probability measure $P$, and we are interested in estimating $p=P(X \in E)$ where $E \subset \mathbb{R}^{d}$ is a rare-event set, i.e., $p$ is a tiny number. Suppose we generate $n$ i.i.d. simulation samples $X_{1}, \ldots, X_{n}$. Then the crude MC estimator is simply $\hat{p}_{n}^{M C}=\frac{1}{n} \sum_{i=1}^{n} I\left(X_{i} \in E\right)$. In this rare-event setting, we would like to control the discrepancy between $\hat{p}_{n}^{M C}$ and $p$, relative to the magnitude of $p$ itself. In other words, we want to control the probability $P\left(\left|\hat{p}_{n}^{M C}-p\right|>\delta p\right)$ for some fixed $\delta<1$.

Note that the crude MC estimator is clearly unbiased. By Chebyshev's inequality, we have for any $\delta>0$,

$$
P\left(\left|\hat{p}_{n}^{M C}-p\right|>\delta p\right) \leq \frac{\operatorname{var}\left(\hat{p}_{n}^{M C}\right)}{\delta^{2} p^{2}}=\frac{\operatorname{var}(I(X \in E))}{n \delta^{2} p^{2}}=\frac{p(1-p)}{n \delta^{2} p^{2}} \leq \frac{1}{n \delta^{2} p} .
$$

Thus, $n \geq \frac{1}{\varepsilon \delta^{2} p}$ guarantees that $P\left(\left|\hat{p}_{n}^{M C}-p\right|>\delta p\right) \leq \varepsilon$ for any $\varepsilon>0$. This reveals that, when $p$ is tiny, the required simulation size which is reciprocal in $p$ can be enormous. Therefore, as widely known in the literature, crude MC is inefficient for rare-event simulation.

Moreover, we note that $\hat{p}_{n}^{M C}$ can only take values in $0,1 / n, \ldots,(n-1) / n, 1$. In particular, $P\left(\hat{p}_{n}^{M C}=0\right)=(1-p)^{n} \approx e^{-n p}$. Thus, when $n$ is not sufficiently large relative to $p$, there is a significant chance that the crude MC estimate is trivially 0.

As discussed in the introduction, we are interested in scenarios where crude MC is the only available simulation format. In this case, we study POT to obtain a better estimate from crude MC
without overwhelming the computational effort.

### 5.2.2 Recap of POT

Now we briefly recap the background for POT. Define the generalized extreme value distribution (GEV) as

$$
H_{\xi}(x)= \begin{cases}\exp \left\{-(1+\xi x)^{-1 / \xi}\right\} & \text { if } \xi \neq 0 \\ \exp \{-\exp \{-x\}\} & \text { if } \xi=0\end{cases}
$$

We say a distribution function $F$ belongs to the maximum domain of attraction of $H_{\xi}$ if there exist constants $c_{n}>0, d_{n} \in \mathbb{R}$ such that $c_{n}^{-1}\left(M_{n}-d_{n}\right) \xrightarrow{d .} H_{\xi}$ where $M_{n}$ denotes the maximum of $n$ i.i.d. samples from $F$. In this case we write $F \in M D A\left(H_{\xi}\right)$. This property is satisfied by a wide variety of distributions, including Cauchy, Pareto, Loggamma, Uniform, Beta, Exponential, Weibull, Gamma, Normal, Lognormal, etc.

Suppose that $Y$ is a random variable with distribution function $F$. The right endpoint is defined as $x_{F}=\sup \{x \in \mathbb{R}: F(x)<1\}$. Here, $x_{F}$ can be either infinite or finite. For a fixed threshold $u<x_{F}$, we call $F_{u}(x)=P(Y-u \leq x \mid Y>u), x \geq 0$ the excess distribution function of $Y$ over the threshold $u$. Finally, we define the GPD as

$$
G_{\xi, \beta}(x)= \begin{cases}1-\left(1+\xi \frac{x}{\beta}\right)^{-1 / \xi} & \text { if } \xi \neq 0 \\ 1-\exp \left\{-\frac{x}{\beta}\right\} & \text { if } \xi=0\end{cases}
$$

where $\xi \in \mathbb{R}, \beta>0$. The support of the distribution is $[0, \infty)$ if $\xi \geq 0$ and $[0,-\beta / \xi]$ if $\xi<0$.
We have the following theorem (see, e.g., [135]):

Theorem 5.1 (Pickands-Balkema-de Haan Theorem). For any $\xi \in \mathbb{R}, F \in M D A\left(H_{\xi}\right)$ if and only if

$$
\lim _{u \uparrow x_{F}} \sup _{0<x<x_{F}-u}\left|F_{u}(x)-G_{\xi, \beta(u)}(x)\right|=0
$$

for some positive function $\beta$.

The theorem shows that under a mild assumption, the excess distribution converges to a GPD with certain parameters as the threshold increases. Therefore, given i.i.d. samples $Y_{1}, \ldots, Y_{n}$ from $F$, we can pick a large threshold $u$, fit the data exceeding $u$ with GPD, and then use the fitted GPD to estimate the target tail quantity. This is called the POT method.

In this chapter, we assume that the target rare event $\{X \in E\}$ can be formulated as $\{f(X) \geq a\}$ where $f$ is a real-valued function and $a$ is a large constant. Under this setting, we estimate the target probability $p$ with the following procedure:

1. Generate i.i.d. simulation samples $X_{1}, \ldots, X_{n}$ and compute $Y_{i}=f\left(X_{i}\right), i=1, \ldots, n$;
2. Pick a threshold $u<a$ following certain criterion;
3. Fit a GPD $G_{\hat{\xi}, \hat{\beta}}$ with the excess data $\left\{Y_{i}-u: 1 \leq i \leq n, Y_{i}>u\right\}$;
4. Output $\hat{p}_{n}^{P O T}=\left(\frac{1}{n} \sum_{i=1}^{n} I\left(Y_{i}>u\right)\right)\left(1-G_{\hat{\xi}, \hat{\beta}}(a-u)\right)$.

As discussed above, the theoretical guarantee for this POT procedure requires that the distribution function of $Y=f(X)$ belongs to $M D A\left(H_{\xi}\right)$ for some $\xi \in \mathbb{R}$. Though this assumption is not verifiable in practice, thanks to its generality the POT method is often applied to real datasets in the extreme event analysis literature. In this chapter, we follow this reasoning and apply this method on simulation data.

We note that after generating some crude MC simulation samples and computing a crude MC estimate, we can always reuse these samples and apply the above procedure to obtain a POT estimate with little additional effort. Therefore, the POT estimate can always serve as a complement to the crude MC estimate, especially when the latter is trivially 0 . In the next section, we will investigate under what conditions the POT estimator works well through multiple experiments.

### 5.3 Numerical Experiments

### 5.3.1 Example 1: Sample Mean

We start from an easy example. Suppose that $X \in \mathbb{R}^{d}$, the simulation output, consists of $d$ i.i.d. random variables. That is, $X^{(1)}, \ldots, X^{(d)}$ are i.i.d. from certain distribution where $X^{(j)}$ denotes the $j$-th element of the vector $X$. We consider $f(X)=\frac{1}{d} \sum_{j=1}^{d} X^{(j)}$, and hence we aim to estimate $P\left(\frac{1}{d} \sum_{j=1}^{d} X^{(j)} \geq a\right)$. With i.i.d. simulation samples $X_{1}, \ldots, X_{n}$, the crude MC estimator is simply $\hat{p}_{n}^{M C}=\frac{1}{n} \sum_{i=1}^{n} I\left(\frac{1}{d} \sum_{j=1}^{d} X_{i}^{(j)} \geq a\right)$. Alternatively, we can use the POT procedure described in Section 5.2.2 to compute $\hat{p}_{n}^{P O T}$.

In order to apply POT, we need to select the threshold $u$ properly, which involves an intrinsic bias-variance tradeoff. Intuitively, if $u$ is too small, then approximating the excess distribution with GPD may bear large bias; on the other hand, if $u$ is too large, then there are too few data exceeding this threshold, which may result in large variance in fitting the GPD parameters. In the literature, some useful tools are proposed to facilitate this selection, such as the mean residual plot [135], but its performance is arguably still case-by-case. In our experiments, for simplicity, we choose $u$ as a certain sample quantile such that it is large but there are still sufficient data above it. In this sample mean example, we set $d=10, n=10^{6}$ and we compare the performance of $u$ being the $0.99 / 0.999 / 0.9999$-th sample quantile. For instance, if $u$ is chosen as the 0.999 -th sample quantile, then there are still $10^{6} \times(1-0.999)=1000$ samples above this threshold to help us fit the GPD parameters. Moreover, it remains an open problem what is the best way to fit the GPD parameters. Throughout this chapter, we use maximum likelihood estimation.

First, we test the methods on a light-tailed distribution, more specifically, normal distribution. We let $X_{i}^{(j)} \sim N(0,1)$. Under this setting, a highly efficient IS scheme is available. That is, we sample $\tilde{X}_{1}, \ldots, \tilde{X}_{n}$ such that $\tilde{X}_{i}^{(j)}$,s are i.i.d. from $N(a, 1)$ instead of $N(0,1)$. Then the IS estimator is computed by $\hat{p}_{n}^{I S}=\frac{1}{n} \sum_{i=1}^{n} I\left(\frac{1}{d} \sum_{j=1}^{d} \tilde{X}_{i}^{(j)} \geq a\right) \exp \left\{-a \sum_{j=1}^{d} \tilde{X}_{i}^{(j)}+d a^{2} / 2\right\}$. We note that $Y=f(X) \sim N(0,1 / d)$, so actually the true probability $p=P(Y \geq a)$ is known, which helps us compare the performance of all these estimators and evaluate the GPD fitting performance.

In order to evaluate these estimators, we generate $n$ simulation samples, compute each estimator, and repeat this process $N=1000$ times. Table 5.1 shows the descriptive statistics of these 1000 estimations with $a=1.5$, where the true probability $p=1.05 \times 10^{-6}$. We see that the carefully designed IS estimator is highly accurate, i.e. the mean is nearly the truth and the standard deviation is small relative to the truth. The mean of the crude MC estimates is close to the truth, which is reasonable since it is actually the mean value of $n \times N=10^{9}$ simulation results. However, the standard deviation of the crude MC estimator is nearly the same as the true probability. In fact, we see that at least $25 \%$ of the crude MC estimates are trivially 0 . For the POT method, we see that POT0.99 (i.e. $u$ is chosen as the 0.99 sample quantile) has large bias but small variance (or standard deviation); POT-0.999 has small bias and the variance is smaller than crude MC; and POT-0.9999 is more unstable than POT- 0.999 as it has larger variance and also gives 0 occasionally. Overall, POT-0.999 appears the best among the three.

Table 5.1: Statistics of each estimator in Example 1 with $N(0,1)$ distribution and $a=1.5 . p=$ $1.05 \times 10^{-6}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.13 \mathrm{E}-06$ | $1.05 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $6.00 \mathrm{E}-06$ |
| IS | $1.05 \mathrm{E}-06$ | $2.39 \mathrm{E}-09$ | $1.04 \mathrm{E}-06$ | $1.05 \mathrm{E}-06$ | $1.05 \mathrm{E}-06$ | $1.05 \mathrm{E}-06$ | $1.06 \mathrm{E}-06$ |
| POT-0.99 | $5.88 \mathrm{E}-07$ | $2.74 \mathrm{E}-07$ | $6.34 \mathrm{E}-08$ | $3.88 \mathrm{E}-07$ | $5.52 \mathrm{E}-07$ | $7.36 \mathrm{E}-07$ | $1.95 \mathrm{E}-06$ |
| POT-0.999 | $1.01 \mathrm{E}-06$ | $6.14 \mathrm{E}-07$ | $1.71 \mathrm{E}-08$ | $5.47 \mathrm{E}-07$ | $8.79 \mathrm{E}-07$ | $1.38 \mathrm{E}-06$ | $3.48 \mathrm{E}-06$ |
| POT-0.9999 | $1.00 \mathrm{E}-06$ | $7.53 \mathrm{E}-07$ | 0 | $4.05 \mathrm{E}-07$ | $8.77 \mathrm{E}-07$ | $1.48 \mathrm{E}-06$ | $3.77 \mathrm{E}-06$ |

Although POT-0.999 still cannot compete with IS, it indeed performs better than crude MC using the same simulation samples. To achieve the same standard deviation as POT-0.999, crude MC requires $n^{\prime}=\left\lceil p(1-p) / s t d^{2}\right\rceil \approx 2.79 \times 10^{6}$, which is almost three times the actual simulation size. Moreover, as mentioned before, crude MC estimator gives 0 frequently, while POT-0.999 can at least give an estimate of a roughly correct magnitude. To illustrate this difference better, we plot the histograms of these two estimators in Figure 5.1, where we find that POT-0.999 is more concentrated around the true probability while over $30 \%$ of the crude MC estimates are 0 .

Now we move $a$ to 2 , so the target event is even rarer $\left(p=1.27 \times 10^{-10}\right.$ ). Table 5.2 shows the descriptive statistics. In this case, since $p$ is too small, crude MC completely fails, in the


Figure 5.1: Histograms of crude MC and POT-0.999 estimates in Example 1 with $N(0,1)$ distribution and $a=1.5$.
sense that all the 1000 estimates are 0 . The IS estimator still performs well. We find that the POT estimators cannot perform well. In order to explain this phenomenon, in Figure 5.2, we plot the fitted GPD tails in five random replications against the true tail. We find that when $a$ is not far from the threshold $u$ (approximately $0.7,1.0,1.2$ respectively for the $0.99,0.999,0.9999$-th sample quantiles), the fitted values are generally close to the true values. When $a$ increases, the fitting becomes more and more inaccurate and unstable, which coincides with our observations on $a=1.5$ and $a=2$. This phenomenon is reasonable, considering that in the POT method, we are leveraging the limited data above $u$, most of which should not be far from $u$, to infer the whole tail. That is, POT may not be suitable if $a$ is too far from the threshold $u$.

Table 5.2: Statistics of each estimator in Example 1 with $N(0,1)$ distribution and $a=2 . p=$ $1.27 \times 10^{-10}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| IS | $1.27 \mathrm{E}-10$ | $3.37 \mathrm{E}-13$ | $1.26 \mathrm{E}-10$ | $1.27 \mathrm{E}-10$ | $1.27 \mathrm{E}-10$ | $1.27 \mathrm{E}-10$ | $1.28 \mathrm{E}-10$ |
| POT-0.99 | $5.72 \mathrm{E}-13$ | $3.68 \mathrm{E}-12$ | 0 | 0 | $1.11 \mathrm{E}-18$ | $7.58 \mathrm{E}-15$ | $7.49 \mathrm{E}-11$ |
| POT-0.999 | $5.65 \mathrm{E}-10$ | $2.25 \mathrm{E}-09$ | 0 | $3.19 \mathrm{E}-18$ | $1.73 \mathrm{E}-12$ | $1.99 \mathrm{E}-10$ | $3.77 \mathrm{E}-08$ |
| POT-0.9999 | $5.29 \mathrm{E}-09$ | $1.72 \mathrm{E}-08$ | 0 | 0 | $1.83 \mathrm{E}-12$ | $1.95 \mathrm{E}-09$ | $1.94 \mathrm{E}-07$ |



Figure 5.2: GPD fitting performance in Example 1 with $N(0,1)$ distribution.

We also investigate the performance in the heavy-tailed case. Let $X_{i}^{(j)} \sim \operatorname{Pareto}(t, \alpha)$ where
$t, \alpha>0$ are the scale and shape parameters. That is, the CDF is $P\left(X_{i}^{(j)} \leq x\right)=1-(t / x)^{\alpha}, x \geq$ $t$. In this case, IS cannot be easily applied. Instead, the CMC estimator defined by $\hat{p}_{n}^{C M C}=$ $\frac{1}{n} \sum_{i=1}^{n} n P\left(X_{i}^{(d)}>\left(\max _{j=1, \ldots, d-1} X_{i}^{(j)}\right) \vee\left(d a-\sum_{j=1}^{d-1} X_{i}^{(j)}\right)\right)$ is known to be unbiased and efficient. We still repeat each estimator $N=1000$ times to make a comparison.

Tables 5.3 to 5.6 show the descriptive statistics where $t=1$ and $\alpha=1.5,2,3,5$. In these experiments, $a$ is tuned manually such that the true probability is close to $10^{-7}$, and hence we could compare the influence of the shape parameter $\alpha$. In all the four experiments, the CMC estimator is highly accurate, so we use the mean of CMC as an approximate to the truth. By contrast, the crude MC estimator performs badly. In fact, in every experiment, over $75 \%$ of the crude MC estimates are 0 . Among the three POT estimators, POT-0.999 still performs generally well, as the mean is close to the truth and the standard deviation is much smaller than crude MC. From the histograms in Figure 5.3, we see that crude MC estimates are almost all 0 while POT-0.999 estimates are more concentrated. Finally, we note that smaller $\alpha$ implies heavier tail, and we observe from these results that POT works better in this case. Especially, when $\alpha=1.5$, the standard deviation of POT-0.999 is only $6.32 \times 10^{-8}$, while it requires $n^{\prime}=\left\lceil p(1-p) / s t d^{2}\right\rceil \approx 2.50 \times 10^{7}$ samples for crude MC to achieve the same standard deviation, which is 25 times the actual $n$.

Table 5.3: Statistics of each estimator in Example 1 with Pareto $(1,1.5)$ distribution and $a=$ 21601.56.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $9.60 \mathrm{E}-08$ | $3.21 \mathrm{E}-07$ | 0 | 0 | 0 | 0 | $2.00 \mathrm{E}-06$ |
| CMC | $9.96 \mathrm{E}-08$ | $1.07 \mathrm{E}-13$ | $9.96 \mathrm{E}-08$ | $9.96 \mathrm{E}-08$ | $9.96 \mathrm{E}-08$ | $9.96 \mathrm{E}-08$ | $9.96 \mathrm{E}-08$ |
| POT-0.99 | $9.68 \mathrm{E}-08$ | $2.29 \mathrm{E}-08$ | $4.54 \mathrm{E}-08$ | $8.09 \mathrm{E}-08$ | $9.45 \mathrm{E}-08$ | $1.11 \mathrm{E}-07$ | $1.88 \mathrm{E}-07$ |
| POT-0.999 | $1.09 \mathrm{E}-07$ | $6.32 \mathrm{E}-08$ | $1.35 \mathrm{E}-08$ | $6.44 \mathrm{E}-08$ | $9.58 \mathrm{E}-08$ | $1.39 \mathrm{E}-07$ | $4.46 \mathrm{E}-07$ |
| POT-0.9999 | $1.78 \mathrm{E}-07$ | $1.75 \mathrm{E}-06$ | $1.11 \mathrm{E}-20$ | $1.58 \mathrm{E}-08$ | $4.64 \mathrm{E}-08$ | $1.11 \mathrm{E}-07$ | $5.13 \mathrm{E}-05$ |

From the observations drawn from this sample mean example, we conclude that POT helps us refine crude MC estimation, especially when the simulation size is relatively large so that there are abundant data above the threshold, the rare-event set boundary $a$ is not far from the threshold, and the tail of the distribution is heavy. In the following subsections, we compare crude MC and POT-0.999 on several other examples to examine the generality of this conclusion.

Table 5.4: Statistics of each estimator in Example 1 with Pareto $(1,2)$ distribution and $a=1000$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $9.70 \mathrm{E}-08$ | $3.22 \mathrm{E}-07$ | 0 | 0 | 0 | 0 | $2.00 \mathrm{E}-06$ |
| CMC | $1.00 \mathrm{E}-07$ | $2.98 \mathrm{E}-13$ | $1.00 \mathrm{E}-07$ | $1.00 \mathrm{E}-07$ | $1.00 \mathrm{E}-07$ | $1.00 \mathrm{E}-07$ | $1.00 \mathrm{E}-07$ |
| POT-0.99 | $9.47 \mathrm{E}-08$ | $2.54 \mathrm{E}-08$ | $4.04 \mathrm{E}-08$ | $7.65 \mathrm{E}-08$ | $9.23 \mathrm{E}-08$ | $1.10 \mathrm{E}-07$ | $2.25 \mathrm{E}-07$ |
| POT-0.999 | $1.12 \mathrm{E}-07$ | $7.22 \mathrm{E}-08$ | $8.15 \mathrm{E}-09$ | $6.10 \mathrm{E}-08$ | $9.41 \mathrm{E}-08$ | $1.45 \mathrm{E}-07$ | $5.01 \mathrm{E}-07$ |
| POT-0.9999 | $1.36 \mathrm{E}-07$ | $1.66 \mathrm{E}-07$ | 0 | $2.17 \mathrm{E}-08$ | $8.31 \mathrm{E}-08$ | $1.82 \mathrm{E}-07$ | $1.52 \mathrm{E}-06$ |

Table 5.5: Statistics of each estimator in Example 1 with Pareto $(1,3)$ distribution and $a=47.625$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $9.90 \mathrm{E}-08$ | $3.25 \mathrm{E}-07$ | 0 | 0 | 0 | 0 | $2.00 \mathrm{E}-06$ |
| CMC | $1.02 \mathrm{E}-07$ | $2.00 \mathrm{E}-12$ | $1.02 \mathrm{E}-07$ | $1.02 \mathrm{E}-07$ | $1.02 \mathrm{E}-07$ | $1.02 \mathrm{E}-07$ | $1.02 \mathrm{E}-07$ |
| POT-0.99 | $7.54 \mathrm{E}-08$ | $2.48 \mathrm{E}-08$ | $2.27 \mathrm{E}-08$ | $5.79 \mathrm{E}-08$ | $7.22 \mathrm{E}-08$ | $8.98 \mathrm{E}-08$ | $1.97 \mathrm{E}-07$ |
| POT-0.999 | $1.20 \mathrm{E}-07$ | $8.48 \mathrm{E}-08$ | $3.66 \mathrm{E}-09$ | $5.82 \mathrm{E}-08$ | $1.00 \mathrm{E}-07$ | $1.58 \mathrm{E}-07$ | $5.07 \mathrm{E}-07$ |
| POT-0.9999 | $1.40 \mathrm{E}-07$ | $1.84 \mathrm{E}-07$ | 0 | $1.60 \mathrm{E}-08$ | $7.63 \mathrm{E}-08$ | $1.93 \mathrm{E}-07$ | $1.80 \mathrm{E}-06$ |

Table 5.6: Statistics of each estimator in Example 1 with Pareto $(1,5)$ distribution and $a=$ 5.234375 .

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $8.60 \mathrm{E}-08$ | $3.01 \mathrm{E}-07$ | 0 | 0 | 0 | 0 | $2.00 \mathrm{E}-06$ |
| CMC | $1.01 \mathrm{E}-07$ | $1.57 \mathrm{E}-11$ | $1.01 \mathrm{E}-07$ | $1.01 \mathrm{E}-07$ | $1.01 \mathrm{E}-07$ | $1.01 \mathrm{E}-07$ | $1.01 \mathrm{E}-07$ |
| POT-0.99 | $2.23 \mathrm{E}-08$ | $1.22 \mathrm{E}-08$ | $2.71 \mathrm{E}-09$ | $1.39 \mathrm{E}-08$ | $1.99 \mathrm{E}-08$ | $2.76 \mathrm{E}-08$ | $9.95 \mathrm{E}-08$ |
| POT-0.999 | $9.55 \mathrm{E}-08$ | $8.17 \mathrm{E}-08$ | $2.29 \mathrm{E}-10$ | $3.81 \mathrm{E}-08$ | $7.39 \mathrm{E}-08$ | $1.29 \mathrm{E}-07$ | $6.42 \mathrm{E}-07$ |
| POT-0.9999 | $1.29 \mathrm{E}-07$ | $1.82 \mathrm{E}-07$ | 0 | $8.10 \mathrm{E}-09$ | $5.87 \mathrm{E}-08$ | $1.79 \mathrm{E}-07$ | $1.58 \mathrm{E}-06$ |

### 5.3.2 Example 2: Random Walk

In this subsection, we consider the overflow of a random walk process. We still suppose that $X^{(1)}, \ldots, X^{(d)}$ are i.i.d. random variables. Then the process $\left\{\sum_{j=1}^{k}\left(X^{(j)}-E X^{(j)}\right): k=\right.$ $1, \ldots, d\}$ is a random walk where the increment at each step has zero mean. We set $f(X)=$ $\max _{k=1, \ldots, d} \sum_{j=1}^{k}\left(X^{(j)}-E X^{(j)}\right)$, and the rare event $\{f(X) \geq a\}$ represents the excursion of this random walk. With $n$ simulation samples $X_{1}, \ldots, X_{n}$, we can compute $\hat{p}_{n}^{M C}$ and $\hat{p}_{n}^{P O T}$. We set $d=10$ and $n=10^{6}$.

Like in Section 5.3.1, we repeat evaluating each estimator for $N=1000$ times. However, without accurate IS or CMC estimators, now we need to estimate the truth in another way. As mentioned before, the average of the $N$ crude MC estimates, denoted by $\overline{\hat{p}_{n}^{M C}}$, is actually the


Figure 5.3: Histograms of crude MC and POT-0.999 estimates in Example 1 with Pareto (1, $\alpha$ ) distribution.
average of $n \times N$ simulation samples from all the $N$ repetitions. Therefore, we can pool these $n \times N$ samples to compute a confidence interval (CI). In particular, in this as well as the following subsections, we use the $95 \%$ CI given by $\overline{\hat{p}_{n}^{M C}} \pm 1.96 \sqrt{\frac{\overline{\hat{p}_{n}^{M C}}\left(1-\overline{\hat{p}_{n}^{M C}}\right)}{n \times N}}$ as an estimate for the true target probability.

Table 5.7 presents the descriptive statistics where $X^{(j)} \sim N(0,1)$ and Tables 5.8 to 5.10 present the statistics where $X^{(j)} \sim \operatorname{Pareto}(1, \alpha)$ for $\alpha=2,3,5$. The rare-event boundary $a$ is tuned such that the true probability is close to $10^{-6}$ for better comparison. Figure 5.4 shows the histograms of the two estimators in these experiments. These results show consistent observations with the previous example. The POT estimator is only slightly biased but the standard deviation is nearly half of that of the crude MC estimator. It also gives non-trivial estimates while about $30 \%$ to $40 \%$ of the crude MC estimates are 0 . Also the POT estimator performs better for heavier-tailed distributions.

Table 5.7: Statistics of each estimator in Example 2 with $N(0,1)$ distribution and $a=15 . \mathrm{CI}$ : $1.23 \times 10^{-6} \pm 6.86 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.23 \mathrm{E}-06$ | $1.09 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $6.00 \mathrm{E}-06$ |
| POT-0.999 | $1.14 \mathrm{E}-06$ | $6.64 \mathrm{E}-07$ | $2.13 \mathrm{E}-08$ | $6.23 \mathrm{E}-07$ | $1.05 \mathrm{E}-06$ | $1.54 \mathrm{E}-06$ | $3.99 \mathrm{E}-06$ |

Table 5.8: Statistics of each estimator in Example 2 with Pareto $(1,2)$ distribution and $a=3000$. CI: $1.06 \times 10^{-6} \pm 6.40 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | max |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.07 \mathrm{E}-06$ | $1.09 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $6.00 \mathrm{E}-06$ |
| POT-0.999 | $1.15 \mathrm{E}-06$ | $4.74 \mathrm{E}-07$ | $2.17 \mathrm{E}-07$ | $8.03 \mathrm{E}-07$ | $1.08 \mathrm{E}-06$ | $1.42 \mathrm{E}-06$ | $3.15 \mathrm{E}-06$ |

Table 5.9: Statistics of each estimator in Example 2 with Pareto $(1,3)$ distribution and $a=200$. CI: $1.17 \times 10^{-6} \pm 6.71 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | max |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.17 \mathrm{E}-06$ | $1.14 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $6.00 \mathrm{E}-06$ |
| POT-0.999 | $1.29 \mathrm{E}-06$ | $5.61 \mathrm{E}-07$ | $1.91 \mathrm{E}-07$ | $8.74 \mathrm{E}-07$ | $1.19 \mathrm{E}-06$ | $1.61 \mathrm{E}-06$ | $3.27 \mathrm{E}-06$ |

Table 5.10: Statistics of each estimator in Example 2 with $\operatorname{Pareto}(1,5)$ distribution and $a=25$. CI: $8.26 \times 10^{-7} \pm 5.63 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $8.26 \mathrm{E}-07$ | $9.23 \mathrm{E}-07$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $1.00 \mathrm{E}-06$ | $6.00 \mathrm{E}-06$ |
| POT-0.999 | $8.72 \mathrm{E}-07$ | $4.60 \mathrm{E}-07$ | $6.98 \mathrm{E}-08$ | $5.39 \mathrm{E}-07$ | $7.86 \mathrm{E}-07$ | $1.10 \mathrm{E}-06$ | $2.75 \mathrm{E}-06$ |



Figure 5.4: Histograms of crude MC and POT-0.999 estimates in Example 2.

### 5.3.3 Example 3: Neural Network

In this example, we consider an example that uses rare-event simulation to evaluate the robustness of neural network classification, a problem that has been recently studied in [14, 125]. We consider the classification problem on the MNIST dataset which contains 70,000 images of handwritten digits. Each image is encoded as a vector $x \in \mathbb{R}^{784}$, and belongs to a class in $\{1, \ldots, 10\}$ (we assume $j=1, \ldots, 9$ corresponds to digit $j$ and $j=10$ corresponds to digit 0 ). The goal of the classification task is to predict the class given the input $x$ accurately. We train a simple neural network with an accuracy of over $97 \%$. Given an input image (encoded as a vector) $x$, this trained neural network outputs the predicted logits $z_{j}(x), j=1, \ldots, 10$, and then this input is classified as $\arg \max _{j} z_{j}(x)$. Suppose that the neural network is able to correctly classify the input $x_{0}$ to its true class $j_{0}$, and we aim to evaluate the robustness of this classification model via simulation. That is, we exert a random perturbation $\varepsilon$ from certain distribution on the input and then compute the simulation output $X=\left(z_{j}\left(x_{0}+\varepsilon\right)\right)_{j=1, \ldots, 10}$. Let $f(X)=\max _{j \neq j_{0}} X^{(j)}-X^{\left(j_{0}\right)}$, and $a=0$. Thus, the target rare event $\{f(X) \geq a\}$ is the event that the classification result changes after the perturbation. We choose the simulation size $n=10^{6}$ and repeat each estimator $N=1000$ times.

We suppose that each element of the random perturbation $\varepsilon$ is i.i.d. with zero mean. For the light-tailed case, we try $N\left(0,0.17^{2}\right)$; for the heavy-tailed case, we try $\operatorname{Pareto}(0.0008,2)$ and $\operatorname{Pareto}(0.02,3)$ minus the mean. From Tables 5.11 to 5.13 , we find that the POT estimator still performs similarly to the previous examples. We skip the histograms as they look similar to the previous ones.

Table 5.11: Statistics of each estimator in Example 3 with $N\left(0,0.17^{2}\right)$ distribution. CI: $1.10 \times$ $10^{-6} \pm 6.49 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.10 \mathrm{E}-06$ | $1.04 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $7.00 \mathrm{E}-06$ |
| POT-0.999 | $1.03 \mathrm{E}-06$ | $5.92 \mathrm{E}-07$ | $3.87 \mathrm{E}-11$ | $5.75 \mathrm{E}-07$ | $9.42 \mathrm{E}-07$ | $1.41 \mathrm{E}-06$ | $3.98 \mathrm{E}-06$ |

Table 5.12: Statistics of each estimator in Example 3 with Pareto $(0.0008,2)$ distribution. CI: $1.07 \times 10^{-6} \pm 6.42 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.07 \mathrm{E}-06$ | $1.03 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $6.00 \mathrm{E}-06$ |
| POT-0.999 | $1.22 \mathrm{E}-06$ | $4.46 \mathrm{E}-07$ | $1.66 \mathrm{E}-07$ | $9.06 \mathrm{E}-07$ | $1.16 \mathrm{E}-06$ | $1.49 \mathrm{E}-06$ | $3.14 \mathrm{E}-06$ |

Table 5.13: Statistics of each estimator in Example 3 with Pareto( $0.02,3$ ) distribution. CI: $1.03 \times$ $10^{-6} \pm 6.28 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | max |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.03 \mathrm{E}-06$ | $1.02 \mathrm{E}-06$ | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ | $5.00 \mathrm{E}-06$ |
| POT-0.999 | $1.33 \mathrm{E}-06$ | $5.19 \mathrm{E}-07$ | $2.43 \mathrm{E}-07$ | $9.59 \mathrm{E}-07$ | $1.27 \mathrm{E}-06$ | $1.63 \mathrm{E}-06$ | $3.66 \mathrm{E}-06$ |

### 5.3.4 Example 4: Queueing System

In this example, we consider a complicated queueing system described as follows. Suppose that customers arrive to the system following a Poisson process with rate $\lambda$. There are $m$ sequential first-in-first-out queues in the system, and the service time of each queue follows a certain distribution. For $k=1, \ldots, m-1$, after the service at the $k$-th queue, each customer joins the $(k+1)$-th queue with probability $p$ and directly leaves the system with probability $1-p$. After the service at the $m$ th queue, each customer leaves the system. Suppose that this simulation model outputs the sojourn times of the first $d$ customers, i.e. $X^{(j)}$ is the sojourn time of the $j$-th customer. We aim to estimate the probability that the maximum sojourn time exceeds $a$, i.e. $f(X)=\max _{j=1, \ldots, d} X^{(j)}$. In the experiments, we fix $\lambda=1, m=10, p=0.8, d=10$. We choose the simulation size $n=10^{6}$ and repeat each estimator $N=100$ times.

Tables 5.14 to 5.17 present the descriptive statistics where the service distribution is chosen as exponential or Pareto distributions. The value of $a$ is not carefully tuned in this example, but we see that the POT estimate still performs generally better than crude MC.

Table 5.14: Statistics of each estimator in Example 4 with $\operatorname{Exp}(1)$ distribution and $a=45$. CI: $3.50 \times 10^{-7} \pm 1.16 \times 10^{-7}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | max |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $3.50 \mathrm{E}-07$ | $5.39 \mathrm{E}-07$ | 0 | 0 | 0 | $1.00 \mathrm{E}-06$ | $2.00 \mathrm{E}-06$ |
| POT-0.999 | $2.95 \mathrm{E}-07$ | $1.93 \mathrm{E}-07$ | $6.35 \mathrm{E}-09$ | $1.23 \mathrm{E}-07$ | $2.75 \mathrm{E}-07$ | $4.41 \mathrm{E}-07$ | $8.34 \mathrm{E}-07$ |

Table 5.15: Statistics of each estimator in Example 4 with $\operatorname{Pareto}(1,2)$ distribution and $a=1000$. CI: $5.30 \times 10^{-7} \pm 1.43 \times 10^{-7}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | max |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $5.30 \mathrm{E}-07$ | $8.22 \mathrm{E}-07$ | 0 | 0 | 0 | $1.00 \mathrm{E}-06$ | $4.00 \mathrm{E}-06$ |
| POT-0.999 | $4.98 \mathrm{E}-07$ | $2.34 \mathrm{E}-07$ | $1.20 \mathrm{E}-07$ | $3.25 \mathrm{E}-07$ | $4.53 \mathrm{E}-07$ | $6.04 \mathrm{E}-07$ | $1.12 \mathrm{E}-06$ |

Table 5.16: Statistics of each estimator in Example 4 with Pareto (1,2) distribution and $a=2000$. CI: $1.40 \times 10^{-7} \pm 7.33 \times 10^{-8}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | max |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $1.40 \mathrm{E}-07$ | $3.77 \mathrm{E}-07$ | 0 | 0 | 0 | 0 | $2.00 \mathrm{E}-06$ |
| POT-0.999 | $1.31 \mathrm{E}-07$ | $7.75 \mathrm{E}-08$ | $2.08 \mathrm{E}-08$ | $7.37 \mathrm{E}-08$ | $1.13 \mathrm{E}-07$ | $1.63 \mathrm{E}-07$ | $3.46 \mathrm{E}-07$ |

Table 5.17: Statistics of each estimator in Example 4 with Pareto $(1,3)$ distribution and $a=50$. CI: $4.40 \times 10^{-7} \pm 1.30 \times 10^{-7}$.

| method | mean | std | $\min$ | $25 \%$ | $50 \%$ | $75 \%$ | $\max$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MC | $4.40 \mathrm{E}-07$ | $7.43 \mathrm{E}-07$ | 0 | 0 | 0 | $1.00 \mathrm{E}-06$ | $3.00 \mathrm{E}-06$ |
| POT-0.999 | $4.50 \mathrm{E}-07$ | $2.28 \mathrm{E}-07$ | $1.07 \mathrm{E}-07$ | $2.89 \mathrm{E}-07$ | $3.85 \mathrm{E}-07$ | $5.68 \mathrm{E}-07$ | $1.10 \mathrm{E}-06$ |

### 5.4 Conclusion

While variance reduction techniques are powerful tools to increase the efficiency of crude MC in rare-event simulation, they often heavily rely on tractable problem structures and careful algorithmic design, which may not be possible for complex practical problems. This motivates us to study the use of POT, a prominent method in extreme event analysis, in rare-event simulation as an alternative to variance reduction. We formulate and test our POT approach on four rare-event simulation examples. Naturally, POT may not be as efficient as carefully designed algorithms such as IS or CMC when they are available. However, it outperforms crude MC in a reasonably wide spectrum of problems. It performs especially well when the simulation size is relatively large, the rare event boundary is not far from the threshold, and the tail of the distribution is heavy. Compared to crude MC, the POT estimator usually has smaller standard deviation with equal simulation size, and it gives an estimate of a roughly correct magnitude instead of a trivial estimate 0 . Therefore, if efficient variance reduction techniques are not available, then our POT procedure can be used to refine and improve the crude MC estimate. For instance, if after generating some simulation sam-
ples, we find that none of them hits the target event but we cannot afford generating more samples, then we can compute the POT estimator with these existing samples to get a non-trivial estimate. In the future, we will investigate how to further improve the performance of our POT procedure, including threshold selection and fitting of GPD parameters.

# Chapter 6: Calibrating Over-parameterized Simulation Models: A Framework via Eligibility Set 

### 6.1 Introduction

Stochastic simulation aims to compute output performances for complex systems that lack analytical tractability. Other than prediction, this evaluation capability facilitates downstream decision-making tasks including optimization, feasibility test and sensitivity analysis, by changing the design parameter value and monitoring the output from the simulation. Simulation modeling as such has been widely used across multiple disciplines in operations research and management [147, 148], finance [34], intelligent systems [149, 150, 17] and model-based learning [151].

The reliability of simulation models, nonetheless, depends crucially on how well these models capture reality, which in turn calls for the correctness of calibration. The latter refers to the task of choosing the model parameter values, and the associated uncertainty quantification on the inference errors. Often times, data represented by the output of the model are only observed at an aggregate level, without full information on the detailed underlying dynamics of the considered system. This is both the strength and challenge of simulation modeling: By building a model for the underlying dynamics, one could answer questions that require looking into unobserved scenarios by suitably tuning the design configuration, which is beyond what standard machine learning or statistics can provide. On the other hand, without direct data, there could be no consistency in estimating these model parameters, and the reliability of the ultimate model prediction could be in doubt.

In stochastic or discrete-event simulation, calibration has been studied mostly assuming direct data on the parameter of interest. In this situation, the focus is on assessing the impact of the statistical error in fitting the parameter that is propagated to the simulation output, known as the input uncertainty or input error (see, e.g., the surveys [152, 153, 154, 155, 156, 157, 158, 159,

160]). This problem has drawn wide attention among the simulation community in recent years, studying methods ranging from the bootstrap [161, 162, 163, 164, 165], finite-difference and delta method [166, 167, 168, 169], robust optimization [170, 171, 172, 173, 174] and relatedly empirical likelihood [175], sectioning [176] and Bayesian methods [177, 178, 179, 180].

Nonetheless, when the simulation model validity is checked against output data instead of assuming direct input data, the problem is considerably much more difficult [181]. This problem arises routinely in the simulation model construction process, yet a rigorous study appears quite open. Conventionally, this problem is treated under the umbrella of model calibration and validation $[182,183]$ which suggests an iterative model refinement process. Once a simulation model is built, it is tested against real data, via two-sample statistical hypothesis tests that compare simulated outputs and real outputs [184] or Turing tests [185] ("validation"). If it is determined from these tests that the model is inadequate, the model would be refined by either expanding the system logic or tuning the model parameters ("calibration"). These two steps are reiterated until the model is satisfactory. Though intuitive, this approach is ad hoc, potentially time-consuming and, moreover, there is no guarantee of a satisfactory model at the end. The ad-hoc-ness arises because just by locating model parameter values that match the simulated versus real outputs in terms of simple hypothesis tests, there is no guarantee that 1 ) there exists a unique set of parameter values that gives the match and 2) the simulation model is good enough for output dimensions different from the one being tested.

Issue 1) above regards how to locate model parameter values, and contains two sub-issues on the existence and uniqueness respectively. The existence of matching parameter values means either the model is correctly specified, or it is parameterized in a rich enough manner so that its degree of freedom is more than necessary to include the "truth", i.e., the model is over-parameterized. In the latter case, uniqueness may not hold as there could be more than one set, or even potentially many sets, of parameter values that give the exact match. This issue is known as non-identifiability [186]. Though this resembles overfitting in machine learning, its underlying challenge and handling methodology are fundamentally different: Whereas one can diagnose overfitting by measur-
ing generalization errors on test sets, in simulation modeling the main goal is to change the design parameters to values that are unobserved before, which by definition have no available test data to start with.

Issue 2), on the other hand, concerns the adequacy in assessing the match between simulated and real data. Challenges arise from the stochasticity and high-dimensionality of outputs in complex simulation models. In stochastic simulation, the outputs are random by definition and are often in the form of serial trajectory. In the literature this stochasticity and high dimensionality is usually avoided by simply using the means of certain outputs as a summary and assessing the match with real data via classical low-dimensional two-sample hypothesis tests [182, 183]. However, even if these tests deem an accurate match, it does not necessarily mean the model parameter takes the correct value, because we may have ignored the many possible other output dimensions which, related to the first issue above, exacerbates the non-identifiability problem.

Our goal in this chapter is to provide a systematic framework to address the two issues above, focusing on over-parameterized regime with high-dimensional stochastic outputs. In other words, we presume that the considered simulation model is sufficiently expressive in resembling the true hidden dynamics, but it can contain many parameters. The real data and the model outputs can be multivariate or in the form of time series. Under these presumptions, we first properly define the target as to find what we call an eligibility set, namely the set of all possible parameter values such that the real and simulated outputs match, in the idealized situation when there are an infinite amount of data on both ends. Note that depending on the over-parameterization and the matching criterion, non-identifiability can arise, and in this case it is theoretically impossible to locate the true parameter value. The eligibility set is defined as the theoretically best (i.e., smallest) set of parameter values that one can possibly aim to obtain under the given degree of non-identifiability. This set contains the true value, and the smaller it is, the more accurately we know regarding the truth. In this way, instead of enforcing a consistent estimation as in classical statistics, we relax the estimation target to a set in order to bypass the non-identifiability issue.

The eligibility set provides a theoretical limit on what we can achieve. When finite data are
available, estimation error kicks in and we could only attain a correct eligibility set up to a certain statistical confidence. Moreover, the data also affects how we produce the matching criterion. Ideally, we want our estimation to be correct (i.e., the set we create indeed contains the truth with high confidence) and non-conservative (i.e., a small set). To this end, we study a framework using feature extraction-then-aggregation as follows. In the extraction phase, we identify the important "summaries" for the outputs, where a summary is regarded as effective if it possesses the ability to distinguish an incorrect model output from the truth. In this chapter, we simply use unsupervised learning tools, such as penultimate layer extraction from neural network models, to locate the important output characteristics as summaries. While there are considerably other potential better approaches, we will see how our suggestion applies well to examples to a reasonable extent.

In the next phase, we compare the match between real and simulated data by testing the similarities of these features. Since the number of features could be large in order to well represent the high-dimensional outputs, the individual hypothesis tests on these features need to be properly aggregated to control familywise error. Here, good choices of the statistics and the aggregation method are based on the tractability in calibrating the acceptance-rejection threshold or analyzing its $p$-value ("correct"), and a small resulting Type II error probability on mistakenly accepting a wrong value ("non-conservative"). Based on these considerations, we propose an approach using an aggregation of the Kolmogorov-Smirnov (KS) statistics, which leads to the following analytics tool: Given a parameter value, we can say, with a statistical confidence, whether this value is inside the eligibility set. If it is not, then the value is not the truth with the prescribed confidence. If it is, then the value is in the smallest possible set that includes the truth (though no conclusion is drawn on whether this value is the truth). In short, the more values are "rejected" or equivalently the smaller is our constructed eligibility set, the more accurate we can locate the truth.

Our approach through eligibility set constructed from feature extraction-then-aggregation described above, though intuitive, faces two admitted limitations. First is our presumption of overparameterized models which means that the "truth" is recovered with the correct parameter values. Often times, however, the model, even when richly parameterized, could still deviate from the true
dynamics and incur a systematic model error - a discrepancy with the truth even when the parameters are best calibrated. Our second limitation concerns computation. When the input dimension is high, it is difficult to fully construct the eligibility set, and thus we refrain ourselves in the task of hypothesis testing instead of set estimation, i.e., given a parameter configuration, we test whether it is inside the eligibility set or in other words "eligible". Both issues appear challenging to us and deserve much follow-up study. Handling the first issue could borrow the Bayesian perspective by incorporating an associated prior on model error (e.g., [187, 188], and also [189, 190] which consider stochastic outputs), or understanding the frequentist behavior of distance minimization [191, 192, 193]. For the second issue, a potential remedy is to recast the goal as to compute confidence bounds on target performance measure, and use robust optimization to compute these bounds by treating the eligibility set as the constraint or the so-called uncertainty set or ambiguity set (see Section 6.2.1). Such optimization problems are in the form of high-dimensional simulation optimization. We point out that while these important limitations are unresolved in the current chapter, a systematic methodology to calibrate models that are over-parameterized (i.e., with negligible model error) with multi-dimensional outputs already appear quite open in the stochastic simulation literature and worth a starting investigation.

Finally, we test our methodology on the calibration of the Agent-Based Interactive DiscreteEvent Simulation Environment (ABIDES) [194, 195], a detailed model to emulate market order dynamics. Faced with an increasing demand of a simulation tool to analyze market participant's behaviors, ABIDES is designed to emulate the interaction of individual agents with an exchange agent (e.g., mimicking NASDAQ) and output the emergent time series of the bid-ask prices and order volumes. ABIDES enables the simulation of tens of thousands of trading agents, many securities, and can involve various behavioral types of trading agents. The trading agents and exchange agents communicate through a kernel to request the information of the limit order book and perform trading order placement. The calibration of the ABIDES model here means finding a configuration of agents such that the generated price series is similar to observed data - arising from either an ABIDES simulation with unknown parameters or historical price series. To test
the power of our approach, we measure the similarity in terms of stylized facts of the limit order book that can be used as realism metrics [195]. The specific stylized facts that we examine include heavy tails and aggregation normality, absence of autocorrelations of the log return distribution, and volatility clustering. In other words, we show that our calibration framework, which operates in "black-box" regime without knowing the explicit realism metrics in advance, is able to select models that match data with respect to them.

### 6.1.1 Comparisons with Existing Approaches

We compare our work with several existing approaches and discuss our relative strengths and limitations. First, as mentioned previously, our approach aims to provide a systematic framework to carry out validation-calibration for stochastic simulation that rigorizes the ad hoc suggestions in the past. To this end, we also point out a related line of work on discrete-event or queueing inference that leverages analytic solutions or heavy-traffic approximations of the model to recover input parameters [196, 197, 198, 199, 200, 201]. Different from these works, our approach does not rely on specific structures of the discrete-event model and assume only a simulatable blackbox input-output relation. The closest to our approach is [202] that uses distributionally robust optimization (DRO) [203, 204, 205, 206] to compute bounds for performance measures. Like us, they consider a match between real and simulated data using the KS statistics. However, they consider single-dimensional output and assume the considered system has a $V$-statistic structure. In contrast, we do not impose assumptions on our system structure and the output can be highdimensional, thus necessitate our feature extraction-then-aggregation framework. Moreover, [202] considers distributional input instead of parametric inputs, the former allowing a tractable (i.e., locally convergent) algorithm for solving the constructed DRO, while conducting an analogous task efficiently in our current setup remains open. We also mention that a preliminary conference version of the current work has appeared in [207], but without the elaborate investigation on the eligibility set guarantees, the feature-based approach to handle high-dimensional outputs, and the detailed numerics.

Second, in computer experiments, the calibration problem admits a long-standing history and is also known as simulation-based or likelihood-free inference, as the likelihood function of data is only accessible via the (black- or grey-box) simulation model [208]. This is regarded as an inverse problem in scientific fields as one attempts to recover physical parameters from surrogate models built on physical laws [209, 210, 211]. The predominant approach uses a Bayesian perspective, by placing priors on the unknown input model parameters and computing posterior for inference [187, 212, 213, 214, 215]. Recent work such as [216] further improves the extrapolation capability by using covariate shift techniques. The Bayesian approach is advantageously automatic in addressing non-identifiability, as the posterior distribution is universally well-defined. However, as the likelihood function is not directly available, the posterior can often be updated only via approximate Bayesian computation (ABC) [217, 218], by running acceptance-rejection on simulated samples that is computationally costly and requires ad hoc tuning [219, 220]. In contrast to this literature, our approach aims for a frequentist guarantee, hinging on our notion of eligibility set to bypass non-identifiability, which relies less heavily on the correctness of the underlying model assumptions such as the prior distribution.

Third, we note that the notion of non-identifiability and the general idea of constructing a confidence set for non-identifiable parameters align with the set identification literature in econometrics [221, 222, 223, 224]. Compared to this literature, our development differs in the extra randomness and complexity introduced by the simulation model. More specifically, in the econometrics literature, the focus is the identified set $\arg \min _{\theta \in \Theta} Q(\theta)$ where $Q(\cdot)$ is an objective function (the distance $d$ introduced later can be seen as $Q$ in their framework) and then confidence sets can be constructed with the sample objective function $Q_{N}(\cdot)$. This idea is similar to ours, but in our case, $Q_{N}(\theta)$ is still random due to the stochasticity in simulation. Moreover, our objective function can be much more complicated and even coming from a black-box simulator.

Finally, machine learning techniques have also been applied to solve calibration problems. This literature focuses on building surrogate models and minimizing the distance with real data. More specifically, when the simulation model is computationally costly, a surrogate model could be built
as a proxy to speed up the evaluation and calibration. For instance, [225] proposes a model-assisted GAN where a convolutional neural network (CNN) is trained as an emulator to mimic the simulator. The similarity between the simulated data and the emulated data is determined by encoding the data with the same CNN and comparing the $L_{2}$ norm of the difference. Alternatively, [226] suggests to use extreme gradient boosted trees (XGBoost) to learn an ensemble of classification and regression trees (CART) which ultimately provides an approximation for agent-based models. After a surrogate is built, the computational load to evaluate and to calibrate by matching with real data could be significantly reduced. We could also integrate such surrogate models in our framework. That is, if the simulation run is time-consuming, then we could build a surrogate model for the original simulation model. However, doing so will affect the statistical guarantee of our eligibility set, because now we should account for the surrogate model error against the original simulation model.

### 6.2 Calibration Framework

We consider a family of output probability distributions $\left\{P^{\theta}: \theta \in \Theta\right\}$ on $\mathbb{R}^{m}$ where $\theta$ is an unknown $d$-dimensional parameter on the parameter space $\Theta \subset \mathbb{R}^{d}$. The unknown true parameter is denoted as $\theta_{0} \in \Theta$. We suppose real-world output data $X_{1}, \ldots, X_{N}$ drawn from $P^{\theta_{0}}$ is available. On the other hand, given any $\theta \in \Theta$, we can generate a simulated random sample $Y_{1}^{\theta}, \ldots, Y_{n}^{\theta}$ from $P^{\theta}$ via a black-box machine. Note that, besides simulatability, we do not make any further assumptions on the relationship between $\theta$ and $P^{\theta}$, which can have a complicated structure as exhibited by most large-scale simulation models. Also note that the setup above implies implicitly that the simulation model represented by $P^{\theta}$ is expressive enough, i.e., over-parameterized, so that it covers the true output distribution.

Our goal is to infer the true value of $\theta$ from the real output data and our simulation capability. For convenience, we use $P_{N}^{\theta_{0}}$ and $P_{n}^{\theta}$ to denote the empirical probability distributions determined by the real output data set $X_{1}, \ldots, X_{N}$ and the simulated output data set $Y_{1}^{\theta}, \ldots, Y_{n}^{\theta}$ respectively.

That is,

$$
P_{N}^{\theta_{0}}(\cdot)=\frac{1}{N} \sum_{k=1}^{N} \delta_{X_{k}}(\cdot), \quad P_{n}^{\theta}(\cdot)=\frac{1}{n} \sum_{j=1}^{n} \delta_{Y_{j}^{\theta}}(\cdot)
$$

where $\delta$ denotes the Dirac measure. Naturally, we also use $F^{\theta_{0}}$ and $F^{\theta}$ to denote the distribution functions of $P^{\theta_{0}}$ and $P^{\theta}$, and correspondingly $F_{N}^{\theta_{0}}$ and $F_{n}^{\theta}$ to denote the respective empirical distribution functions.

### 6.2.1 Remedying Non-Identifiability via Eligibility Sets

To infer $\theta_{0}$, the standard statistical approach would be to obtain a point estimate as close to $\theta_{0}$ as possible. However, when the simulation model is over-parameterized as in the setup above, there could be many values other than $\theta_{0}$ that give rise to the same output pattern, an issue known as non-identifiability. More precisely, the observability of output-level data allows us to only obtain information on the output distribution $P^{\theta_{0}}$. Consider any statistical distance $d(\cdot, \cdot)$ between two probability distributions that is valid (in particular, the distance is zero if two distributions are identical). The best we can achieve is to find $\theta$ such that $d\left(P^{\theta}, P^{\theta_{0}}\right)=0$. If there exists $\theta_{0}^{\prime} \in \Theta$ such that $\theta_{0} \neq \theta_{0}^{\prime}$ but $d\left(P^{\theta_{0}^{\prime}}, P^{\theta_{0}}\right)=0$, the output data form fundamentally prevents us from identifying which one is the true parameter. This non-identifiability stems from the unobservability of the detailed dynamics that create these data.

Our idea to bypass the non-identifiability issue is to construct a region that contains the true parameter value instead of trying to get a "best" point estimation. We call this region the eligibility set. To start with, imagine we have infinitely many real data so that $P^{\theta_{0}}$ is fully known. In this case, the set $\left\{\theta \in \Theta: d\left(P^{\theta}, P^{\theta_{0}}\right)=0\right\}$ clearly contains the true parameter value $\theta_{0}$, and moreover it contains the most refined information on the parameter value coming from the output data.

In practice where we only have a finite real data set of size $N$, we construct a statistically confident eligibility set by relaxing the distance from zero to a small threshold, namely

$$
\begin{equation*}
\mathcal{E}=\left\{\theta \in \Theta: d\left(P^{\theta}, P_{N}^{\theta_{0}}\right) \leq \eta\right\} \tag{6.1}
\end{equation*}
$$

where $\eta \in \mathbb{R}^{+}$is a suitable constant. This construction is highly intuitive: If $P^{\theta}$ is sufficiently close to the empirical distribution $P_{N}^{\theta_{0}}$, then we take $\theta$ as an acceptable candidate, and vice versa. The use of (6.1) is highlighted in the following simple lemma:

Lemma 6.1. If $d(\cdot, \cdot)$ satisfies

$$
\begin{equation*}
\mathbb{P}\left(d\left(P^{\theta_{0}}, P_{N}^{\theta_{0}}\right) \leq \eta\right) \geq 1-\alpha \tag{6.2}
\end{equation*}
$$

where $\mathbb{P}$ refers to the probability with respect to the real data $X_{1}, \ldots, X_{N}$, then $\mathcal{E}=\{\theta \in \Theta$ : $\left.d\left(P^{\theta}, P_{N}^{\theta_{0}}\right) \leq \eta\right\}$ is a $(1-\alpha)$-level confidence region for $\theta_{0}$, i.e.,

$$
\begin{equation*}
\mathbb{P}\left(\theta_{0} \in \mathcal{E}\right) \geq 1-\alpha \tag{6.3}
\end{equation*}
$$

Moreover, if (6.2) holds asymptotically as $N \rightarrow \infty$, then (6.3) holds asymptotically as well.

Proof of Lemma 6.1. The result follows by noting that $d\left(P^{\theta_{0}}, P_{N}^{\theta_{0}}\right) \leq \eta$ implies $\theta_{0} \in \mathcal{E}$.

The main implication of Lemma 6.1 is that, in order to obtain a statistically valid confidence set for parameter $\theta$, it suffices to construct a statistical distance $d(\cdot, \cdot)$ in the nonparametric space of probability distributions. Suppose we have knowledge on the sampling distribution on $d\left(P^{\theta_{0}}, P_{N}^{\theta_{0}}\right)$, then we can obtain $\eta$ as the quantile of this distribution so that $\left\{Q: d\left(Q, P_{N}^{\theta_{0}}\right) \leq \eta\right\}$ is a $(1-\alpha)$ level confidence region for $P^{\theta_{0}}$, consequently giving rise to a confidence region for $\theta_{0}$ by "lifting" from the nonparametric to parametric space. The nonparametric viewpoint here is beneficial as it is typically extremely difficult to find the right parametric class for simulation models (hence the term "likelihood-free" inference as discussed in Section 6.1.1). On the other hand, it is relatively easy to construct $d(\cdot, \cdot)$ such that we know its universal sampling distribution of $d\left(Q, Q_{N}\right)$, for general output distribution $Q$ and its associated empirical distribution $Q_{N}$.

An easy but handy observation, which we will utilize heavily in Section 6.4, is that $d(\cdot, \cdot)$ above can be a semi-metric, i.e., $d(P, Q)$ can be 0 not only when $P=Q$ but potentially when $P \neq Q$. Because of this, we can extract a certain dimension or transformation of the output in
constructing this $d$. For instance, let $\Pi P$ be the distribution of a transformation of the output generated from $P$. Then we can consider $d(\Pi P, \Pi Q)$ for any statistical distance $d$ defined on the suitable probability distribution space. If we adopt such a semi-metric, then the non-identifiability should correspondingly be defined with respect to this semi-metric between $P$ and $Q$.

Before moving to more details on $d(\cdot, \cdot)$ and $\eta$, we discuss the relation of (6.1) to the robust optimization (RO) literature [227, 228]. The latter advocates decision-making under the worst-case scenario as a principled approach to handle optimization under uncertainty, where the worst case is taken over the uncertain parameters in the model. In computing the worst case, RO places the target performance measure in the objective, and constrain the uncertain parameters to lie in the so-called uncertainty set or ambiguity set, which is a set believed to contain the true parameter with high confidence. In DRO, the uncertain parameters are the underlying probability distributions in a stochastic problem [203, 205, 204]. In this setting, a common approach to construct the uncertainty set is a neighborhood ball measured by statistical distance such as the Kolmogorov-Smirnov (KS) distance [229] that we will utilize, $\phi$-divergence [230, 206, 170, 173, 174, 231] and Wasserstein distance [232, 233, 234]. It can also be constructed based on moment matching [204, 235, 171] or imposition of shape information [236, 237, 238]. Like DRO, our set (6.1) is a confidence region that is constructed to cover the truth with high confidence and, although our target parameters are finite-dimensional, the set is created via scrutinizing the unknown distributions as in DRO. However, different from this literature, our uncertainty set is constructed by matching the model and the real data at the output level as a way to bypass non-identifiability.

In addition to the similarity between eligibility set and uncertainty set, the worst-case notion in (D)RO that gives rise to bounds on performance measure also plays a role in our framework. Although we do not pursue in this chapter, this role is important for two reasons. First is that when the input is high-dimensional, it is difficult to compute the entire eligibility set $\mathcal{E}$, and focusing on bounds for relevant performance measure simplifies the problem. Second, in simulation analysis our goal often is to evaluate a target performance measure or conduct a downstream task that utilizes such an evaluation. For concreteness, say this target is $\psi\left(\theta_{0}\right)$ where $\psi$ is a function dependent
on $\theta$. The optimization pair

$$
\begin{align*}
& \text { maximize/minimize } \psi(\theta)  \tag{6.4}\\
& \qquad \text { subject to } \theta \in \mathcal{E}
\end{align*}
$$

which utilizes the eligibility set $\mathcal{E}$ as the feasible region, results in valid confidence bounds for $\psi\left(\theta_{0}\right)$. We rigorize this as:

Lemma 6.2. If $\mathcal{E}$ is a $(1-\alpha)$-level confidence set for the true parameter $\theta_{0}$, i.e., $\mathbb{P}\left(\theta_{0} \in \mathcal{E}\right) \geq 1-\alpha$ where $\mathbb{P}$ refers to the probability with respect to the data, then the optimal values of (6.4), denoted $Z^{*}$ and $Z_{*}$ respectively, satisfy

$$
\mathbb{P}\left(Z_{*} \leq \psi\left(\theta_{0}\right) \leq Z^{*}\right) \geq 1-\alpha
$$

That is, $Z^{*}$ and $Z_{*}$ constitute $(1-\alpha)$-level upper and lower confidence bounds for $\psi\left(\theta_{0}\right)$.

Proof of Lemma 6.2. The result follows by noting that $\theta_{0} \in \mathcal{E}$ implies $Z_{*} \leq \psi\left(\theta_{0}\right) \leq Z^{*}$, so that

$$
\mathbb{P}\left(Z_{*} \leq \psi\left(\theta_{0}\right) \leq Z^{*}\right) \geq \mathbb{P}\left(\theta_{0} \in \mathcal{E}\right) \geq 1-\alpha .
$$

That is, by combining the nonparametric distance $d$ in Lemma 6.1 and the DRO formulation in Lemma 6.2, we arrive at confidence bounds for target performance measure $\psi\left(\theta_{0}\right)$, from which one can utilize for many tasks such as optimization and feasibility analyses. While using DRO is not the focus of this chapter, it serves to highlight the downstream use of our eligibility set and, moreover, gives an approach to use it even if the eligibility set cannot be fully constructed which is important when the input parameter space is huge.

### 6.2.2 Constructing Eligibility Sets: An Elementary Case

To implement our proposed framework, we need to resolve two questions. First is the choice of distance measure $d$ and the associated constant $\eta$ in order to achieve the guarantee depicted in Lemma 6.1. Second, at least in elementary cases where we want to approximate the entire $\mathcal{E}$, we need to be computationally able to determine whether $d\left(P^{\theta}, P_{N}^{\theta_{0}}\right) \leq \eta$ or not for any $\theta \in \Theta$.

For the first question, a good choice of $d$ should satisfy that (i) the resulting uncertainty set is non-conservative, i.e., it is as small as possible in terms of "volume"; (ii) the associated threshold $\eta$ that satisfies (6.2) is obtainable; (iii) $d$ is efficient to compute. For one-dimensional output, we consider the Kolmogorov-Smirnov (KS) distance

$$
d\left(P_{1}, P_{2}\right)=\sup _{x \in \mathbb{R}}\left|F_{1}(x)-F_{2}(x)\right|
$$

where $P_{1}, P_{2}$ are probability distributions and $F_{1}, F_{2}$ are respectively their cumulative distribution functions. Since $d\left(P^{\theta_{0}}, P_{N}^{\theta_{0}}\right)$ is exactly the KS statistic for the goodness-of-fit for $F^{\theta_{0}}$, its sampling distribution is well-known and the threshold can be readily set as $\eta=q_{1-\alpha} / \sqrt{N}$, where $q_{1-\alpha}$ is the ( $1-\alpha$ )-quantile of $\sup _{t \in[0,1]}|B B(t)|$ and $B B(\cdot)$ denotes a standard Brownian bridge. This choice of ( $d, \eta$ ) satisfies (6.2) asymptotically as $N$ increases, and thus Theorem 6.1 applies.

In practice, the simulation size is also finite as constrained by computational capacity. This requires us to use $P_{n}^{\theta}$ and $F_{n}^{\theta}$ to approximate $P^{\theta}$ and $F^{\theta}$, and hence $d\left(P_{n}^{\theta}, P_{N}^{\theta_{0}}\right)$ to approximate $d\left(P^{\theta}, P_{N}^{\theta_{0}}\right)$. To summarize, the eligibility set for $\theta$ is constructed as

$$
\begin{equation*}
\mathcal{E}=\left\{\theta \in \Theta: \sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \frac{q_{1-\alpha}}{\sqrt{N}}\right\} \tag{6.5}
\end{equation*}
$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup _{t \in[0,1]}|B B(t)|$. Given any $\theta \in \Theta$, we can generate a simulated sample $Y_{1}^{\theta}, \ldots, Y_{n}^{\theta} \sim P^{\theta}$ and check whether it is in $\mathcal{E}$ defined by (6.5). If there are infinitely many values in the parameter space, then we can choose a finite number of values, say $\theta_{1}, \ldots, \theta_{m} \in \Theta$ as "representatives", which can be either deterministic grid points or randomly
sampled.
It is worth discussing why we choose to use KS over other perceivably natural candidates for model distances such as the Kullback-Leibler (KL) divergence. The KL divergence, and more generally $\phi$-divergence, require likelihood ratio that is usually unavailable for simulation models. Thus, to empirically estimate these divergences, we need to either discretize the distributions or use variational representation, but even for well-designed consistent KL estimators, the convergence rate can be arbitrarily slow [239]. An alternative choice is the Wasserstein distance, which is relatively easy to compute thanks to its linear programming representation. However, the corresponding threshold $\eta$ cannot be easily calibrated.

Putting everything together, we propose Algorithm 6.1 to compute the eligibility set for the unknown parameter $\theta$. In the following, we present some theoretical results that guide the amount of simulation runs needed to produce the guarantees offered by Lemma 6.1, and the conservativeness level of the resulting eligibility set. Theorem 6.1 first gives a sufficient condition on the simulation size.

```
Algorithm 6.1: Constructing eligibility set of \(\theta\).
    Input: The real output data \(X_{1}, \cdots, X_{N}\). The number of candidate \(\theta\) 's \(m\). The simulation
        replication size \(n\). The confidence level \(1-\alpha\).
    Output: An approximate eligibility set \(\hat{\mathcal{E}}_{m}\).
    1 Generate \(\theta_{1}, \cdots, \theta_{m} \in \Theta\);
    \({ }_{2}\) For \(i=1, \cdots\), \(m\) do
    3 Generate an i.i.d. random sample of simulated data \(Y_{1}^{\theta_{i}}, \cdots, Y_{n}^{\theta_{i}} \sim P^{\theta_{i}}\);
    \(4 \quad\) Compute \(c_{i}=\sup _{x \in \mathbb{R}}\left|\frac{1}{n} \sum_{j=1}^{n} I\left(Y_{j}^{\theta_{i}} \leq x\right)-\frac{1}{N} \sum_{l=1}^{N} I\left(X_{l} \leq x\right)\right| ;\)
    End
    6 Return the eligibility set \(\hat{\mathcal{E}}_{m}=\left\{\theta_{i}: c_{i} \leq q_{1-\alpha} / \sqrt{N}\right\}\).
```

Theorem 6.1. Suppose that $X_{1}, \cdots, X_{N}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta_{0}}, \cdots, Y_{n}^{\theta_{0}}$ is an i.i.d. simulated sample from $P^{\theta_{0}} . F_{N}^{\theta_{0}}$ and $F_{n}^{\theta_{0}}$ are respectively the empirical distribution functions of the two random samples. If $n=\omega(N)$ as $N \rightarrow \infty$, then

$$
\liminf _{n, N \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \frac{q_{1-\alpha}}{\sqrt{N}}\right) \geq 1-\alpha
$$

From the view of hypothesis testing, Theorem 6.1 analyzes the probability of Type I error in validating the simulation model outputs. Here, Type I error refers to the eligibility set $\hat{\mathcal{E}}_{m}$ excluding the true parameter $\theta_{0}$ given that $\theta_{0}$ is chosen as a representative. The theorem states that the Type I error probability is asymptotically bounded by $\alpha$ as long as the simulation size $n$ is of larger order than the real data size $N$. In this sense, we may regard the discretized eligibility set $\hat{\mathcal{E}}_{m}$ computed with Algorithm 6.1 as an approximate confidence region for $\theta_{0}$.

On the other hand, we also analyze Type II error, i.e., some representative that does not have the same output distribution as the truth is accepted. Theorem 6.2 provides an upper bound for the probability that $F^{\theta} \neq F^{\theta_{0}}$ yet $\theta$ is eligible.

Theorem 6.2. Suppose that $X_{1}, \cdots, X_{N}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta}$ is an i.i.d. simulated sample from $P^{\theta} . F_{N}^{\theta_{0}}$ and $F_{n}^{\theta}$ are respectively the empirical distribution functions of the two random samples. $F^{\theta_{0}}$ and $F^{\theta}$ denote the cumulative distribution functions of $P^{\theta_{0}}$ and $P^{\theta}$. Suppose that $\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|>0$. For any $\varepsilon_{1}, \varepsilon_{2}>0$ such that $\varepsilon_{1}+\varepsilon_{2}<\sup _{x \in \mathbb{R}} \mid F^{\theta}(x)-$ $F^{\theta_{0}}(x) \mid$, if

$$
N>\left(\frac{q_{1-\alpha}}{\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\varepsilon_{1}-\varepsilon_{2}}\right)^{2}
$$

then

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \leq q_{1-\alpha} / \sqrt{N}\right) \leq 2\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right) .
$$

Theorem 6.2 states that if a representative $\theta$ gives a different output distribution from $\theta_{0}$, then for sufficiently large $N$, we have a finite-sample upper bound for the probability that $\theta$ is still included in $\hat{\mathcal{E}}_{m}$. Note that the smaller is $\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|$, i.e. the closer is the simulation output distribution to the truth, the larger are the required simulation size $n$ and real data size $N$ in order to guarantee that $\theta$ is distinguished from $\theta_{0}$ with high probability, which coincides with intuition. With Theorem 6.2, we may easily develop a Type II error guarantee for Algorithm 6.1, as shown in Theorem 6.3.

Theorem 6.3. We follow Algorithm 6.1 to obtain $\hat{\mathcal{E}}_{m}$. For any $\varepsilon, \varepsilon_{1}, \varepsilon_{2}>0$ such that $\varepsilon_{1}+\varepsilon_{2}<\varepsilon$, if

$$
N>\left(\frac{q_{1-\alpha}}{\varepsilon-\varepsilon_{1}-\varepsilon_{2}}\right)^{2}
$$

then

$$
\mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon, \theta_{i} \in \hat{\mathcal{E}}_{m}\right) \leq 2 m\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right) .
$$

Here, Type II error probability is characterized by the probability that there exists some representative $\theta_{i}$ such that $\sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon$ (i.e. the output distribution given by $\theta_{i}$ is not too close to the truth) yet it is accepted in the eligibility set $\hat{\mathcal{E}}_{m}$. Theorem 6.3 gives a finite-sample upper bound for this probability. With chosen $\varepsilon, \varepsilon_{1}, \varepsilon_{2}$ and sufficiently large real data size $N$, we could suitably choose the simulation size $n$ and the number of representatives $m$ such that with high probability, the eligibility set $\hat{\mathcal{E}}_{m}$ only includes representatives that are close enough to the truth in terms of output distribution. In this way, Theorem 6.3 provides guidance on how to choose $n$ and $m$ according to the real data size $N$. For example, we have the following corollaries.

Corollary 6.1. We follow Algorithm 6.1 to obtain $\hat{\mathcal{E}}_{m}$. If $\log m=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then for any $\varepsilon>0$,

$$
\lim _{m, n, N \rightarrow \infty} \mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon, \theta_{i} \in \hat{\mathcal{E}}_{m}\right)=0 .
$$

Corollary 6.2. We follow Algorithm 6.1 to obtain $\hat{\mathcal{E}}_{m}$. If $m=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then

$$
\lim _{m, n, N \rightarrow \infty} \mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\sqrt{\frac{\log m}{m}}, \theta_{i} \in \hat{\mathcal{E}}_{m}\right)=0 .
$$

Corollary 6.1 shows that if $\log (m)=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then for any $\varepsilon>0$, asymptotically almost surely any eligible representative $\theta_{i}$ in $\hat{\mathcal{E}}_{m}$ satisfies that $\sup _{x \in \mathbb{R}} \mid F^{\theta_{i}}(x)-$ $F^{\theta_{0}}(x) \mid \leq \varepsilon$. If further we have $m=o(N)$, then Corollary 6.2 suggests that asymptotically almost surely any $\theta_{i}$ in $\hat{\mathcal{E}}_{m}$ satisfies $\sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right| \leq \sqrt{\log m / m}$, where $\sqrt{\log m / m}$ shrinks in $m$. Intuitively, as long as the simulation size is sufficient and the number of representatives is moderate, we could confidently conclude that any eligible representative is hardly distinct from the truth.

Finally, we introduce a two-sample variant of the eligibility set. We say $\theta$ is eligible if

$$
\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log (\alpha / 2)} .
$$

Here, we are considering the two-sample KS statistic $\sqrt{\frac{n N}{n+N}} \sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right|$ and approximating $q_{1-\alpha}$ with $\sqrt{-\frac{1}{2} \log (\alpha / 2)}$. We summarize the two-sample version calibration framework in Algorithm 6.2.

```
Algorithm 6.2: Constructing eligibility set of \(\theta\) (two-sample variant).
    Input: The real output data \(X_{1}, \cdots, X_{N}\). The number of candidate \(\theta\) 's \(m\). The simulation
                replication size \(n\). The confidence level \(1-\alpha\).
    Output: An approximate eligibility set \(\hat{\mathcal{E}}_{m}^{v}\).
    \({ }_{1}\) Generate \(\theta_{1}, \cdots, \theta_{m} \in \Theta\);
    \({ }_{2}\) For \(i=1, \cdots\), \(m\) do
    3 Generate an i.i.d. random sample of simulated data \(Y_{1}^{\theta_{i}}, \cdots, Y_{n}^{\theta_{i}} \sim P^{\theta_{i}}\);
    \(4 \quad\) Compute \(c_{i}=\sup _{x \in \mathbb{R}}\left|\frac{1}{n} \sum_{j=1}^{n} I\left(Y_{j}^{\theta_{i}} \leq x\right)-\frac{1}{N} \sum_{l=1}^{N} I\left(X_{l} \leq x\right)\right|\);
    5 End
    6 Return The eligibility set \(\hat{\mathcal{E}}_{m}^{v}=\left\{\theta_{i}: c_{i} \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log (\alpha / 2)}\right\}\).
```

Similarly, we first provide a theoretical guarantee for Type I error probability.
Theorem 6.4. Suppose that $X_{1}, \cdots, X_{N}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta_{0}}, \cdots, Y_{n}^{\theta_{0}}$ is an i.i.d. simulated sample from $P^{\theta_{0}} . F_{N}^{\theta_{0}}$ and $F_{n}^{\theta_{0}}$ are respectively the empirical distribution functions
of the two random samples. If $n=N \geq 4$, then

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log \left(\frac{\alpha}{2}\right)}\right) \geq 1-1.085 \alpha .
$$

If $n=N \geq 458$, then

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log \left(\frac{\alpha}{2}\right)}\right) \geq 1-\alpha .
$$

In contrast to Theorem 6.1, Theorem 6.4 provides finite-sample bounds for the Type I error probability instead of asymptotic guarantees. Moreover, while Theorem 6.1 requires that $n$ grows in a higher order than $N$, here we let $n=N$. From the theorem, as long as $n=N \geq 4$, Type I error probability is close to achieving the nominal level $\alpha$, and as long as $n=N \geq 458$, Type I error probability is fully controlled by $\alpha$.

Regarding Type II error probability, we also have the following theorems.
Theorem 6.5. Suppose that $X_{1}, \cdots, X_{N}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta}$ is an i.i.d. simulated sample from $P^{\theta} . F_{N}^{\theta_{0}}$ and $F_{n}^{\theta}$ are respectively the empirical distribution functions of the two random samples. $F^{\theta_{0}}$ and $F^{\theta}$ denote the cumulative distribution functions of $P^{\theta_{0}}$ and $P^{\theta}$. Suppose that $\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|>0$. For any $\varepsilon_{1}, \varepsilon_{2}>0$ such that $\varepsilon_{1}+\varepsilon_{2}<\sup _{x \in \mathbb{R}} \mid F^{\theta}(x)-$ $F^{\theta_{0}}(x) \mid$, if

$$
\frac{n N}{n+N}>\frac{-\frac{1}{2} \log (\alpha / 2)}{\left(\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\varepsilon_{1}-\varepsilon_{2}\right)^{2}}
$$

then

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log (\alpha / 2)}\right) \leq 2\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right) .
$$

Theorem 6.6. We follow Algorithm 6.2 to obtain $\hat{\mathcal{E}}_{m}^{v}$. For any $\varepsilon, \varepsilon_{1}, \varepsilon_{2}>0$ such that $\varepsilon_{1}+\varepsilon_{2}<\varepsilon$,
if

$$
\frac{n N}{n+N}>\frac{-\frac{1}{2} \log (\alpha / 2)}{\left(\varepsilon-\varepsilon_{1}-\varepsilon_{2}\right)^{2}}
$$

then

$$
\mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon, \theta_{i} \in \hat{\mathcal{E}}_{m}^{v}\right) \leq 2 m\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right) .
$$

The implications of Theorems 6.5 and 6.6 are similar to Theorems 6.2 and 6.3, except that now $n$ and $N$ need to together satisfy some conditions in order that the finite-sample bounds stated in the theorems are valid. We could also derive corollaries on how to choose $m$ and $n$ according to $N$ to control Type II error probability. Indeed, Corollaries 6.1 and 6.2 still hold if we replace $\hat{\mathcal{E}}_{m}$ with $\hat{\mathcal{E}}_{m}^{v}$. That is, we have the following corollaries:

Corollary 6.3. We follow Algorithm 6.2 to obtain $\hat{\mathcal{E}}_{m}^{v}$. If $\log m=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then for any $\varepsilon>0$,

$$
\lim _{m, n, N \rightarrow \infty} \mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon, \theta_{i} \in \hat{\mathcal{E}}_{m}^{v}\right)=0 .
$$

Corollary 6.4. We follow Algorithm 6.2 to obtain $\hat{\mathcal{E}}_{m}^{v}$. If $m=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then

$$
\lim _{m, n, N \rightarrow \infty} \mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\sqrt{\frac{\log m}{m}}, \theta_{i} \in \hat{\mathcal{E}}_{m}^{v}\right)=0
$$

Generally, we would recommend using the two-sample variant compared to one-sample since the former does not require that the simulation size $n$ grows faster than the real size $N$. Instead, $n$ could be equal to $N$, which reduces the computational load. However, when we already have a
large number of simulation runs $(n \gg N)$ (e.g., used to train the "features" for high-dimensional outputs discussed in Section 6.4), we are contented with the one-sample version.

### 6.3 Elementary Numerics

In this section, we apply our calibration framework to calibrate simple queueing models. We will show how our methodology recovers the parameters which are nearly non-identifiable from the truth.

### 6.3.1 An M/M/1 Example

First, we apply our methodology on a simple M/M/1 queueing model example. The interarrival time distribution is $\operatorname{Exp}(\lambda)$ and the service time distribution is $\operatorname{Exp}(\mu)$. Moreover, we follow the first-come first-served (FCFS) discipline. Our goal is to calibrate $(\lambda, \mu)$ from the average sojourn time of the first 100 customers. The parameter space is set as $(0,2)^{2}$ and the true value is $(0.5,1)$.

First of all, to show that our framework could recover the true parameter when the problem is identifiable, we suppose that the true value of $\lambda$ is known and we only calibrate $\mu$. We use the two-sample version of Algorithm 6.1 to obtain the eligibility set with $m=500, n=N=100$ and $\alpha=0.05$. Figure 6.1 shows the approximated KS distance against the value of $\mu$. The red line represents the threshold $\eta$, and thus the points below the red line form the eligibility set. From the figure, we see that the eligibility set centers around the true value 1 .


Figure 6.1: Numerical result for calibrating $\mu$ in the $\mathrm{M} / \mathrm{M} / 1$ example: KS distance against $\mu$.

Now we calibrate $(\lambda, \mu)$ simultaneously with $m=1000, n=N=100$ and $\alpha=0.05$. Figure 6.2 a shows how the KS distance changes with $\lambda$ and $\mu$. Figure 6.2 b visualizes the resulting eligibility set.


Figure 6.2: Numerical results for calibrating $(\lambda, \mu)$ in the $\mathrm{M} / \mathrm{M} / 1$ example.

Judging from Figure 6.2, we observe that this problem is close to non-identifiable. In fact, it is well known that when $\lambda / \mu<1$, the steady-state sojourn time distribution is $\operatorname{Exp}(\mu-\lambda)$. Our simulation output is the average sojourn time of the first 100 customers, which serves as an approximation to the steady-state. Thus, intuitively, the distance between the simulated data and the true data mainly depends on the distance between $\mu-\lambda$ and the truth. In Figure 6.2b, the dashed line represents $\mu-\lambda=0.5$. The points in the eligibility set indeed lie around this line.

In such a case where the problem is close to non-identifiable while the data size and the simulation size is not sufficiently large, if we carry out a stochastic gradient descent (SGD) or other optimization approaches attempting to minimize the distance between the true data and the simulated data, then it is possible that we will arrive at a misleading point which is actually far from the true value. However, with the concept of eligibility set, we could construct a region that covers the true parameter with high confidence. This demonstrates the key motivation of our approach in addressing non-identiafiability by relaxing point to set-level estimation.

### 6.3.2 A G/G/1 Example

Now we generalize the problem setting to a G/G/1 queue with interarrival time distribution $\operatorname{Gamma}(k, \theta)$ ( $k$ is the shape parameter and $\theta$ is the scale parameter) and service time distribution Lognormal $\left(\mu, \sigma^{2}\right)$. Suppose that the output of the simulation model is the sojourn time of the first $K=10$ customers. We aim to calibrate the parameter $(k, \theta, \mu, \sigma)$ from the output data. The true value is set as $(1,1,-2,2)$ and the parameter space is chosen as $(0,5) \times(0,5) \times(-5,5) \times(0,5)$. We construct eligibility sets with $\alpha=0.05, m=100,000$ and $n=N=100,500,1000$. Figure 6.3 shows the evolution of the eligibility set as the data size and the simulation size grow. The diagonal graphs present the histograms of the points in the eligibility set in each dimension while the off-diagonal ones visualize the eligibility set via pairwise scatter plots.

From the figure, we find that as the data size and the simulation size grow, the eligibility set gradually becomes more concentrated around the true value. Moreover, we observe that though $\mu$ and $\sigma$ could be tuned relatively accurately, $(k, \theta)$ is close to non-identifiable. Even when $n=N=$ 1000, the relationship between the first two dimensions of points in the eligibility set still looks like a reciprocal curve. Intuitively, the mean value of $\operatorname{Gamma}(k, \theta)$ is $k \theta$, and thus as long as $k \theta$ is close to the true value 1 , it is hard to well distinguish whether $(k, \theta)$ is also close to the truth only judging from the customers' sojourn time.

Now we fix $\mu=-2$ and $\sigma=2$ and focus on calibrating $(k, \theta)$. We randomly sample $m=1000$ points in the parameter space $(0,5)^{2}$ and for each point, we test whether it is in the eligibility set with $n=N=1000,2000,5000,10000$. Figure 6.4 shows how the eligibility set shrinks as $n$ and $N$ grow. Though the problem is not structurally non-identifiable, it indeed requires a large number of samples in order to accurately locate the truth.

### 6.4 Feature-Based Construction of Eligibility Sets

We present our recipe to calibrate simulation models beyond the simple cases considered in Sections 6.2.2 and 6.3. This requires suitable construction of the distance measure $d$. We propose a


Figure 6.3: Evolution of the eligibility set as the data size and the simulation size increase.
two-stage procedure for this construction, first a feature extraction stage that reduces the dimension of the outputs to a manageable number, and second a feature aggregation stage to put together the extracted features into a statistical distance $d$ that satisfy the three attractive properties discussed in Section 6.2.2. The latter requires analyzing the statistical properties of the aggregation schemes as we will illustrate. For the first stage, we extract features using the penultimate layer of a neural network, trained using three approaches: auto-encoder, generative adversarial network (GAN) and its variant Wasserstein GAN (WGAN). For the second stage, we aggregate features using three methods: supremum of Kolmogorov-Smirnov statistics (SKS), supremum of sample mean differences (SSMD), and ellipsoidal sample mean differences (ESMD). These aggregated statistics


Figure 6.4: Numerical results for the G/G/1 example.
give rise to a statistical distance, which we compare against computable critical values (at a given confidence level) to decide the eligibility of a parameter value. Figure 6.5 gives an overview of our feature-based eligibility decision framework.

### 6.4.1 Feature Extraction

Our feature-based calibration starts by extracting features from the output $Z$, which from now on are assumed to be potentially high-dimensional. These features are defined via a summary function $f$ acted on the original input space. That is, we find a proper summary function $f$ such that $f(Z) \in \mathbb{R}^{K}$, in which case we extract $K$ features from the output.

In general, there are many possible methods to extract features and the choice of method does not affect the correctness. More specifically, with summary function $f$, the eligibility set is $\{\theta \in$


Figure 6.5: Feature-based calibration framework diagram.
$\left.\Theta: d\left(P^{\theta} \circ f^{-1}, P_{N}^{\theta_{0}} \circ f^{-1}\right) \leq \eta\right\}$ (or approximately $\left\{\theta \in \Theta: d\left(P_{n}^{\theta} \circ f^{-1}, P_{N}^{\theta_{0}} \circ f^{-1}\right) \leq \eta\right\}$ ). As long as the threshold $\eta$ is calibrated properly according to the distance measure $d$, the eligibility set is guaranteed to cover the true parameter $\theta_{0}$ with high probability. However, even though such a correctness guarantee can be readily ensured by properly choosing $\eta$, the choice of $f$ affects the conservativeness in terms of the size of the eligibility set. Consider the case that $Z_{1} \sim P^{\theta_{1}}$, $Z_{2} \sim P^{\theta_{2}}$ and $P^{\theta_{1}} \neq P^{\theta_{2}}$. If we choose the summary function $f$ such that $f\left(Z_{1}\right)$ and $f\left(Z_{2}\right)$ have the same distribution, then we cannot distinguish between $\theta_{1}$ and $\theta_{2}$ only with the extracted features, though they should have been distinguishable. Therefore, a good method should have the power to distinguish different parameters.

Here we leverage several machine learning methods to extract features, including auto-encoder [240], GAN [241], and WGAN [242]. The details can be found in Section 6.7, and here we provide some key explanation and how we use these methods. Auto-encoder is a bottleneck-like structure, where the information is compressed by the first half of the network and then reconstructed by the second half. By minimizing the mean squared error between the original and reconstructed inputs, the network is trained to learn the latent representation in the middle layer. GAN contains two competing neural networks, where one of them is a generative model, which learns how to generate samples similar to the input distribution, and the other one is a discriminative model, which learns how to differentiate the generated samples from the input ones. The training procedure of

GAN can be framed as a minimax two-layer game, where the discriminator is trained to minimize the probability of incorrectly classifying the input data and the simulated data, and the generator is trained to maximize the probability of the discriminator to make a classification mistake. [241] shows that when the discriminator is trained to be optimal for any given generator, then the training of GAN can be viewed as minimizing the Jensen-Shannon (JS) divergence. However, [243] mentions that the instability of GAN's training is caused by vanishing gradient when JS divergence is applied to distributions supported by low dimensional manifolds. As an improvement, [242] proposes WGAN, where it shares similar network architecture with GAN but the objective is about minimizing the Wasserstein- 1 distance instead. Though GAN or WGAN were originally proposed to generate data with the same distribution as the true data, its generator or discriminator could serve as a feature extractor. Specifically, the discriminator-based feature extractor has been successfully applied in various fields [244, 245].

In our work, we consider five feature extraction techniques, including auto-encoder, GAN discriminator output, GAN hidden layer, WGAN critic output and WGAN hidden layer. With a pre-trained auto-encoder, one can apply its encoder part to extract hidden feature from the input data. For GAN discriminator output, a usual way is to directly use the trained discriminator for extraction, which can summarize the output with a single number. GAN hidden layer serves an alternative way to leverage the discriminator, that is, using the output from the last hidden layer instead of the final output. Similarly, since WGAN shares a similar network architecture, one can also use the direct output from the trained critic part (like the discriminator in GAN), or apply a similar trick to obtain output from the last hidden layer. Intuitively, using GAN discriminator output or WGAN critic output to summarize a high dimensional data into a single number might result in much loss of information and thus bring conservativeness. However, using the output of the last hidden layer not only helps in dimensionality reduction without losing too much information but also introduces less conservativeness. We will make a comparison and discuss more in our experimental results in Section 6.5.

### 6.4.2 Feature Aggregation

From now on, we suppose that we have found a way to extract $K$ features from the output. To facilitate presentation, we abuse the notation $X_{i}$ in this subsection to refer to the extracted features, i.e., the summary function $f$ applied to the raw real or simulated samples. Suppose that we have the true sample features $X_{1}, \ldots, X_{N} \in \mathbb{R}^{K}$ with parameter $\theta_{0}$ and simulated sample features $Y_{1}^{\theta}, \ldots, Y_{n}^{\theta} \in \mathbb{R}^{K}$ with parameter $\theta$. We need a way to aggregate the features to judge whether $\theta=\theta_{0}$.

We will introduce three methods and analyze their correctness and conservativeness. The first method, SKS, is a generalization of our approach in Section 6.2.2 to multiple output dimensions by using the Bonferroni correction. We will provide theoretical guarantees for its errors. The second method is SSMD that applies the Bonferroni correction to sample mean differences, and the third method ESMD uses the sum of squares of sample mean differences. For the last two methods, we calibrate the thresholds or critical values with normal approximations borrowed from standard statistics inference tools. However, it is challenging to analyze the Type II error probabilities for them, so we will investigate them via examples and draw insights therein.

Intuitively, using more features usually implies more information, and it is easier to distinguish different parameters. However, with more features the threshold is also higher, which in turn makes it more difficult to detect wrong parameters. Generally, for each method, there is no certain conclusion regarding whether using more or less features would be better in terms of conservativeness. Overall, we would recommend using SKS with all the features since theoretically it does not introduce much conservativeness, whereas SSMD and ESMD do not, at least currently, have the same level of theoretical guarantees. Moreover, SSMD and ESMD only compare the mean and the covariance matrix while SKS compares the whole distribution, and hence accounts for more information.

## Supremum of Kolmogorov-Smirnov Statistics (SKS).

This is a generalization of the calibration approach in Section 6.2 .2 from $K=1$ to $K>1$, by using the Bonferroni correction. We focus on the one-sample version, as the results for the two-sample version could be adapted similarly. In this case, we say $\theta$ is eligible (i.e., inside the eligibility set) if

$$
\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta}(x)-F_{N, k}^{\theta_{0}}(x)\right| \leq \frac{q_{1-\alpha / K}}{\sqrt{N}}, \forall k=1, \ldots, K
$$

where $F_{N, k}^{\theta_{0}}$ and $F_{n, k}^{\theta}$ are the empirical distribution functions for $X_{1, k}, \ldots, X_{N, k}$ and $Y_{1, k}^{\theta}, \ldots, Y_{n, k}^{\theta}$. We note that this method compares the marginal distributions over each dimension, but not the joint distribution. Having the same marginal distributions does not ensure the same joint distribution, thus injecting conservativeness to our method. Alternatively, directly comparing the joint distributions can be less conservative (e.g., using the multivariate Dvoretzky-Kiefer-Wolfowitz inequality [246]), but such a comparison statistic is computationally challenging for high-dimensional distributions.

To justify the correctness of our methods, we analyze the probability of Type I error, i.e., $\theta_{0}$ is eligible according to our criterion. Special focus will be on how the required simulation size depends on the number of features $K$, that is, whether the requirement is substantial when facing high-dimensionality. In the one-dimensional case, we have stated Theorem 6.1. Now suppose that we jointly consider $K$ features and we use the Bonferroni correction. Then we have the following theorem.

Theorem 6.7. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta_{0}}, \cdots, Y_{n}^{\theta_{0}} \in$ $\mathbb{R}^{K}$ is an i.i.d. simulated sample from $P^{\theta_{0}} . F_{N, k}^{\theta_{0}}$ and $F_{n, k}^{\theta_{0}}$ are respectively the empirical distribution functions of the $k$-th component of the two random samples. If $n=\omega(N)$ as $N \rightarrow \infty$, then

$$
\liminf _{n, N \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right| \leq \frac{q_{1-\alpha / K}}{\sqrt{N}}, \forall 1 \leq k \leq K\right) \geq 1-\alpha
$$

Similar to Theorem 6.1, Theorem 6.7 justifies that the eligibility set is an asymptotically valid confidence region for any fixed $K$ as long as $n$ is of higher order than $N$. Next, to see how $n$ should scale with a growing $K$, we state Theorem 6.8 , which shows that the asymptotic guarantee still holds as long as $n / N$ grows in a higher order than $\log ^{2} K$, which is a relatively slow rate.

Theorem 6.8. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta_{0}}, \cdots, Y_{n}^{\theta_{0}} \in$ $\mathbb{R}^{k}$ is an i.i.d. simulated sample from $P^{\theta_{0}} . F_{N, k}^{\theta_{0}}$ and $F_{n, k}^{\theta_{0}}$ are respectively the empirical distribution functions of the $k$-th component of the two random samples. If $n / N=\omega\left(\log ^{2} K\right)$ as $K \rightarrow \infty$, then

$$
\liminf _{K \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right| \leq \frac{q_{1-\alpha / K}}{\sqrt{N}}, \forall 1 \leq k \leq K\right) \geq 1-\alpha
$$

The above theorems provide us some guidance on how to choose the simulation size $n$ according to the data size $N$ and number of features $K$ in order to ensure the correctness. Note that in the high-dimensional case, by replacing $q_{1-\alpha}$ with $q_{1-\alpha / K}$, we are increasing the threshold. However, this increase is not that substantial. Indeed, from [247], we know that $\mathbb{P}\left(\sup _{0 \leq t \leq 1}|B B(t)| \leq z\right)=$ $1-2 \sum_{v=1}^{\infty}(-1)^{v-1} e^{-2 v^{2} z^{2}}$ for $z>0$, and thus $\alpha=2 \sum_{v=1}^{\infty}(-1)^{v-1} e^{-2 v^{2} q_{1-\alpha}^{2}} \leq 2 e^{-2 q_{1-\alpha}^{2}}$. Hence we get that $q_{1-\alpha}^{2} \leq-\log (\alpha / 2) / 2$. Therefore, $q_{1-\alpha / K}^{2}=O(\log K)$ as $K \rightarrow \infty$. That is to say, when we simultaneously consider a large number of features, using Bonferroni correction will not bring much conservativeness.

Parallel to the one-dimensional case, we are also interested in whether our methods can efficiently detect the wrong parameters. Thus, we analyze the probability of Type II error, which is characterized by the scenario that $\theta \neq \theta_{0}$ is eligible. Intuitively, the difficulty still depends on the discrepancy between the distribution functions. If $\theta \neq \theta_{0}$ but they result in the same output distribution, then we cannot identify which is the true parameter only by comparing the output data. The following theorem, which generalizes Theorem 6.2, shows how the conservativeness depends on this discrepancy as well as $n, N, K$.

Theorem 6.9. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta} \in$
$\mathbb{R}^{K}$ is an i.i.d. simulated sample from $P^{\theta} . F_{N, k}^{\theta_{0}}$ and $F_{n, k}^{\theta}$ are respectively the empirical distribution functions of the $k$-th component of the two random samples. $F_{k}^{\theta_{0}}$ and $F_{k}^{\theta}$ denote the cumulative distribution functions of the $k$-th component under $P^{\theta_{0}}$ and $P^{\theta}$. Suppose that $\max _{1 \leq k \leq K} \sup _{x \in \mathbb{R}} \mid F_{k}^{\theta}(x)-$ $F_{k}^{\theta_{0}}(x) \mid>0$. For any $\varepsilon_{1}, \varepsilon_{2}>0$ such that $\varepsilon_{1}+\varepsilon_{2}<\max _{1 \leq k \leq K} \sup _{x \in \mathbb{R}}\left|F_{k}^{\theta}(x)-F_{k}^{\theta_{0}}(x)\right|$, if

$$
N>\left(\frac{q_{1-\alpha / K}}{\max _{1 \leq k \leq K} \sup _{x \in \mathbb{R}}\left|F_{k}^{\theta}(x)-F_{k}^{\theta_{0}}(x)\right|-\varepsilon_{1}-\varepsilon_{2}}\right)^{2},
$$

then

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta}(x)-F_{N, k}^{\theta_{0}}(x)\right| \leq \frac{q_{1-\alpha / K}}{\sqrt{N}}, \forall 1 \leq k \leq K\right) \leq 2\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right) .
$$

Comparing with Theorem 6.2, we note that $q_{1-\alpha / K}$ grows only logarithmically in $K$ as $K \rightarrow \infty$, so the minimum required $N$ does not increase much. Overall, using the Bonferroni correction does not bring much conservativeness compared to only using one feature, and thus appears superior as it accounts for more information. In fact, numerical experiments in Section 6.5 shows that the Bonferroni correction generally works well.

## Supremum of Sample Mean Differences (SSMD).

Another natural idea to aggregate features is to directly compare their sample means. We denote $\bar{X}_{k}=\frac{1}{N} \sum_{j=1}^{N} X_{j, k}$ and $\bar{Y}_{k}=\frac{1}{n} \sum_{i=1}^{n} Y_{i, k}^{\theta}$. Assume that $\operatorname{var}\left(X_{1, k}\right), \operatorname{var}\left(Y_{1, k}^{\theta}\right)<\infty$ for any $k$. We say $\theta$ is eligible if

$$
\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta_{k}, \forall k=1, \ldots, K
$$

where

$$
\eta_{k}=z_{1-\alpha /(2 K)} \sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \operatorname{var}\left(X_{1, k}\right)}
$$

with $z_{1-\alpha /(2 K)}$ being the $(1-\alpha /(2 K))$-quantile of standard normal distribution. In practice, $\operatorname{var}\left(X_{1, k}\right)$ could be estimated using the sample variance.

Indeed, by the central limit theorem (CLT), we know that for sufficiently large $N$ and $n, \bar{X}_{k}-\bar{Y}_{k}$
is approximately distributed as $N\left(E\left(X_{1, k}\right)-E\left(Y_{1, k}\right), \frac{\operatorname{var}\left(X_{1, k}\right)}{N}+\frac{\operatorname{var}\left(Y_{1, k}^{\theta}\right)}{n}\right)$. If $\theta=\theta_{0}$, then the approximate distribution is $N\left(0,\left(\frac{1}{N}+\frac{1}{n}\right) \operatorname{var}\left(X_{1, k}\right)\right)$. We combine this normal approximation with the Bonferroni correction. Note that the idea is similar to the two-sample hypothesis testing, and thus when $n$ and $N$ are large, the approximate correctness is implied. More concretely, we have the following theorem:

Theorem 6.10. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta_{0}}, \cdots, Y_{n}^{\theta_{0}} \in$ $\mathbb{R}^{K}$ is an i.i.d. simulated sample from $P^{\theta_{0}}$. For any $k=1, \ldots, K$, we denote $\bar{X}_{k}=\frac{1}{N} \sum_{j=1}^{N} X_{j, k}, \bar{Y}_{k}=$ $\frac{1}{n} \sum_{i=1}^{n} Y_{i, k}^{\theta_{0}}$ and assume that var $\left(X_{1, k}\right)=\operatorname{var}\left(Y_{1, k}^{\theta_{0}}\right)<\infty$. Use $\widehat{\operatorname{var}}_{k}$ to denote the sample variance of $X_{j, k}$ 's. If $n /(n+N) \rightarrow \rho$ where $0 \leq \rho \leq 1$ as $n, N \rightarrow \infty$, then

$$
\liminf _{n, N \rightarrow \infty} \mathbb{P}\left(\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq z_{1-\alpha /(2 K)} \sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \widehat{\text { var }}_{k}}, \forall k=1, \ldots, K\right) \geq 1-\alpha
$$

Like Theorem 6.7 for SKS, Theorem 6.10 controls the Type I error of SSMD. Regarding Type II error that drives the level of conservativeness, SSMD appears challenging to fully analyze. Nonetheless, we will look at special cases to draw insights. First, we give an example to show that using Bonferroni correction can be conservative in terms of relative difference in the Type II error probability. That is, the ratio of the Type II error probability of using all the $K$ features to the error of using only one feature could be pretty large. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. random sample from $N\left(\theta_{0}, I_{K}\right)$ (where $I_{K}$ denotes the $K \times K$ identity matrix) and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta} \in \mathbb{R}^{K}$ is an i.i.d. random sample from $N\left(\theta, I_{K}\right)$. Given a confidence level $1-\alpha$, we define $\eta$ and $\eta^{\prime}$ by

$$
\mathbb{P}\left(\left|N\left(0, \frac{1}{N}+\frac{1}{n}\right)\right| \leq \eta\right)=1-\alpha, \mathbb{P}\left(\left|N\left(0, \frac{1}{N}+\frac{1}{n}\right)\right| \leq \eta^{\prime}\right)=1-\frac{\alpha}{K} .
$$

We suppose that $\Delta_{1}=\theta_{01}-\theta_{1}>0$ and $\theta_{0 k}=\theta_{k}$ for $k=2, \ldots, K\left(\theta_{0 k}\right.$ and $\theta_{k}$ respectively denote the $k$-th component of $\theta_{0}$ and $\theta$ ). If we only use the first feature, that is, we reject $\theta=\theta_{0}$ if
$\left|\bar{X}_{1}-\bar{Y}_{1}\right|>\eta$, then the Type II error probability is

$$
p_{1}:=\mathbb{P}\left(\left|\bar{X}_{1}-\bar{Y}_{1}\right| \leq \eta\right)=\mathbb{P}\left(\left|N\left(\Delta_{1}, \frac{1}{N}+\frac{1}{n}\right)\right| \leq \eta\right) .
$$

If we use all the $K$ features, that is, we reject $\theta=\theta_{0}$ if $\exists k,\left|\bar{X}_{k}-\bar{Y}_{k}\right|>\eta^{\prime}$, then the Type II error probability is

$$
p_{2}:=\mathbb{P}\left(\forall k,\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta^{\prime}\right)=\left(1-\frac{\alpha}{K}\right)^{K-1} \mathbb{P}\left(\left|N\left(\Delta_{1}, \frac{1}{N}+\frac{1}{n}\right)\right| \leq \eta^{\prime}\right) .
$$

The following theorem shows that in this setting, $p_{2} / p_{1}$ grows exponentially:

Theorem 6.11. We suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $N\left(\theta_{0}, I_{K}\right)$ and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta} \in \mathbb{R}^{K}$ is an i.i.d. simulated sample from $N\left(\theta, I_{K}\right)$. We suppose that $\Delta_{1}=\theta_{01}-\theta_{1}>0$ and $\theta_{0 k}=\theta_{k}$ for $k=2, \ldots, K . p_{1}$ and $p_{2}$ are respectively the Type II error probability of only using the first feature and using all the $K$ features. For fixed $K>1$, as $N, n \rightarrow \infty$, we have that $p_{2} / p_{1}$ grows exponentially in $N$ and $n$.

Next, we show that using all the features with Bonferroni correction is not too conservative in terms of the absolute difference in the Type II error probability. Here we consider a more general setting. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. random sample from $N\left(\theta_{0}, \Sigma\right)$ (where $\Sigma$ is a positive definite matrix) and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta} \in \mathbb{R}^{K}$ is an i.i.d. random sample from $N(\theta, \Sigma)$. Given a confidence level $1-\alpha$, we define $\eta_{k}$ and $\eta_{k}^{\prime}$ by

$$
\mathbb{P}\left(\left|N\left(0,\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}\right)\right| \leq \eta_{k}\right)=1-\alpha, \mathbb{P}\left(\left|N\left(0,\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}\right)\right| \leq \eta_{k}^{\prime}\right)=1-\frac{\alpha}{K} .
$$

We suppose that $\theta \neq \theta_{0}$. We denote $\Delta_{k}=\theta_{0 k}-\theta_{k}$. If we only use the $k$-th feature, that is, we reject $\theta=\theta_{0}$ if $\left|\bar{X}_{k}-\bar{Y}_{k}\right|>\eta_{k}$, then the Type II error probability is

$$
p_{1}:=\mathbb{P}\left(\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta_{k}\right)=\mathbb{P}\left(\left|N\left(\Delta_{k},\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}\right)\right| \leq \eta_{k}\right) .
$$

If we use all the $K$ features, that is, we reject $\theta=\theta_{0}$ if $\exists k,\left|\bar{X}_{k}-\bar{Y}_{k}\right|>\eta_{k}^{\prime}$, then the Type II error probability is

$$
p_{2}:=\mathbb{P}\left(\forall k,\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta_{k}^{\prime}\right) .
$$

Theorem 6.12. We suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $N\left(\theta_{0}, \Sigma\right)$ and $Y_{1}^{\theta}, \cdots, Y_{n}^{\theta} \in \mathbb{R}^{K}$ is an i.i.d. simulated sample from $N(\theta, \Sigma) . p_{1}$ and $p_{2}$ are respectively the Type II error probability of only using the $k$-th feature and using all the $K$ features. If $\Delta_{k} \neq 0$, then for fixed $K>1$, as $N, n \rightarrow \infty$, both $p_{1}$ and $p_{2}$ converge to 0 exponentially in $N$ and $n$. If further we have $N=\omega(\log K)$ and $n=\omega(\log K)$ as $K \rightarrow \infty$, then $p_{2} \rightarrow 0$.

Theorem 6.12 shows that in this more general setting, the Type II error probability of either using only one feature or using all the features decays exponentially in $N$ and $n$. Moreover, if the number of features $K$ is also growing, the Type II error probability of using all the features still converges to 0 as long as $N$ and $n$ grow in a higher order than $\log K$.

Though the above theorems are developed for the special case of Gaussian distribution, they convey some information regarding the conservativeness of using more features. Intuitively, in the general case we may apply CLT and then the sample means are approximately Gaussian for sufficiently large sample sizes. Our conclusion is that even though in some cases using more features makes the method more conservative, overall we can still get an acceptable Type II error probability.

## Ellipsoidal Sample Mean Difference (ESMD).

Now we consider further aggregating the sample mean difference of each feature with sum of squares. We say that $\theta$ is eligible if

$$
\sum_{k=1}^{K}\left(\bar{X}_{k}-\bar{Y}_{k}\right)^{2} \leq \eta
$$

where $\eta$ is the $(1-\alpha)$-quantile of the generalized chi-square distribution given by $\left(\frac{1}{N}+\frac{1}{n}\right) Z^{T} \Sigma_{X} Z$ with $Z \sim N\left(0, I_{K}\right)$ and $\Sigma_{X}$ being the covariance matrix of $X_{1}$. In practice, $\Sigma_{X}$ could be esti-
mated using the samples and the quantile could be numerically evaluated. Again, the approximate correctness is implied by CLT. We have the following theorem:

Theorem 6.13. Suppose that $X_{1}, \cdots, X_{N} \in \mathbb{R}^{K}$ is an i.i.d. true sample from $P^{\theta_{0}}$ and $Y_{1}^{\theta_{0}}, \cdots, Y_{n}^{\theta_{0}} \in$ $\mathbb{R}^{K}$ is an i.i.d. simulated sample from $P^{\theta_{0}}$. For any $k=1, \ldots, K$, we denote $\bar{X}_{k}=\frac{1}{N} \sum_{j=1}^{N} X_{j, k}, \bar{Y}_{k}=$ $\frac{1}{n} \sum_{i=1}^{n} Y_{i, k}^{\theta_{0}}$. Assume that the covariance matrix of $X_{1}$ exists and denote it as $\Sigma_{X}$. Use $\hat{\Sigma}$ to denote the sample covariance matrix of $X_{j}$ 's. If $n /(n+N) \rightarrow \rho$ where $0 \leq \rho \leq 1$ as $n, N \rightarrow \infty$, then

$$
\liminf _{n, N \rightarrow \infty} \mathbb{P}\left(\sum_{k=1}^{K}\left(\bar{X}_{k}-\bar{Y}_{k}\right)^{2} \leq \eta\right) \geq 1-\alpha
$$

where $\eta$ is the $(1-\alpha)$-quantile of the generalized chi-square distribution given by $\left(\frac{1}{N}+\frac{1}{n}\right) Z^{T} \hat{\Sigma} Z$ with $Z \sim N\left(0, I_{K}\right)$.

Theorem 6.13 gives a guarantee on Type I error. Regarding Type II error or conservativeness, we give an example to show that including or excluding more features do not lead to generally dominant results.

Suppose we have a random sample $X_{1}, \ldots, X_{N} \sim N\left(\theta_{0}, I_{K}\right)$ and a random sample $Y_{1}^{\theta}, \ldots, Y_{n}^{\theta} \sim$ $N\left(\theta, I_{K}\right)$ where $\theta_{0}, \theta \in \mathbb{R}^{K}$. If $\theta=\theta_{0}$, then $\bar{X}_{k}-\bar{Y}_{k} \sim N\left(0, \frac{1}{N}+\frac{1}{n}\right)$ and $\sum_{k=1}^{K}\left(\bar{X}_{k}-\bar{Y}_{k}\right)^{2} /\left(\frac{1}{N}+\frac{1}{n}\right) \sim \chi_{K}^{2}$. Given a confidence level $1-\alpha$, we define $\eta$ and $\eta^{\prime}$ by

$$
\mathbb{P}\left(\left|N\left(0, \frac{1}{N}+\frac{1}{n}\right)\right| \leq \eta\right)=1-\alpha, \mathbb{P}\left(\chi_{K}^{2} \leq \frac{\eta^{\prime}}{\frac{1}{N}+\frac{1}{n}}\right)=1-\alpha .
$$

Now we suppose that $\theta \neq \theta_{0}$. For any $k=1, \ldots, K$, we denote $\Delta_{k}=\theta_{0 k}-\theta_{k}$. If we only use the $k$-th feature, that is, we reject $\theta=\theta_{0}$ if $\left|\bar{X}_{k}-\bar{Y}_{k}\right|>\eta$, then the Type II error probability is

$$
p_{1}:=\mathbb{P}\left(\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta\right)=\mathbb{P}\left(\left|N\left(\Delta_{k}, \frac{1}{N}+\frac{1}{n}\right)\right| \leq \eta\right) .
$$

If we use all the features, that is, we reject $\theta=\theta_{0}$ if $\sum_{k=1}^{K}\left(\bar{X}_{k}-\bar{Y}_{k}\right)^{2}>\eta^{\prime}$, then the Type II error
probability is

$$
p_{2}:=\mathbb{P}\left(\sum_{k=1}^{K}\left(\bar{X}_{k}-\bar{Y}_{k}\right)^{2} \leq \eta^{\prime}\right)=\mathbb{P}\left(\chi_{K, v}^{2} \leq \frac{\eta^{\prime}}{\frac{1}{N}+\frac{1}{n}}\right)
$$

where

$$
v=\frac{\sum_{k=1}^{K} \Delta_{k}^{2}}{\frac{1}{N}+\frac{1}{n}} .
$$

Given the value of $\alpha, k, K, n, N$ and $\Delta_{k}$ 's, we are able to numerically compute $p_{1}$ and $p_{2}$. Below are some results from the numerical computation.

First, we set $\alpha=0.05, k=1, K=10$ and $\Delta=0.1 *(1,1, \ldots, 1)$. We also let $n=N$. Then the change of Type II error probabilities with the sample size is shown in Figure 6.6a. We find that it is better to use all the features in this case. Next, we set $\Delta=0.1 *(1,0, \ldots, 0)$. The result is shown in Figure 6.6b. We find that in this case it is better to only use the first feature. The results are interpretable. In the first case, the marginal distributions in each dimension are different for the two samples, so using more features help us detect the wrong parameter. In the second case, however, the distributions are the same except for the first dimension, so including more features in turn makes it more difficult to distinguish the parameters. Therefore, there is no certain conclusion that one choice is better than the other.


Figure 6.6: Change of Type II error probabilities with the sample size. $p_{1}$ : Single-feature type II error probability. $p_{2}$ : All-feature type II error probability.

We are also interested in how the Type II error probability will change with the dimension $K$. We fix $n=N=1000$ and the results are shown in Figure 6.7a and 6.7b. From the figures, we
see that in both cases, the Type I error probability of only using the first feature does not change with the number of features. In the first case, the Type II error probability of using all the features decreases as the number of features grows. As explained, each dimension contributes to detecting the wrong parameter, so it is easier with more features. By comparison, in the second case, the Type II error probability increases as the number of features grows, as the additional features do not help us distinguish the parameters but nullify the difference in the first dimension. This result further supports that neither of the methods dominates the other one.


Figure 6.7: Change of Type II error probabilities with the number of features. $p_{1}$ : Single-feature type II error probability. $p_{2}$ : All-feature type II error probability.

### 6.5 Multi-Agent Simulator Calibration

In this section, we apply our feature-based calibration framework in Section 6.4 to the AgentBased Interactive Discrete-Event Simulation (ABIDES) environment [194, 195] which, among other uses, can be used to simulate limit order book exchange markets such as NASDAQ or New York Stock Exchange. ABIDES provides a selection of background agent types (such as market makers, noise agents, value agents, etc.), a NASDAQ-like exchange agent which lists any number of securities for trade against a limit order book with price-then-FIFO matching rules, and a simulation kernel which manages the flow of time and handles all inter-agent communication.

We use background agent implementation introduced in [195]. Value agents are designed to simulate the actions of fundamental traders that trade according to their belief of the exogenous
value of a stock (also called fundamental price - which in our model follows a mean reverting price time series). Value traders choose to buy or sell a stock depending on whether it is cheap or expensive relative to their noisy observation of a fundamental price. The fundamental follows a discrete-time mean-reverting Ornstein-Uhlenbeck process [248]. Noise agents are designed to emulate the actions of consumer agents who trade on demand without any other considerations; they arrive to the market at times that are uniformly distributed throughout the trading day and place an order of random size in random direction. Market makers act as liquidity providers by placing limit orders on both sides of the limit order book with a constant arrival rate.

We test our calibration algorithm on ABIDES against a synthetic configuration that is designated as a ground truth. Our experimental analysis consists of two parts. The first is comparison of different feature extraction and aggregation techniques in deciding the acceptance/rejection of the eligibility of configurations. The second part is the study of how the eligible configurations conform with the stylized facts, which are the important realism metrics in financial markets. Stylized fact similarity to the ground truth for the accepted configurations indicates that they are indeed more "realistic" than the rejected models.

### 6.5.1 Comparisons of Feature Extraction and Aggregation Techniques

We first test the various approaches in Section 6.4 on a specific example. Then we vary the number of input parameters and conduct ablation studies on our training methods.

## A Basic Example.

We generate limit order book time series that result from simulating the market from 9:30 to 10:00 AM with configurations that consist of an exchange agent, a market maker agent, $n_{\text {noise }}$ noise agents and $n_{\text {value }}$ value agents that each follows a noisy observation of an Orstein-Uhlenbeck fundamental with the mean reversion rate $\kappa$, mean fundamental value $\bar{r}$ and arrival rate $\lambda_{a}$ (defined as in [248]). The representative time series is taken as the log returns of the resultant mid-price sampled at every second, $r_{t}:=\log \left(m_{t+1} / m_{t}\right)$, where $m_{t}$ is the mid price at second $t$ - hence, in our
experiment, each data point from which to extract relevant features is a time series of length 1799 .
We run $N=100,000$ simulations for each configuration. The feature extraction network architectures are shown in Tables 6.3, 6.4 and 6.5 in Section 6.8. The parameters that we aim to calibrate are $n_{\text {noise }}, n_{\text {value }}, \bar{r}, \kappa$ and $\lambda_{a}$.

We compare combinations of the five feature extraction methods: autoencoder hidden layer, GAN hidden layer, WGAN hidden layer, GAN discriminator output and WGAN critic output, and three aggregation methods: SKS, SSMD and ESMD. We set the confidence level $1-\alpha=0.95$. Figures 6.8, 6.9 and 6.10 summarize the calibration results of SKS, SSMD and ESMD respectively. In all figures, Models 1 to 17 in the $x$-axis refer to 17 different configurations on ABIDES (see Table 6.2 for the details). Model 1 represents the ground truth configuration. Models 2 to 5 are the configurations close to the truth, because their parameters are chosen with smallest perturbations from the ground truth. The remaining models refer to configurations with parameters that differ significantly. In Figure 6.8, we use the value of the SKS distance as the $y$-axis, so that an SKS value below the cutoff line of the corresponding aggregation method indicates eligibility. The accepted configurations are marked as a circle, and the rejected configurations are shown without any markers. In Figures 6.9 and 6.10, we convert the scale of SSMD and ESMD into a " $p$-value" on the $y$-axis under a normal approximation, as discussed in Section 6.4.2. The " $p$-value" here measures how often the distance is smaller than the critical value. Correspondingly, a dot value above the cutoff 0.05 indicates eligibility. This $p$-value scaling for SSMD and ESMD serves to facilitate easier visual comparison than in the original scale.

From these results, we see that SKS exhibits the most reasonable trend across the board, where configurations close to the ground truth are classified as eligible while others are not. In comparison, with SSMD and ESMD, some methods either cannot recover the truth or misclassify ineligible configurations as eligible. Even though some extraction methods with SSMD or ESMD are able to recover the truth, they also classify many of the other configurations as eligible. Thus, with the same feature extraction techniques, SSMD and ESMD appear to underperform compared to SKS in terms of conservativeness.


Figure 6.8: Comparisons of different feature extraction methods with SKS. Eligible configurations are the ones below the thresholds (shown in dash lines) and plotted with dots.


Figure 6.9: Comparisons of different feature extraction methods with SSMD. Eligible configurations are the ones above the thresholds (shown in dash lines).


Figure 6.10: Comparisons of different feature extraction methods with ESMD. Eligible configurations are the ones above the thresholds (shown in dash lines).

## Calibration with Different Numbers of Model Parameters.

The discussion above indicates that combining SKS with different feature extraction methods outperforms other aggregation approaches, not only in recovering the truth but also in eliminating false candidates. This subsection aims to investigate the robustness of this observation. In particular, we investigate the performance of SKS in calibrating different numbers of model parameters.

Table 6.1 shows five sets of parameters to be calibrated. For each set, we investigate a total of 20 configuration candidates. Here, the data we calibrate against is the trading mid-price - the setting is otherwise the same. Moreover, to fairly compare the performance of different feature extractors, from now on we choose more similar network architectures as in Tables 6.6 to 6.8.

| Case No. | 1 | 2 | 3 | 4 | 5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Parameters | $n_{\text {value }}$ | $\kappa, \lambda_{a}$ | $n_{\text {value }}, \kappa, \lambda_{a}$ | $n_{\text {value }}, \bar{r}, \kappa, \lambda_{a}$ | $n_{\text {value }}, n_{\text {noise }}, \bar{r}, \kappa, \lambda_{a}$ |

Table 6.1: Parameters to calibrate in each case


Figure 6.11: Comparisons of different feature extraction method with SKS statistics. The annotations over the bars denote whether the true case can be recovered. " T " means true and " F " means false.

Figure 6.11 shows the calibration results on 1 to 5 parameter sets. If a method recovers the true parameter and simultaneously maintains a small eligibility set (among the 20 configurations),
then it is viewed as superior. In terms of the first metric, correctness, all of the methods are able to recover the truth. In terms of the second metric, conservativeness, GAN hidden features appears to be the best, since it consistently achieves the minimal eligibility set size while able to uncover the truth. Autoencoder follows next by achieving the smallest set size in 3 cases and the second smallest set size in 2 cases. WGAN hidden features performs a bit more conservatively than autoencoder because in almost all cases, WGAN hidden feature is able to attain the smallest or the second smallest set size except for case 1 . GAN discriminator output is also slightly more conservative than those methods mentioned before, and WGAN critic output is the most conservative one among all methods. Note that the calibration performance comparisons in terms of correctness and conservativeness among different methods are similar within each case, which further showcases the robustness of our methodologies.

Therefore, in order to achieve a low conservativeness, we would first recommend using GAN hidden layer as a feature extraction method, followed by the autoencoder then WGAN hidden layer. GAN discriminator output or WGAN critic output are less recommended because of their over-conservativeness as compared with the first three methods.

## Ablation Study.

We further study the performances of different approaches when the number of extracted features, $K$, varies. The magnitude of $K$ is controlled by the size of the final hidden layer in each neural network. We vary the size of this layer to $15,29,57,113,225$, and 450 . The network architecture can be adjusted by adding or removing a block of 〈Convolutional, Leaky ReLU, Dropout〉 layers in order to increase or decrease the network size by a factor of 2 . In each setting, we measure the Type I error probability and $q-$ SKS over 500 experiments, where $q$ is calculated as the critical value of SKS. Intuitively, the magnitude of the quantity $q$ - SKS measures the gap between the SKS statistic and the threshold $q$. The larger the gap is, the more configurations can be considered as eligible, meaning that the method is more conservative.

Figures 6.12 and 6.13 present the results of the ablation study. First, we do not observe a clear
trend of Type I error probability as the number of extracted features changes, and it is not obvious to see any methods consistently outperforming others. For $q-$ SKS, there is no clear pattern to distinguish its performance either. Nonetheless, all methods have Type I error probabilities less than or equal to the nominal level $\alpha=0.05$. Furthermore, all methods constantly obtain a positive average value of $q-$ SKS, meaning that the ground truth can still be recovered even with different hidden feature dimensions. These two observations once again verify the power of our methodology.


Figure 6.12: Type I error probability of Autoencoder, GAN hidden features extraction, WGAN hidden features extraction combined with SKS statistics as the dimension changes.

### 6.5.2 Realism Metrics

Properties of limit order book market behavior that are repeated across a wide range of instruments, markets, and time periods are referred to as stylized facts [249, 250, 251]. Evaluating the statistical properties of simulated asset returns, order volumes, order arrival times, order cancellations, etc and comparing them to those generated from real historical data allows us to infer the level of fidelity of a simulation; hence, stylized facts can be used as simulated realism metrics [195]. In this subsection, we investigate how our eligible models can match the ground truth in terms of these metrics.


Figure 6.13: The difference between critical value and SKS of Autoencoder, GAN hidden features extraction, WGAN hidden features extraction combined with SKS statistics as the dimension changes. To get a sense of the estimation uncertainty, we also show shaded areas which have width 0.2 times the standard deviation.

We investigate the following metrics. Denote the midprice at time $t$ as $m_{t}$, and the log return at time scale $\Delta t$ as $r_{t, \Delta t}=\log m_{t+\Delta t}-\log m_{t}$ :

- Heavy tails and aggregational normality: The distribution of the asset prices exhibits a fat tail. However, as $\Delta t$ increases, the distribution tends to show a slimmer tail, and more like a normal distribution.
- Absence of autocorrelation: The autocorrelation function $\operatorname{corr}\left(r_{t+\tau, \Delta t}, r_{t, \Delta t}\right)$ becomes insignificant as $\tau$ gets longer than 20 minutes.
- Volatility clustering: High-volatility events usually cluster in time. The autocorrelation function $\operatorname{corr}\left(r_{t+\tau, \Delta t}^{2}, r_{t, \Delta t}^{2}\right)$ is used to measure volatility clustering. Empirical results on various equities indicate that this quantity remains positive over several days.

We examine the realism metrics of the 17 configurations studied in Section 6.5.1. The simulation sample size for each configuration is 100,000 . For the first metric, we examine minutely $\log$ return; for the second metric, we examine the autocorrelation of minutely $\log$ returns and take
$\Delta t$ to be 25 minutes; for the third metric, we examine the 10 -second log return and the lags of the autocorrelation function ranging from 1 to 10 .

Figures $6.14,6.15$ and 6.16 show the realism metrics exhibited by the 17 configurations. Recall that, according to Figure 6.8, the three eligible configurations under our framework are Model 1 (accepted by all methods), Model 2 (accepted only by WGAN critic output with SKS), and Model 3 (accepted only by GAN discriminator output with SKS). Model 1, in particular, is the true configuration. We observe that the distribution of the autocorrelation function value at $\Delta=20$ minutes has most of the mass around 0 . Therefore, from this perspective, the true configuration follows the empirical rule of the market. For volatility clustering, the average autocorrelation function of squared returns among 100, 000 samples decays as the lag (based on 10 seconds) increases but remains positive. However, we have also found that the autocorrelation function can become negative as we gradually increase the length of the lag, thus deviating from the empirical rule in this regime. Nonetheless, the true configuration still loosely follows the empirical rule when the lag is not too large.

In terms of the similarity between the ground truth and the simulated configurations, we see that Models 1 and 3 behave almost the same as the truth in all three metrics. Model 2 also behaves similarly in log return autocorrelation and volatility clustering (even closer to the truth compared with Model 3 in these two metrics), but it has a slimmer tail in the $\log$ return distribution as shown in Figure 6.14. On the other hand, other models acting statistically like the truth include Models 4, 5, and 12. In detail, Model 4 is similar but slightly different from the truth in all these three metrics. Models 5 and 12 are close to the truth in the log return distribution and the log return autocorrelation, but deviate greatly in volatility clustering. For all the other ineligible configurations, their realism metric performances are disparate from the truth in more than two of the metrics. Some examples of this are Models 14 and 9. Model 14 is similar to the truth in log return distribution, but distinguishable in the log return autocorrelation and volatility clustering, and Model 9 is similar to the truth in log return autocorrelation, but bears a substantial difference in the log return distribution and volatility clustering. In summary, the accepted models are able to
match the truth in terms of realism metrics to a great extent, while the rejected models either have a lot of difference in one of the metrics or have differences in two or more metrics.


Figure 6.14: One-minute log return distributions. Accepted models are the first three in the first row, which are shown in red.

### 6.6 Supplementary A: Proofs of Theorems

Proof of Theorem 6.1. Since

$$
\begin{aligned}
\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right| & \leq \sup _{x \in \mathbb{R}}\left(\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|+\left|F^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right|\right) \\
& \leq \sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|+\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|,
\end{aligned}
$$



Figure 6.15: Distributions of return autocorrelation at 25 minutes. Accepted models are the first three in the first row, which are shown in red.
we get that

$$
\begin{aligned}
& \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right|>q_{1-\alpha} / \sqrt{N}\right) \\
& \leq \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|+\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>q_{1-\alpha} / \sqrt{N}\right) \\
& \leq \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>\lambda q_{1-\alpha} / \sqrt{N}\right)+\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>(1-\lambda) q_{1-\alpha} / \sqrt{N}\right)
\end{aligned}
$$

for any $\lambda \in(0,1)$. It is known that $\sqrt{n} \sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right| \Rightarrow \sup _{t \in \mathbb{R}}\left|B B\left(F^{\theta_{0}}(t)\right)\right|$ and similarly $\sqrt{N} \sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right| \Rightarrow \sup _{t \in \mathbb{R}}\left|B B\left(F^{\theta_{0}}(t)\right)\right|$ as $n, N \rightarrow \infty$ where $\Rightarrow$ stands for


Figure 6.16: Average autocorrelation of squared returns as a function of time lag from 1 to 10. Squared returns are estimated for every 10 seconds. Accepted models are the first three in the first row, which are shown in red.
convergence in distribution. $n=\omega(N)$ as $N \rightarrow \infty$ implies that $n / N \rightarrow \infty$ as $N \rightarrow \infty$, and thus

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>\lambda q_{1-\alpha} / \sqrt{N}\right)=\mathbb{P}\left(\sqrt{n} \sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>\lambda q_{1-\alpha} \sqrt{n / N}\right) \rightarrow 0
$$

as $N \rightarrow \infty$ for any $\lambda \in(0,1)$. Hence,

$$
\begin{aligned}
& \limsup _{N \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right|>q_{1-\alpha} / \sqrt{N}\right) \\
\leq & \limsup _{N \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>(1-\lambda) q_{1-\alpha} / \sqrt{N}\right) \\
= & \mathbb{P}\left(\sup _{t \in \mathbb{R}}\left|B B\left(F^{\theta_{0}}(t)\right)\right|>(1-\lambda) q_{1-\alpha}\right) \\
\leq & \mathbb{P}\left(\sup _{t \in[0,1}|B B(t)|>(1-\lambda) q_{1-\alpha}\right) .
\end{aligned}
$$

By the arbitrariness of $\lambda$ and the definition of $q_{1-\alpha}$, we finally get that

$$
\limsup _{N \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta_{0}}(x)-F_{N}^{\theta_{0}}(x)\right|>q_{1-\alpha} / \sqrt{N}\right) \leq \mathbb{P}\left(\sup _{t \in[0,1]}|B B(t)| \geq q_{1-\alpha}\right)=\alpha
$$

Proof of Theorem 6.2. We know that

$$
\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \geq\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|-\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|,
$$

and then we get that

$$
\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \geq \sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|-\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right| .
$$

Therefore,

$$
\begin{aligned}
& \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \leq q_{1-\alpha} / \sqrt{N}\right) \\
& \leq \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|-\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right| \leq q_{1-\alpha} / \sqrt{N}\right) \\
& \leq \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|>\varepsilon_{1}\right)+\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon_{2}\right) .
\end{aligned}
$$

The last inequality is obtained using the fact that $q_{1-\alpha} / \sqrt{N}<\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\varepsilon_{1}-\varepsilon_{2}$ under
the conditions in the theorem. By the refined Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [252], we get that

$$
\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|>\varepsilon_{1}\right) \leq 2 e^{-2 n \varepsilon_{1}^{2}}, \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon_{2}\right) \leq 2 e^{-2 N \varepsilon_{2}^{2}}
$$

which concludes the proof.

Proof of Theorem 6.3. By applying Theorem 6.2, we have that

$$
\begin{aligned}
& \mathbb{P}\left(\exists i=1, \cdots, m \text { s.t. } \sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon, \theta_{i} \in \hat{\mathcal{E}}_{m}\right) \\
\leq & \sum_{i=1}^{m} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F^{\theta_{i}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon, \theta_{i} \in \hat{\mathcal{E}}_{m}\right) \\
\leq & 2 m\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right) .
\end{aligned}
$$

Proof of Corollary 6.1. For any $\varepsilon>0$, we pick $\varepsilon_{1}=\varepsilon_{2}=\varepsilon / 3$. If $\log m=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then

$$
2 m\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right)=2\left(e^{\log m-2 n \varepsilon_{1}^{2}}+e^{\log m-2 N \varepsilon_{2}^{2}}\right) \rightarrow 0
$$

as $N \rightarrow \infty$. By applying Theorem 6.3, we prove the corollary.

Proof of Corollary 6.2. Let $\varepsilon=\sqrt{\log m / m}$ and $\varepsilon_{1}=\varepsilon_{2}=\varepsilon / 3$. If $m=o(N)$ and $n=\Omega(N)$ as $N \rightarrow \infty$, then

$$
2 m\left(e^{-2 n \varepsilon_{1}^{2}}+e^{-2 N \varepsilon_{2}^{2}}\right)=2\left(e^{\log m-2 n \log m /(9 m)}+e^{\log m-2 N \log m /(9 m)}\right) \rightarrow 0
$$

as $N \rightarrow \infty$. By applying Theorem 6.3, we prove the corollary.

Proof of Theorem 6.4. See Theorem 1 in [253].

Proof of Theorem 6.5. Follow the proof of Theorem 6.2. We could get that

$$
\begin{aligned}
& \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F_{N}^{\theta_{0}}(x)\right| \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log (\alpha / 2)}\right) \\
\leq & \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F^{\theta}(x)-F^{\theta_{0}}(x)\right|-\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|-\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right| \leq \sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log (\alpha / 2)}\right) \\
\leq & \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n}^{\theta}(x)-F^{\theta}(x)\right|>\varepsilon_{1}\right)+\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{N}^{\theta_{0}}(x)-F^{\theta_{0}}(x)\right|>\varepsilon_{2}\right) .
\end{aligned}
$$

where the last inequality is obtained using the fact that $\left.\sqrt{\frac{n+N}{n N}} \sqrt{-\frac{1}{2} \log (\alpha / 2)}<\sup _{x \in \mathbb{R}} \right\rvert\, F^{\theta}(x)-$ $F^{\theta_{0}}(x) \mid-\varepsilon_{1}-\varepsilon_{2}$ under the conditions in the theorem. Then we apply the refined DKW inequality to conclude the proof.

Proof of Theorem 6.6. Follow the proof of Theorem 6.3 except that we apply Theorem 6.5 here.

Proof of Corollary 6.3. Follow the proof of Corollary 6.1 except that we apply Theorem 6.6 here.

Proof of Corollary 6.4. Follow the proof of Corollary 6.2 except that we apply Theorem 6.6 here.

Proof of Theorem 6.7. From Theorem 6.1, we know that for any $k$,

$$
\limsup _{n, N \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right|>q_{1-\alpha / K} / \sqrt{N}\right) \leq \alpha / K
$$

and thus

$$
\limsup _{n, N \rightarrow \infty} \mathbb{P}\left(\exists k, \sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right|>q_{1-\alpha / K} / \sqrt{N}\right) \leq \alpha
$$

Proof of Theorem 6.8. Define $\lambda=(N / n)^{1 / 4} /(\log K)^{1 / 2}$. As $K \rightarrow \infty, \lambda=o(1 / \log K)$ and $\lambda=$
$\omega(\sqrt{N / n})$. Without loss of generality, we assume that $0<\lambda<1$. We have that

$$
\begin{aligned}
& \limsup _{K \rightarrow \infty} \mathbb{P}\left(\exists k, \sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right|>q_{1-\alpha / K} / \sqrt{N}\right) \\
\leq & \limsup _{K \rightarrow \infty} \sum_{k=1}^{K} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right|>q_{1-\alpha / K} / \sqrt{N}\right) \\
\leq & \limsup _{K \rightarrow \infty} \sum_{k=1}^{K}\left(\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{k}^{\theta_{0}}(x)\right|>\lambda q_{1-\alpha / K} / \sqrt{N}\right)\right. \\
& \left.+\mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{N, k}^{\theta_{0}}(x)-F_{k}^{\theta_{0}}(x)\right|>(1-\lambda) q_{1-\alpha / K} / \sqrt{N}\right)\right) \\
\leq & \limsup _{K \rightarrow \infty} \sum_{k=1}^{K} 2\left(e^{-2 \lambda^{2} q_{1-\alpha / K}^{2} n / N}+e^{-2(1-\lambda)^{2} q_{1-\alpha / K}^{2}}\right) \\
= & \limsup _{K \rightarrow \infty} 2 K\left(e^{-2 \lambda^{2} q_{1-\alpha / K}^{2} n / N}+e^{-2(1-\lambda)^{2} q_{1-\alpha / K}^{2}}\right) .
\end{aligned}
$$

We know that $\alpha=2 \sum_{v=1}^{\infty}(-1)^{v-1} e^{-2 v^{2} q_{1-\alpha}^{2}}$, and hence $\alpha /\left(2 K e^{-2 q_{1-\alpha / K}^{2}}\right) \rightarrow 1$ as $K \rightarrow \infty$. Thus, $2 q_{1-\alpha / K}^{2}-\log K \rightarrow-\log (\alpha / 2)$ as $K \rightarrow \infty$. We have that

$$
K e^{-2 \lambda^{2} q_{1-\alpha}^{2} n / N}=e^{\log K-\left(2 q_{1-\alpha / K}^{2}-\log K\right) \sqrt{n / N} / \log K-\sqrt{n / N}} \rightarrow 0
$$

and

$$
K e^{-2(1-\lambda)^{2} q_{1-\alpha / K}^{2}}=e^{\log K-(1-\lambda)^{2}\left(2 q_{1-\alpha / K}^{2}-\log K\right)-(1-\lambda)^{2} \log K} \rightarrow \alpha / 2
$$

Therefore,

$$
\limsup _{K \rightarrow \infty} \mathbb{P}\left(\exists k, \sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta_{0}}(x)-F_{N, k}^{\theta_{0}}(x)\right|>q_{1-\alpha / K} / \sqrt{N}\right) \leq \alpha
$$

Proof of Theorem 6.9. Suppose that $k^{*}=\arg \max _{1 \leq k \leq K} \sup _{x \in \mathbb{R}}\left|F_{k}^{\theta}(x)-F_{k}^{\theta_{0}}(x)\right|$. Then we have
that

$$
\begin{aligned}
& \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k}^{\theta}(x)-F_{N, k}^{\theta_{0}}(x)\right| \leq q_{1-\alpha / K} / \sqrt{N}, \forall 1 \leq k \leq K\right) \\
\leq & \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{n, k^{*}}^{\theta}(x)-F_{N, k^{*}}^{\theta_{0}}(x)\right| \leq q_{1-\alpha / K} / \sqrt{N}\right) .
\end{aligned}
$$

By applying Theorem 6.2, we get the result.

Proof. Proof of Theorem 6.10 It is well known that by CLT,

$$
\frac{\bar{X}_{k}-\bar{Y}_{k}}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \operatorname{var}\left(X_{1, k}\right)}} \Rightarrow N(0,1)
$$

Then by Slutsky's theorem,

$$
\frac{\bar{X}_{k}-\bar{Y}_{k}}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \widehat{v a r}_{k}}} \Rightarrow N(0,1)
$$

which concludes the proof.

Proof of Theorem 6.11. We use $\Phi$ to denote the CDF of standard normal distribution and denote $q_{1}=\Phi^{-1}\left(1-\frac{\alpha}{2}\right), q_{2}=\Phi^{-1}\left(1-\frac{\alpha}{2 K}\right)$. Then $\eta=q_{1} \sqrt{\frac{1}{N}+\frac{1}{n}}$ and $\eta^{\prime}=q_{2} \sqrt{\frac{1}{N}+\frac{1}{n}}$. We have that

$$
p_{1}=\Phi\left(q_{1}-\frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}\right)-\Phi\left(-q_{1}-\frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}\right)
$$

and

$$
p_{2}=\left(1-\frac{\alpha}{K}\right)^{K-1}\left(\Phi\left(q_{2}-\frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}\right)-\Phi\left(-q_{2}-\frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}\right)\right) .
$$

For fixed $K>1$ (which implies $q_{1}<q_{2}$ ), as $N, n \rightarrow \infty$, we have that

$$
\frac{p_{2}}{p_{1}} \sim\left(1-\frac{\alpha}{K}\right)^{K-1} \frac{-q_{1}+\frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}}{-q_{2}+\frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}} \exp \left(\left(q_{2}-q_{1}\right) \frac{\Delta_{1}}{\sqrt{\frac{1}{N}+\frac{1}{n}}}+\frac{q_{1}^{2}-q_{2}^{2}}{2}\right),
$$

which grows exponentially in $N$ and $n$.

Proof of Theorem 6.12. We have that

$$
p_{2}:=\mathbb{P}\left(\forall k,\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta_{k}^{\prime}\right) \leq \mathbb{P}\left(\left|\bar{X}_{k}-\bar{Y}_{k}\right| \leq \eta_{k}^{\prime}\right)=\mathbb{P}\left(\left|N\left(\Delta_{k},\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}\right)\right| \leq \eta_{k}^{\prime}\right)=: p_{2}^{\prime} .
$$

We still use the notations of $q_{1}$ and $q_{2}$. Then $\eta_{k}=q_{1} \sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}$ and $\eta_{k}^{\prime}=q_{2} \sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}$. We have that

$$
p_{1}=\Phi\left(q_{1}-\frac{\Delta_{k}}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}}\right)-\Phi\left(-q_{1}-\frac{\Delta_{k}}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}}\right)
$$

and

$$
p_{2}^{\prime}=\Phi\left(q_{2}-\frac{\Delta_{k}}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}}\right)-\Phi\left(-q_{2}-\frac{\Delta_{k}}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}}\right) .
$$

Clearly, if $\Delta_{k} \neq 0$, then for fixed $K>1$, as $N, n \rightarrow \infty$, both $p_{1}$ and $p_{2}^{\prime}$ converge to 0 exponentially in $N$ and $n$. Compared to $p_{2}^{\prime}, p_{1}$ decreases exponentially faster.

Now we only analyze $p_{2}^{\prime}$. It is known that as $K \rightarrow \infty, q_{2}=O(\sqrt{\log K})$. If $\Delta_{k} \neq 0$ and $N=\omega(\log K), n=\omega(\log K)$ as $K \rightarrow \infty$, then

$$
p_{2}^{\prime} \leq \bar{\Phi}\left(-q_{2}+\frac{\left|\Delta_{k}\right|}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right) \Sigma_{k k}}}\right) \rightarrow 0 .
$$

Proof of Theorem 6.13. Use $\bar{X}$ and $\bar{Y}$ to denote the sample mean vectors. It is well known that

$$
\frac{(\bar{X}-\bar{Y})}{\sqrt{\left(\frac{1}{N}+\frac{1}{n}\right)}} \Rightarrow N\left(0, \Sigma_{X}\right)
$$

and hence

$$
\frac{\sum_{k=1}^{K}\left(\bar{X}_{k}-\bar{Y}_{k}\right)^{2}}{\frac{1}{N}+\frac{1}{n}}=\frac{(\bar{X}-\bar{Y})^{T}(\bar{X}-\bar{Y})}{\frac{1}{N}+\frac{1}{n}} \Rightarrow Z^{T} \Sigma_{X} Z
$$

Since $\hat{\Sigma} \rightarrow \Sigma_{X}$ almost surely, we get that $\frac{\eta}{\frac{1}{N}+\frac{1}{n}}$ converges to the $(1-\alpha)$-quantile of $Z^{T} \Sigma_{X} Z$ almost surely, which concludes the proof.

### 6.7 Supplementary B: Details on Machine Learning Models and Feature Extraction

Auto-encoder: Auto-encoder is a unsupervised learning method [240]. It is made up of two components, an encoder $e: \mathcal{X} \rightarrow \mathcal{H}_{X}$, and an decoder $r: \mathcal{H}_{X} \rightarrow \mathcal{X}$. Given a target dimension $K$, encoder $e$ finds the hidden feature $H \in \mathbb{R}^{K}$ of the input $X \in \mathcal{X}$, and afterward decoder $r$ reconstructs the hidden feature $H$ and outputs the reconstruction $\hat{X} \in \mathcal{X}$. The bottleneck structure allows the network to find out the hidden feature of the input by itself, and the training procedure minimizes the Euclidean distance between the input samples and the reconstructed samples. The training of the network can be summarized as follows:

$$
\min _{\theta, \hat{\theta}} \sum_{X}\|r(e(X ; \theta) ; \hat{\theta})-X\|_{2}^{2}
$$

where $\theta, \hat{\theta}$ are the weights of encoder $e$ and decoder $r$.

GAN: GAN is another type of unsupervised learning method [241]. There are two adversarial modules contained in the GAN structure, which are discriminator $D$ and generator $G$. Given the real inputs $X_{1}, \ldots, X_{N} \in \mathcal{X}$ from the real distribution $p_{\text {real }}$, the GAN network aims at learning the real distribution $p_{\text {real }}$ and outputs samples similar to the real inputs. Discriminator $D$ has input $X \in \mathcal{X}$ and summarizes $X$ to a single value $D(X)$, which measures the probability that input $X$ is a sample from the real distribution. The direct output of discriminator, $D(X)$, is able to serve as a hidden feature of $X$. Otherwise, we can also extract the output from the last hidden layer. This is not as concise as the direct output $D(X)$, but preserves more information about $X$. Generator $G$ has input $z$, as a sample from the latent space $\mathcal{Z}$, and outputs $\hat{X} \in \mathcal{X}$. Similar to two players in game theory, both discriminator $D$ and generator $G$ have their own objectives. The training process is a competition between $D$ and $G$, where $D$ is trained to tell the difference between the real inputs
and the simulated ones, and $G$ is trained to learn the real distribution $p_{\text {real }}$ and fool discriminator $D$. For discriminator $D$, it minimizes the cross entropy,

$$
\mathcal{L}_{D}=-\mathbb{E}_{x \sim p_{\text {real }}} \log D(x)-\mathbb{E}_{z \sim \mathcal{Z}} \log (1-D(G(z))) .
$$

As a competitor against $D$, generator $G$ maximizes the objective of $D$, so the loss function of $G$ is

$$
\mathcal{L}_{G}=-\mathcal{L}_{D}
$$

Therefore, the optimization of GAN can be describe as a minimax game:

$$
\min _{G} \max _{D} \mathbb{E}_{x \sim p_{\text {real }}} \log D(x)+\mathbb{E}_{z \sim z} \log (1-D(G(z)))
$$

WGAN: There are many variants of GAN, and a popular one is WGAN. Specifically, with an optimal discriminator $D^{*}$, the objective of GAN quantifies the similarity between the real input distribution and the generative distribution using the Jensen-Shannon (JS) divergence. However, the JS divergence becomes less sensitive in measuring the statistical distance between distributions when their supports are on low-dimensional manifolds, which further causes the instability in training [242]. Therefore, [242] proposes to adopt Wasserstein-1 distance as a better metric, which is

$$
W\left(p_{\text {real }}, p_{\text {gen }}\right)=\inf _{\gamma \in \Pi\left(p_{\text {real }}, p_{\text {gen }}\right)} \mathbb{E}_{(x, y) \sim \gamma}[\|x-y\|]
$$

where $\Pi\left(p_{\text {real }}, p_{\text {gen }}\right)$ is the set of all possible joint distributions with marginal distributions $p_{\text {real }}$ and $p_{\text {gen }}$. However, due to the intractability of Wasserstein distance calculation, an alternative is to leverage the Kantorovich-Rubinstein duality, where the problem becomes

$$
W\left(p_{\text {real }}, p_{\text {gen }}\right)=\sup _{\|f\|_{L} \leq 1} \mathbb{E}_{x \sim p_{\text {real }}}[f(x)]-\mathbb{E}_{x \sim p_{\text {gen }}}[f(x)] .
$$

In practice, one can always extend the 1-Lipschitz class functions to $L$-Lipschitz and the objective
becomes $L \cdot W\left(p_{\text {real }}, p_{\text {gen }}\right)$. Meanwhile, the $L$-Lipschitz continuity of neural network can be obtained by weight clipping [242]. Therefore, the optimization of WGAN can be described as follows:

$$
\min _{G} \max _{\|D\|_{L} \leq L} \mathbb{E}_{x \sim p_{\text {real }}}[D(x)]-\mathbb{E}_{z \sim}[D(G(z))] .
$$

Since WGAN has a similar architecture to GAN, we can also use the direct output from the critic function or output from the last hidden layer as the hidden feature of the input.

### 6.8 Supplementary C: Experimental Details

| Config. | $m$ | $n$ | $\bar{r}$ | $\kappa$ | $\lambda_{a}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 100 | 1000 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 2 | 105 | 1050 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-14}$ |
| 3 | 90 | 900 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 4 | 70 | 700 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 5 | 95 | 950 | $1.1 \times 10^{5}$ | $1.5 \times 10^{-12}$ | $1.1 \times 10^{-13}$ |
| 6 | 500 | 1000 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 7 | 70 | 700 | $10^{5}$ | $8 \times 10^{-1}$ | $10^{-12}$ |
| 8 | 100 | 1000 | $10^{5}$ | $5 \times 10^{-2}$ | $10^{-12}$ |
| 9 | 200 | 2000 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 10 | 10 | 100 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 11 | 50 | 500 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 12 | 105 | 1050 | $9 \times 10^{4}$ | $1.8 \times 10^{-12}$ | $8 \times 10^{-14}$ |
| 13 | 100 | 3000 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 14 | 10 | 10 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-12}$ |
| 15 | 200 | 1500 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-12}$ |
| 16 | 10 | 10 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-13}$ |
| 17 | 50 | 500 | $10^{5}$ | $1.67 \times 10^{-12}$ | $10^{-11}$ |

Table 6.2: Agent configurations

| Encoder |  |
| :---: | :---: |
| Layer | Network Architecture |
| 1 | Dense (1000, tanh) |
| 2 | Dense (500, tanh) |
| 3 | Dense (200, tanh) |
| 4 | Dense (100, tanh) |
| 5 | Dense (50, sigmoid) |
| Decoder |  |
| Layer | Network Architecture |
| 1 | Dense (100, tanh) |
| 2 | Dense (200, tanh) |
| 3 | Dense (500, tanh) |
| 4 | Dense (1000, tanh) |
| 5 | Dense (1799, sigmoid) |

Table 6.3: Autoencoder architecture (version 1)

| Generator |  |
| :---: | :---: |
| Layer | Network Architecture |
| 1-3 | Dense (9600), Leaky ReLU (slope $=0.2$ ), Reshape ( $(75,128)$ ) |
| 4-6 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 7-9 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 10-12 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same ), Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 13-15 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 16-17 | Convolutional 1D (1, kernel size $=100$, stride $=1$, padding $=$ same ), Sigmoid |
| Discriminator |  |
| Layer | Network Architecture |
| 1-3 | Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 4-6 | Convolutional 1D ( 128 , kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 7-9 | Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 10-12 | Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 13-14 | Flatten, Dense (1, sigmoid) |

Table 6.4: GAN architecture (version 1)

| Generator |  |
| :---: | :---: |
| Layer | Network Architecture |
| 1-3 | Dense (9600), Leaky ReLU (slope $=0.2$ ), Reshape ( $(75,128)$ ) |
| 4-6 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same ), Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 7-9 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 10-12 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same ), Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 13-15 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 16-17 | Convolutional 1D (1, kernel size $=100$, stride $=1$, padding $=$ same , Sigmoid |
| Discriminator |  |
| Layer | Network Architecture |
| 1-3 | Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 4-6 | Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 7-9 | Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 10-12 | Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 13-14 | Flatten, Dense (1, linear) |

Table 6.5: WGAN architecture (version 1)

| Encoder |  |
| :---: | :---: |
| Layer | Network Architecture |
| 1-3 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=1$ ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 4-6 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=1$ ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 7-9 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=1$ ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 10-12 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=2$ ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 13-15 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=2$ ), Leaky ReLU (slope $=0.2$ ), Dropout $($ probability $=0.2)$ |
| Decoder |  |
| Layer | Network Architecture |
| 1-2 | Transposed Convolutional 1D (256, kernel size $=5$, stride $=2$, padding $=2$ ), Leaky ReLU (slope $=0.2$ ) |
| 3-4 | Transposed Convolutional 1D (256, kernel size $=5$, stride $=2$, padding $=2$ ), Leaky ReLU (slope $=0.2$ ) |
| 5-6 | Transposed Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=1$ ), Leaky ReLU $($ slope $=0.2)$ |
| 7-8 | Transposed Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=1$ ), Leaky ReLU (slope $=0.2$ ) |
| 9-10 | Transposed Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=1$ ), Leaky ReLU (slope $=0.2$ ) |
| 11-12 | Convolutional 1D (1, kernel size $=3$, padding = 1), Sigmoid |

Table 6.6: Autoencoder architecture (version 2)

| Generator |  |
| :---: | :---: |
| Layer | Network Architecture |
| 1-3 | Dense (3200), Leaky ReLU (slope $=0.2$ ), Reshape ( $(25,128)$ ) |
| 4-6 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 7-9 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 10-12 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 13-15 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 16-18 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 19-20 | Convolutional 1D (1, kernel size $=100$, stride $=1$, padding $=$ same ), Sigmoid |
| Discriminator |  |
| Layer | Network Architecture |
| 1-3 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 4-6 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 7-9 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 10-12 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 13-15 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 16-17 | Flatten, Dense (1, sigmoid) |

Table 6.7: GAN architecture (version 2)

| Generator |  |
| :---: | :---: |
| Layer | Network Architecture |
| 1-3 | Dense (3200), Leaky ReLU (slope = 0.2), Reshape ( 25,128 ) |
| 4-6 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 7-9 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 10-12 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=2$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 13-15 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 16-18 | Transposed Convolutional 1D (128, kernel size $=4$, stride $=3$, padding $=$ same $)$, Batch normalization, Leaky ReLU (slope $=0.2$ ) |
| 19-20 | Convolutional 1D (1, kernel size $=100$, stride $=1$, padding $=$ same , Sigmoid |
| Discriminator |  |
| Layer | Network Architecture |
| 1-3 | Convolutional 1D $(256$, kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 4-6 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same ), Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 7-9 | Convolutional 1D ( 256 , kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 10-12 | Convolutional 1D ( 256 , kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 13-15 | Convolutional 1D (256, kernel size $=4$, stride $=2$, padding $=$ same $)$, Leaky ReLU (slope $=0.2$ ), Dropout (probability $=0.2$ ) |
| 16-17 | Flatten, Dense (1, linear) |

Table 6.8: WGAN architecture (version 2)

# Chapter 7: Model Calibration via Distributionally Robust Optimization: On the NASA Langley Uncertainty Quantification Challenge 

We consider the NASA Langley Uncertainty Quantification (UQ) Challenge problem [254] where, given a set of "output" data and under both aleatory and epistemic uncertainties, we aim to infer a region that contains the true values of the associated variables. These steps allow us to investigate the reduction of uncertainty by obtaining further information and estimate the failure probabilities of related systems. To tackle these challenges, we study a methodology based on an integration of robust optimization (RO), more specifically, a recent line of research known as distributionally robust optimization (DRO), and importance sampling in Monte Carlo simulation. We will see that the main computation machinery in this integrated methodology boils down to solving sampled linear programs (LPs). In this chapter, we will explain our methodology, introduce theoretical statistical guarantees via connections to nonparametric hypothesis testing, and present the numerical results on this UQ Challenge.

We briefly introduce the Challenge and notations, where details can be found in [254]. The uncertainty model in the Challenge is given by $\left\langle f_{a}, E\right\rangle$, where $a \sim f_{a}$ is an aleatory variable following a probability density $f_{a}$ and probability distribution function $F_{a}$, and $e \in E$ is an epistemic variable inside the deterministic set $E$. Both the true distribution of $a$ and the true value of $e$ are unknown. Initially, we are given $E_{0} \supset E$ and data $D_{1}=\left\{y^{(i)}(t)\right\}, i=1, \ldots, n_{1}$ in the form of a discrete-time trajectory observed at $n_{t}$ time points from 0 to $T$. We have the computational capability to simulate $y(a, e, t)$ for given values of $a \in A, e \in E_{0}$. The task is to calibrate the distribution of $a$ and value of $e$ with uncertainty quantification, as well as using them to conduct downstream decision and risk evaluation tasks.

### 7.1 Overview of Our Methodology (Problem A)

We first give a high-level overview of our methodology in extracting a region $E$ that contains the true epistemic variables. For convenience, we call this region an "eligibility set" of $e$. For each value of $e$ inside $E$, we also have a set (in the space of probability distributions) that contains "eligible" distributions for the random variable $a$. For the sake of computational tractability (as we will see shortly), the eligibility set of $e$ is represented by a set of sampled points in $E_{0}$ that approximate its shape, whereas the eligibility set of $a$ is represented by probability weights on sampled points on $A$. The eligibility set $E$ and the corresponding eligibility set of distributions for $a$ are obtained by solving an array of LPs that are constructed from these properly sampled points, and then deciding eligibility by checking the LP optimal values against a threshold that resembles the " $p$-value" approach in hypothesis testing. As another key ingredient, this methodology involves a dimension-collapsing transformation $\mathbf{S}$, applied on the raw data, which ultimately allows using the Kolgomorov-Smirnov (KS) statistic to endow rigorous statistical guarantees.

Algorithm 7.1 is a procedural description of our approach to construct the eligibility set $E$, which also gives as a side product an eligibility set of the distributions of $a$ for each $e$, represented by weights in the set (7.11). In the following, we explain the elements and terminologies in this algorithm in detail.

## Algorithm 7.1: Constructing eligibility set $E$.

Input: Data $D_{1}=\left\{\left(y^{(i)}(t)\right)_{t=0, \ldots, T}\right\}_{i=1, \ldots, n_{1}}$. A uniformly sampled set of $e^{(l)}, l=1, \ldots, n_{2}$ over $E_{0}$. A uniformly sampled set of $a^{(j)}, j=1, \ldots, k$ over $A$. A summary function $\mathbf{S}(\cdot): \mathbb{R}^{n_{t}+1} \rightarrow \mathbb{R}^{m}$. A target confidence level $1-\alpha$.
1 Simulate outputs from the baseline distribution: Evaluate $\left(y\left(a^{(j)}, e^{(l)}, t\right)\right)_{t=0, \ldots, T}$ for $j=1, \ldots, k, l=1, \ldots, n_{2}$.
2 Summarize the outputs: Evaluate $\mathbf{s}^{(i)}=\mathbf{S}\left(\left(y^{(i)}(t)\right)_{t=0, \ldots, T}\right)$ for $i=1, \ldots, n_{1}$, and $\left.\mathbf{S}\left(y\left(a^{(j)}, e^{(l)}, t\right)\right)_{t=0, \ldots, T}\right)$ for $j=1, \ldots, k, l=1, \ldots, n_{2}$.
3 Compute the degree of eligibility: For each $l=1, \ldots, n_{2}$, solve optimization problem Eq. (7.10) to obtain $q_{l}^{*}$.
${ }_{4}$ Construct the eligibility set: Output $E=\left\{e^{(l)}: q_{l}^{*} \leq q_{1-\alpha / m}\right\}$. Smooth the set if needed.

### 7.2 A DRO Perspective

Our starting idea is to approximate the set

$$
\begin{equation*}
E=\left\{e \in E_{0}: \text { there exists } P_{e} \text { s.t. } d\left(P_{e}, \hat{P}\right) \leq \eta\right\} \tag{7.1}
\end{equation*}
$$

where $P_{e}$ is the probability distribution of $\{y(a, e, t)\}_{t=0, \ldots, T}$, namely the outputs of the simulation model $\{y(a, e, t)\}_{t=0, \ldots, T}$ at a fixed $e$ but random $a . \hat{P}$ denotes the empirical distribution of $D_{1}$, more concretely the distribution given by

$$
\hat{P}(\cdot)=\frac{1}{n_{1}} \sum_{i=1}^{n_{1}} \delta_{\left(y^{(i)}(t)\right)_{t=0, \ldots, T}}(\cdot),
$$

where $\delta_{\left(y^{(i)}(t)\right)_{t=0, \ldots, T}}(\cdot)$ denotes the Dirac measure at $\left(y^{(i)}(t)\right)_{t=0, \ldots, T} . d(\cdot, \cdot)$ denotes a discrepancy between two probability distributions, and $\eta \in \mathbb{R}_{+}$is a suitable constant. Intuitively, $E$ in Eq. (7.1) is the set of $e$ such that there exists a distribution for the outputs that is close enough to the empirical distribution from the data. If for a given $e$ there does not exist any possible output distribution that is close to $\hat{P}$, then $e$ is likely not the truth. The following gives a theoretical justification for using Eq. (7.1):

Theorem 7.1. Suppose that the true distribution of the output $(y(t))_{t=0, \ldots, T}$, called $P_{\text {true }}$, satisfies $d\left(P_{\text {true }}, \hat{P}\right) \leq \eta$ with confidence level $1-\alpha$, i.e., we have

$$
\begin{equation*}
\mathbb{P}\left(d\left(P_{\text {true }}, \hat{P}\right) \leq \eta\right) \geq 1-\alpha, \tag{7.2}
\end{equation*}
$$

where $\mathbb{P}$ denotes the probability with respect to the data. Then the set $E$ in Eq. (7.1) satisfies $\mathbb{P}\left(e_{\text {true }} \in E\right) \geq 1-\alpha$, where $e_{\text {true }}$ denotes the true value of $e$. Similar deduction holds if $E q$. (7.2) holds asymptotically (as the data size grows), in which case the same asymptotic modification holds for the conclusion.

The proof of Theorem 7.1 comes from a straightforward set inclusion.

Proof. Note that $d\left(P_{\text {true }}, \hat{P}\right) \leq \eta$ implies $e_{\text {true }} \in E$. Thus we have $\mathbb{P}\left(e_{\text {true }} \in E\right) \geq \mathbb{P}\left(d\left(P_{\text {true }}, \hat{P}\right) \leq\right.$ $\eta) \geq 1-\alpha$. Similar derivation holds for the asymptotic version.

In Eq. (7.1), the set of distributions $\left\{P_{e}: d\left(P_{e}, \hat{P}\right) \leq \eta\right\}$ is analogous to the so-called uncertainty set or ambiguity set in the RO literature (e.g., $[228,255]$ ), which is a set postulated to contain the true values of uncertain parameters in a model. RO generally advocates decision-making under uncertainty that hedges against the worst-case scenario, where the worst case is over the uncertainty set (and thus often leads to a minimax optimization problem). DRO, in particular, focuses on problems where the uncertainty is on the probability distribution of an underlying random variable (e.g., [205, 204]). This is the perspective that we are taking here, where $a$ has a distribution that is unknown, in addition to the uncertainty on $e$. Moreover, we also take a generalized view of RO or DRO here as attempting to construct an eligibility set of $e$ instead of finding a robust decision via a minimax optimization.

Theorem 7.1 focuses on the situation where the uncertainty set is constructed and calibrated from data, which is known as data-driven RO or DRO ([256, 257]). If such an uncertainty set has the property of being a confidence region for the uncertain parameters or distributions, then by solving RO or DRO, the confidence guarantee can be translated to the resulting decision, or the eligibility set in our case. Here we have taken a nonparametric and frequentist approach, as opposed to other potential Bayesian methods.

In implementation we choose $\alpha=0.05$, so that the eligibility set $E$ has the interpretation of approximating a $95 \%$ confidence set for $e$. In the above developments, $d\left(P_{e}, \hat{P}\right) \leq \eta$ can in fact be replaced with a more general set $P_{e} \in \mathcal{U}$ where $\mathcal{U}$ is calibrated from the data. Nonetheless, the distance-based set (or "ball") surrounding the empirical distribution is intuitive to understand, and our specific choice of the set below falls into such a representation.

To use Eq. (7.1), there are two immediate questions:

1. What $d(\cdot, \cdot)$ should and can we use, and how do we calibrate $\eta$ ?
2. How do we determine whether there exists $P_{e}$ that satisfies $d\left(P_{e}, \hat{P}\right) \leq \eta$ for a given $e$ ?

In the following two sections, we address the above two questions respectively which would then lead us to Algorithm 7.1.

### 7.3 Constructing Discrepancy Measures

For the first question, we first point out that in theory many choices of $d$ could be used (basically, any $d$ that satisfies the confidence property in Theorem 7.1). But, a poor choice of $d$ would lead to a more conservative result, i.e., larger $E$, than others. A natural choice of $d$ should capture the discrepancy of the distributions efficiently. Moreover, the choice of $d$ should also account for the difficulty in calibrating $\eta$ such that the assumption in Theorem 7.1 can be satisfied, as well as the computational tractability in solving the eligibility determination problem in Eq. (7.1).

Based on the above considerations, we construct $d$ and calibrate $\eta$ as follows. First, we "summarize" the data $D_{1}$ into a lower-dimensional representation, say $\left\{s_{1}^{(i)}, \ldots, s_{m}^{(i)}\right\}, i=1, \ldots, n_{1}$, where $s_{v}^{(i)}=S_{v}\left(y^{(i)}(t)_{t=0, \ldots, T}\right)$ for some function $S_{v}(\cdot)$. For convenience, we denote $\mathbf{S}(\cdot)=$ $\left(S_{1}(\cdot), \ldots, S_{m}(\cdot)\right): \mathbb{R}^{n_{t}+1} \rightarrow \mathbb{R}^{m}$, and $\mathbf{s}^{(i)}=\left(s_{1}^{(i)}, \ldots, s_{m}^{(i)}\right)$. We call $\mathbf{S}(\cdot)$ the "summary function" and $\mathbf{s}^{(i)}$ the "summaries" of the $i$-th output. $\mathbf{S}(\cdot)$ attempts to capture important characteristics of the raw data (we will see later that we use the positions and values of the peaks extracted from Fourier analysis). Also, the low dimensionality of $\mathbf{s}^{(i)}$ is important to calibrate $\eta$ well.

Next, we define

$$
\begin{equation*}
d\left(P_{e}, \hat{P}\right)=\max _{v=1, \ldots, m} \sup _{x \in \mathbb{R}}\left|F_{e, v}(x)-\hat{F}_{v}(x)\right| \tag{7.3}
\end{equation*}
$$

where $\hat{F}_{v}(x)=\frac{1}{n_{1}} \sum_{i=1}^{n_{1}} I\left(s_{v}^{(i)} \leq x\right)$, with $I(\cdot)$ denoting the indicator function, is the empirical distribution function of $s_{v}^{(i)}$ (i.e., the distribution function of $\hat{P}$ projected onto the $v$-th summary). $F_{e, v}(x)$ is the probability distribution function of the $v$-th summary of the simulation model output $S_{v}(y(a, e, t))_{t=0, \ldots, T}$ (i.e., the distribution function of the projection of $P_{e}$ onto the $v$-th summary). We then choose $\eta=q_{1-\alpha / m} / \sqrt{n_{1}}$ as the $(1-\alpha / m)$-quantile of the Kolmogorov-Smirnov (KS) statistic, namely that $q_{1-\alpha / m}$ is the $(1-\alpha / m)$-quantile of $\sup _{x \in[0,1]}|B B(x)|$ where $B B(\cdot)$ denotes a standard Brownian bridge [258, 259].

To understand Eq. (7.3), note that the set of $P_{e}$ that satisfies $d\left(P_{e}, \hat{P}\right) \leq \eta$ is equivalent to $P_{e}$ that satisfies

$$
\begin{equation*}
\sup _{x \in \mathbb{R}}\left|F_{e, v}(x)-\hat{F}_{v}(x)\right| \leq \frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}, \quad v=1, \ldots, m, \tag{7.4}
\end{equation*}
$$

Here, $\sup _{x \in \mathbb{R}}\left|F_{e, v}(x)-\hat{F}_{v}(x)\right|$ is the KS-statistic for a goodness-of-fit test against the distribution $F_{e, v}(x)$, using the data on the $v$-th summary. Since we have $m$ summaries and hence $m$ tests, we use a Bonferroni correction and deduce that

$$
\liminf _{n_{1} \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|F_{\text {true }, v}(x)-\hat{F}_{v}(x)\right| \leq \frac{q_{1-\alpha / m}}{\sqrt{n_{1}}} \text { for } v=1, \ldots, m\right) \geq 1-\alpha
$$

where $F_{\text {true, }}$ denotes the true distribution function of the $v$-th summary. Thus, the (asymptotic version of the) assumption in Theorem 7.1 holds.

Note that here the quality of the summaries does not affect the statistical correctness of our method (in terms of overfitting), but it does affect crucially the resulting conservativeness (in the sense of getting a larger $E$ ). Moreover, in choosing the number of summaries $m$, there is a tradeoff between the conservativeness coming from representativeness and simultaneous estimation. On one end, using more summaries mean more knowledge we impose on $P_{e}$, which translates into a smaller feasible set for $P_{e}$ and ultimately a smaller eligibility set $E$. This relation, however, is true only if there is no statistical noise coming from the data. In the case of finite data size $n_{1}$, then more summaries also means that constructing the feasible set for $P_{e}$ requires more simultaneous estimations in calibrating its size, which is manifested in the Bonferroni correction whose degree increments with each additional summary. In our implementation (see Section 7.7), we find that using 12 summaries seems to balance well this representativeness versus simultaneous estimation error tradeoff.

We also note that there could be other approaches than Bonferroni to aggregate the statistics of individual summaries to obtain an overall familywise guarantee, which could potentially reduce its conservativeness especially when the summaries are strongly correlated. These alternatives could be done by estimating the dependence among the summaries, or using $p$-value aggregation
methods that do not require dependence information [260, 261, 262]. We choose Bonferroni here since it is easy to use and perform reasonably in our implementation.

### 7.4 Determining Existence of an Aleatory Distribution

Now we address the second question on how we can decide, for a given $e$, whether a $P_{e}$ exists such that $d\left(P_{e}, \hat{P}\right) \leq \eta$. We first rephrase the representation with a change of measure. Consider a "baseline" probability distribution, say $P_{0}$, that is chosen by us in advance. A reasonable choice, for instance, is the uniform distribution over $A$, the support of $a$. Then we can write $d\left(P_{e}, \hat{P}\right) \leq \eta$ as

$$
\begin{equation*}
\sup _{x \in \mathbb{R}}\left|\int_{S_{v}(u) \leq x} W_{e}(u) d P_{0}(u)-\hat{F}_{v}(x)\right| \leq \frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}, v=1, \ldots, m, \tag{7.5}
\end{equation*}
$$

where $W_{e}(\cdot)=d P_{e} / d P_{0}$ is the Radon-Nikodym derivative of $P_{e}$ with respect to $P_{0}$, and we have used the change-of-measure representation $F_{e, v}(x)=\int_{S_{v}(u) \leq x} W_{e}(u) d P_{0}(u)$. Here we have assumed that $P_{0}$ is suitably chosen such that absolute continuity of $P_{e}$ with respect to $P_{0}$ holds. Eq. (7.5) turns the determination of the existence of eligible $P_{e}$ into the existence of an eligible Radon-Nikodym derivative $W_{e}(\cdot)$.

The next step is to utilize Monte Carlo simulation to approximate $P_{0}$. More specifically, given $e$, we run $k$ simulation runs under $P_{0}$ to generate $\left(y\left(a^{(j)}, e, t\right)\right)_{t=0, \ldots, T}$ for $j=1, \ldots, k$. Then Eq. (7.5) can be approximated by

$$
\begin{equation*}
\sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\left(y\left(a^{(j)}, e, t\right)\right)_{t=0, \ldots, T}\right) \leq x\right)-\hat{F}_{v}(x)\right| \leq \frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}, v=1, \ldots, m \tag{7.6}
\end{equation*}
$$

where $W_{j}=(1 / k)\left(d P_{e} / d P_{0}\left(\left(y\left(a^{(j)}, e, t\right)\right)\right)\right)$ represents the (unknown) sampled likelihood ratio from the view of importance sampling [23, 64]. Our task is to find a set of weights, $W_{j}, j=$ $1, \ldots, k$, such that Eq. (7.6) holds. These weights should approximately satisfy the properties of the Radon-Nikodym derivative, namely positivity and integrating to one. Thus, we seek for
$W_{j}, j=1, \ldots, k$ such that

$$
\begin{gather*}
\sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\left(y\left(a^{(j)}, e, t\right)\right)_{t=0, \ldots, T}\right) \leq x\right)-\hat{F}_{v}(x)\right| \leq \frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}, v=1, \ldots, m  \tag{7.7}\\
\sum_{j=1}^{k} W_{j}=1, W_{j} \geq 0 \text { for } j=1, \ldots, k, \tag{7.8}
\end{gather*}
$$

where Eq. (7.8) enforces the weights to lie in a probability simplex. If $k$ is much larger than $n_{1}$, then the existence of $W_{j}, j=1, \ldots, k$ satisfying Eq. (7.7) and Eq. (7.8) would determine that the considered $e$ is in $E$. To summarize, we have:

Theorem 7.2. Suppose $k=\omega\left(n_{1}\right)$, and $P_{\text {true }}$ is absolutely continuous with respect to $P_{0}$ and that $\left\|d P_{\text {true }} / d P_{0}\right\|_{\infty} \leq C$ for some constant $C>0$ and $\|\cdot\|_{\infty}$ denotes the essential supremum. Suppose, for each $e$, we generate $k$ simulation replications to get $\left.\left(y\left(a^{(j)}, e, t\right)\right)_{t=0, \ldots, T}\right), j=1, \ldots, k$, where $a^{(j)}$ are drawn from $P_{0}$ in an i.i.d. fashion. Then the set

$$
E=\left\{e: \text { there exists } W_{j}, j=1, \ldots, k \text { such that Eq. (7.7) and Eq. (7.8) hold }\right\}
$$

will satisfy

$$
\liminf _{n_{1} \rightarrow \infty, k / n_{1} \rightarrow \infty} \mathbb{P}\left(e_{\text {true }} \in E\right) \geq 1-\alpha
$$

The proof of Theorem 7.2 is in Section 7.14. Note that this theorem is an asymptotic statement when $k$ is chosen much larger than $n_{1}$. In our subsequent implementation, we configure our simulation size so that the latter condition holds.

Note that in Theorem 7.2, $W_{j}$ 's represent the unknown sampled likelihood ratios such that, together with the $a^{(j)}$,s generated from $P_{0}$, the function

$$
\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\left(y\left(a^{(j)}, e, t\right)\right)_{t=0, \ldots, T}\right) \leq \cdot\right)
$$

approximates the unknown true $v$-th summary distribution function $F_{\text {true }, v}$.
To use the above $E$ and elicit the guarantee in Theorem 7.2, we still need some steps in order to conduct feasible numerical implementation. First, we need to discretize or sufficiently sample $e$ 's over $E_{0}$, since checking the existence of eligible $W_{j}$ 's for all $e$ is computationally infeasible. In our implementation we draw $n_{2}=1000 e$ 's uniformly over $E_{0}$, call them $e^{(1)}, \ldots, e^{\left(n_{2}\right)}$, and then put together the geometry of $E$ from the eligible $e^{(l)}$ 's. Second, the current representation of the KS constraint Eq. (7.7) involves entire distribution functions. We can write Eq. (7.7) as a finite number of linear constraints, given by

$$
\begin{equation*}
\hat{F}_{v}\left(s_{v}^{(i)}+\right)-\frac{q_{1-\alpha / m}}{\sqrt{n_{1}}} \leq \sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\left(y\left(a^{(j)}, e, t\right)\right)_{t=0, \ldots, T}\right) \leq s_{v}^{(i)}\right) \leq \hat{F}_{v}\left(s_{v}^{(i)}-\right)+\frac{q_{1-\alpha / m}}{\sqrt{n_{1}}} \tag{7.9}
\end{equation*}
$$

for $i=1, \ldots, n_{1}, v=1, \ldots, m$, where $s_{v}^{(i)}, i=1, \ldots, n_{1}$ are the $v$-th summary of the $i$-th data point, and $s_{v}^{(i)}+$ and $s_{v}^{(i)}$ - denote the right and left limits of the empirical distribution at $s_{v}^{(i)}$.

Thus, putting everything together, we solve, for each $e^{(l)}, l=1, \ldots, n_{2}$, the feasibility problem:

Find $W_{j}, j=1, \ldots, k$ such that Eq. (7.9) and Eq. (7.8) hold.

If there exists feasible $W_{j}, j=1, \ldots, k$, then $e^{(l)}$ is eligible. The set $\left\{e^{(l)}: e^{(l)}\right.$ is eligible $\}$ is an approximation of $E$. Note that this is a "sampled" subset of $E$. In general, without running the simulation at the other points of $E$, there is no guarantee whether these other points are eligible or not. However, if the distribution of $\{y(a, e, t)\}_{t=0, \ldots, T}$ is continuous in $e$ in some suitable sense, then it is reasonable to believe that the neighborhood of an eligible point $e^{(l)}$ is also eligible (and vice versa). In this case, we can "smooth" the discrete set of $\left\{e^{(l)}: e^{(l)}\right.$ is eligible $\}$ if needed (e.g., by doing some clustering and taking the convex hull of each cluster). Finally, note that the feasibility problem above is a linear problem in the decision variables $W_{j}$ 's.

### 7.5 Towards the Main Procedure

To link to our main Algorithm 7.1, we offer an equivalent approach to the above feasibility-problem-based procedure that allows further flexibility in choosing the threshold $q_{1-\alpha / m}$, which currently is set as the Bonferroni-adjusted KS critical value. This equivalent approach leaves this choice of threshold open and can determine the set of eligible $e^{(l)}$ as a function of the threshold, thus giving some room to improve conservativeness should the formed approximate $E$ turns out to be too loose according to other expert opinion. Here, we solve, for each $e^{(l)}, l=1, \ldots, n_{2}$, the optimization problem

$$
\begin{align*}
q_{l}^{*}=\min & q \\
\text { s.t. } & \hat{F}_{v}\left(s_{v}^{(i)}+\right)-\frac{q}{\sqrt{n_{1}}} \\
& \leq \sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\left(y\left(a^{(j)}, e^{(l)}, t\right)\right)_{t=0, \ldots, T}\right) \leq s_{v}^{(i)}\right)  \tag{7.10}\\
& \leq \hat{F}_{v}\left(s_{v}^{(i)}-\right)+\frac{q}{\sqrt{n_{1}}} \text { for } i=1, \ldots, n_{1}, v=1, \ldots, m ; \\
& \sum_{j=1}^{k} W_{j}=1, W_{j} \geq 0 \text { for } j=1, \ldots, k,
\end{align*}
$$

where the decision variables are $W_{j}, j=1, \ldots, k$ and $q$. If the optimal value $q_{l}^{*}$ satisfies $q_{l}^{*} \leq$ $q_{1-\alpha / m}$, then $e^{(l)}$ is eligible (This can be seen by checking its equivalence to the feasibility problem via the monotonicity of the feasible region for $W_{j}$ 's in Eq. (7.10) as $q$ increases). The rest then follows as above that $\left\{e^{(l)}: e^{(l)}\right.$ is eligible $\}$ is an approximation of $E$. Like before, Eq. (7.10) is an LP. Moreover, here $q_{l}^{*}$ captures in a sense the "degree of eligibility" of $e^{(l)}$, and allows convenient visualization by plotting $q_{l}^{*}$ against $e^{(l)}$ to assess the geometry of $E$. For these reasons we prefer to use Eq. (7.10) over the feasibility problem before. These give the full procedure in Algorithm 7.1. Note that Algorithm 7.1 has a variant where we re-generate a sample of $a^{(j)}$ 's for each different $e^{(l)}$. It is clear that the correctness guarantee (Theorem 7.2) still holds in this case.

Now we analyze the complexity of our calibration procedure. To implement Algorithm 7.1, we evaluate the output $y$ for $k \times n_{2}$ times and use FFT to summarize the outputs $k \times n_{2}$ times. Then we solve the linear optimization for $n_{2}$ times, where the linear optimization problem includes $k+1$
decision variables and $2 m \times n_{1}+k+1$ linear constraints.
Moreover, we also present how to find eligible distributions of $a$ for an eligible $e^{(l)}$. The set of eligible distributions of $a$ is approximated by the weights $W_{j}$ 's that satisfy Eq. (7.9) and Eq. (7.8), namely

$$
\begin{align*}
\left\{W_{j}, j=1, \ldots, k:\right. & \hat{F}_{v}\left(s_{v}^{(i)}+\right)-\frac{q_{1-\alpha / m}}{\sqrt{n_{1}}} \\
& \leq \sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\left(y\left(a^{(j)}, e^{(l)}, t\right)\right)_{t=0, \ldots, T}\right) \leq s_{v}^{(i)}\right) \\
& \leq \hat{F}_{v}\left(s_{v}^{(i)}-\right)+\frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}, \text { for } i=1, \ldots, n_{1}, v=1, \ldots, m ; \\
& \left.\sum_{j=1}^{k} W_{j}=1, W_{j} \geq 0 \text { for } j=1, \ldots, k\right\}, \tag{7.11}
\end{align*}
$$

where $W_{j}$ is the probability weight on $a^{(j)}$. From this, one could also obtain approximate bounds for quantities related to the distribution of $a$. For instance, to get approximate bounds for the mean of $a$, we can maximize and minimize $\sum_{j} W_{j} a^{(j)}$ subject to constraint (7.11).

### 7.6 Related Literature

Before we discuss our numerical findings, we discuss some related literature on the problem setting and our proposed methodology.

The model calibration problem that infers input from output data has been studied across different disciplines. In scientific areas it is viewed as an inverse problem [186], in which Bayesian methodologies are predominantly used (e.g., [212, 213, 187, 214, 215]). Our presented approach is an alternative to Bayesian methods that aim to provide frequentist guarantees in the form of confidence regions. In addition to Bayesian approaches, other alternative methods include entropy maximization [263] that use the entropy as a criterion to select the "best" distribution, but it does not have the frequentist guarantee in recovering the true distribution that we provide in this UQ Challenge.

We point out that model calibration has also been investigated in the stochastic simulation community [264, 183]. In this setting, model calibration is often viewed together with model validation. To validate a model, the conventional approach is to use statistical tests such as the two-sample mean-difference tests [184] or others like the Schruben-Turing test [185] that decides whether the simulated output data and historical real output data are close enough. If not, then the simulation model is re-calibrated, and this process is repeated until the gap between simulation and real data is sufficiently close. Though having a long history, the development of rigorous frameworks to conduct model calibration and validation has been quite open with relatively few elaborate discussions in the literature [181].

In terms of methodology, our approach is closely related to RO, which is an established method for optimization under uncertainty that advocates the representation of unknown or uncertain parameters in the model as a (deterministic) set (e.g., [228, 255]). This set is often called an uncertainty set or an ambiguity set. In the face of decision-making, RO optimizes the decision over the worst-case scenario within the uncertainty set, which usually comes in the form of a minimax problem with the outer optimization on the decision while the inner optimization on the worst case scenario. DRO, a recently active branch of RO, considers stochastic optimization where the underlying probability distribution is uncertain (e.g., [203, 205, 204]). In this case, the uncertainty set lies in the space of probability distributions and one attempts to make decisions under the worstcase distribution. In this chapter we take a generalized view of RO or DRO as attempting to find a set of eligible "decisions", namely the $e$, so it does not necessarily involve a minimax problem but instead a set construction.

In data-driven RO or DRO, the uncertainty set is constructed or calibrated from data. If such a set has the property of being a confidence region for the uncertain parameters or distributions, then by solving the RO or DRO, the confidence guarantee can be translated to bounds on the resulting decision, and in our case the eligibility set. This approach of constructing uncertainty sets, by viewing them as confidence regions or via hypothesis testing, has been the main approach in data-driven RO or DRO [229]. Recently, alternate approaches have been studied to reduce the
conservativeness in set calibration, by utilizing techniques from empirical likelihood [265, 266, 267], Wasserstein profile function [268], Bayesian perspectives [269] and data splitting [257, 270].

In our development, we have constructed an uncertainty set for the unknown distribution $P_{e}$ via a confidence region associated with the KS goodness-of-fit test. This uncertainty set has been proposed in [229]. Other distance-based uncertainty sets, including $\phi$-divergence [230, 206, 170, 173, 174, 231] and Wasserstein distance [232, 233, 234], have also been used, as well as sets based on moment [204, 235, 171] or distributional shape information [236, 237, 238]. We use a simultaneous group of KS statistics with Bonferroni correction, motivated by the tractability in the resulting integration with the importance weighting. The closest work to our framework is the stochastic simulation inverse calibration problem studied in [202], but they consider singledimensional output and parameter to calibrate the input distributions, in contrast to our "summary" approach via Fourier analysis and the multi-dimensional settings we face.

Another important ingredient in our approach is importance sampling. This is often used as a variance reduction tool (e.g., [51, 52]; [54] Chapter 5; [34] Chapter 4) and is shown to be particularly effective in rare-event simulation (e.g., [22, 63, 23, 64]). It operates by sampling a random variable from a different distribution from the true underlying distribution, and applies a so-called likelihood ratio to de-bias the resulting estimate. Other than variance reduction, importance sampling is also used in risk quantification in operations research and mathematical finance that uses a robust optimization perspective (e.g., [170, 172, 173]), which is more closely related to our use in this chapter. Additionally, it is used in Bayesian computation [271], and more recently in machine learning contexts such as covariate shift estimation [272, 273] and off-policy evaluation in reinforcement learning [274, 275].

In the remainder of this chapter, we illustrate the use of our methodology and report our numerical results on the UQ Challenge.

### 7.7 Summarizing Discrete-Time Histories using Fourier Transform

By observing the plot of the outputs $y^{(i)}, i=1, \ldots, n_{1}$ (see Figure 7.1), we judge that these time series are highly seasonal. Naturally, we choose to use discrete Fourier transform to summarize $(y(t))_{t=0, \ldots, T}$, and we may write $y(t)$ in the form $y(t)=\sum_{k=-(N-1)}^{N-1} C_{k} e^{i k \omega_{0} t}$ where $\omega_{0}=\frac{2 \pi n_{t}}{N T}$. In order to apply Fast Fourier Transform (FFT) more easily, we pick $N=2^{\left\lceil\log _{2}\left(n_{t}+1\right)\right\rceil}$ and pad each $y^{(i)}$ with trailing zeros to length $N$.


Figure 7.1: The plot of $y^{(i)}, i=1, \ldots, n_{1}$

First we apply FFT to $y^{(i)}, i=1, \ldots, n_{1}$. For each $y^{(i)}$, we compute the $C_{k}$ 's. Figure 7.2 shows the real part and the imaginary part of $C_{k}$ 's against the corresponding frequencies. Note that we only show $C_{k}$ 's with positive $k$ since $C_{-k}$ is the complex conjugate of $C_{k}$.

For the real part, we see that there is a large positive peak, a large negative peak, a small positive peak and a small negative peak. After testing, we confirm that for any $i$, the large peaks lie in the first 14 terms (from 0 Hz to 1.59 Hz ), while the small peaks lie between the 15 th term and the 50th term (from 1.71 Hz to 5.98 Hz ). For the imaginary part, we see that there is a large negative peak and a small positive peak. The large peak is also located in the first 14 terms and the small peak between the 15 th term and the 50th one.

Therefore, we choose to use the following method to summarize $y$ (i.e., construct the function $\mathbf{S}(\cdot)$ ): first, we apply the Fourier transform to compute $C_{k}$ 's and the corresponding frequencies; second, we compute the real part and the imaginary part of $C_{k}$ 's; third, for the real part, we find the maximum value and the minimum value over $[0 \mathrm{~Hz}, 1.59 \mathrm{~Hz}]$ and $[1.71 \mathrm{~Hz}, 5.98 \mathrm{~Hz}]$, as well


Figure 7.2: The real part and the imaginary part of $C_{k}$ 's against the corresponding frequencies
as their corresponding frequencies; fourth, for the imaginary part, we find the minimum value over $[0 \mathrm{~Hz}, 1.59 \mathrm{~Hz}]$ and the maximum value over $[1.71 \mathrm{~Hz}, 5.98 \mathrm{~Hz}]$ as well as their corresponding frequencies. Then we use these 12 parameters as the summaries of $y$.

To illustrate how well these summaries fit $y$, Figure 7.3a shows the comparison for $y^{(1)}$. The fit qualities of other time series are similar to this example. Though they are not extremely close to each other, the fitted curves do resemble the original curves. Note that it is entirely possible to improve the fitting if we keep more frequencies even if they are not as significant as the main peaks. For instance, Figure 7.3 b shows the improved fitting curve if for both real part and imaginary part, we respectively keep the 20 frequencies with the largest values. It can be seen that now the fit quality is quite good. On the other hand, as discussed in Section 7.2, using a larger number of summaries both represents more knowledge of $P_{e}$ (better fitting) but also leads to more simultaneous estimation error when using the Bonferroni correction needed in calibrating the set for $P_{e}$. To balance the conservativeness of our approach coming from representativeness versus simultaneous estimation, we choose to use the 12-parameter summaries depicted before.


Figure 7.3: Fitting $y^{(1)}$ with different number of parameters

### 7.8 Uncertainty Reduction (Problem B)

### 7.8.1 Ranking Epistemic Parameters (B.1 and B.2)

Now we implement Algorithm 7.1 with $n_{2}=k=1000$ and the summary function $\mathbf{S}(\cdot)$ defined in the previous section. The dimension of the summary function is $m=12$. We choose $\alpha$ to be 0.05. Thus, following the algorithm, for each $l=1, \ldots, n_{2}$, we compute $q_{l}^{*}$ and then compare it with $q_{1-\alpha / m}=q_{1-0.05 / 12}=1.76$.

In Figure 7.4, we plot the $q_{l}^{*}$ 's against each dimension of $e$. The red horizontal lines in the graphs correspond to $q_{1-\alpha / m}=1.76$. Thus the dots below the red lines constitute the eligible $e$ 's. We rank the epistemic parameters according to these graphs, namely we rank higher the parameter whose range can potentially be reduced the most if we remove the region with no eligible $e$. Note that this ranking scheme can be summarized using more rigorous metrics related to the expected amount of eligible $e$ 's after range shrinkage, but since there are only four dimensions, using the graphs directly seem sufficient for our purpose here.

We find that the values of $e_{2}$ and $e_{4}$ of the eligible $e$ 's broadly range from 0 to 2 , which implies that reducing the ranges of these two dimensions could hardly reduce our uncertainty. By contrast, the values of $e_{1}$ and $e_{3}$ of the eligible $e$ 's are both concentrated in the lower part of [0,2]. Thus, our ranking of the epistemic parameters according to their ability to improve the predictive ability
is $e_{3}>e_{1}>e_{2}>e_{4}$.
Chances are that the true values of $e_{1}$ and $e_{3}$ are relatively small. In order to further pinpoint the true values of $e_{1}$ and $e_{3}$, we choose to make two uncertainty reductions: increase the lower limits of the bounding interval of $e_{1}$ and $e_{3}$.


Figure 7.4: $q_{l}^{*}$ against each epistemic variable

### 7.8.2 Impact of the value of $n_{1}$ (A.2)

To investigate the impact of the value of $n_{1}$, for different values of $n_{1}$ we randomly sample $n_{1}$ outputs without replacement. Then we take these outputs as the new data set. By repeatedly implementing Algorithm 7.1, we find that the larger is $n_{1}$, the smaller is the proportion of eligible $e$ 's. It is intuitive that as the data size grows, $e$ can be better pinpointed. Moreover, except for $e_{4}$,
the range of each epistemic variable of eligible $e$ 's obviously shrinks as $n_{1}$ increases, which further confirms that $e_{4}$ is the least important epistemic variable.

### 7.8.3 Updated Parameter Ranking (B.3)

After the epistemic space is reduced, we repeat the process in Section 7.8.1 but now $e$ 's are generated uniformly from $E_{1}$. From the associated scatter plots (Figure 7.5), the updated ranking of the epistemic parameters is $e_{2}>e_{3}>e_{1}>e_{4}$.


Figure 7.5: $q_{l}^{*}$ against each epistemic variable (refined)

### 7.9 Reliability of Baseline Design (Problem C)

### 7.9.1 Failure Probabilities and Severity (C.1, C. 2 and C.5)

Combining the refined range of $e$ provided by the host with our Algorithm 7.1, we construct $E \subset E_{1}$. To estimate $\min _{e \in E} / \max _{e \in E} \mathbb{P}\left(g_{i}(a, e, \theta) \geq 0\right)$, we run simulations to respectively solve

$$
\begin{gather*}
\min / \max \sum_{j=1}^{k} W_{j} I\left(g_{i}\left(a^{(j)}, e, \theta\right) \geq 0\right)  \tag{7.12}\\
\text { s.t. } e \in E, W \in U_{e}
\end{gather*}
$$

where $U_{e}$ is the set of $\left(W_{1}, \cdots, W_{k}\right)$ in Eq. (7.11) that is dependent on the value of $e$. These give the range of $R_{i}(\theta)$. We use the same method to approximate $R(\theta)$, the failure probability for any requirement. Note that in our implementation the $E$ in the formulations above is represented by discrete points $e^{(l)}$,s. As discussed previously, under additional smoothness assumptions, we could "smooth" these points to obtain a continuum. Nonetheless, under sufficient sampling of $e^{(l)}$, the discretized set should be a good enough approximation in the sense that the optimal values from the "discretized" problems are close to those using the continuum.

Using the above method, we get that the ranges of $R_{1}(\theta), R_{2}(\theta), R_{3}(\theta)$ and $R(\theta)$ are approximately $[0,0.6235],[0,0.7320],[0,0.5270]$ and $[0,0.8217]$. Though the ranges seem to be quite wide, they can provide us useful information to be utilized next.

To evaluate $s_{i}(\theta)$, the severity of each individual requirement violation, similarly we simulate

$$
\max _{e \in E} \max _{W \in U} \sum_{j=1}^{k} W_{j} g_{i}\left(a^{(j)}, e, \theta\right) I\left(g_{i}\left(a^{(j)}, e, \theta\right) \geq 0\right)
$$

The results for $s_{1}(\theta), s_{2}(\theta)$ and $s_{3}(\theta)$ are respectively $0.1464,0.0493$ and 3.5989. Clearly the violation of $g_{3}$ is the most severe one while the violation of $g_{2}$ is the least.

### 7.9.2 Rank for Uncertainties (C.3)

Our analysis on the rank for epistemic uncertainties is based on the range of $R(\theta)$ obtained above. In our computation, we obtain

$$
\min _{W \in U} / \max _{W \in U} \sum_{j=1}^{k} W_{j} I\left(g_{i}\left(a^{(j)}, e, \theta\right) \geq 0 \text { for some } i=1,2,3\right)
$$

for each eligible $e \in E$. For simplicity, we use $R_{\min }$ and $R_{\max }$ to denote these two values for each eligible $e \in E$ respectively.

Our approach is to scrutinize the plots of $R_{\min }$ and $R_{\max }$ against each epistemic variable (Figure 7.6 and 7.7). For $R_{\text {min }}$, large values are notable, since it means that any distribution that provides similarity to the original data is going to fail with a large probability. Therefore the most ideal reduction is to avoid the region of $e$ such that all $R_{\min }$ 's are large. For $R_{\max }$, the largest $R_{\max }$ for the region denotes the maximum failure probability that one can have. So we pay attention to the epistemic variables that could potentially reduce the "worst-case" failure probability. Note that it is possible to convert this approach to a quantitative metric by setting a threshold to define large $R_{\text {min }}$ and $R_{\text {max }}$, and checking the reduced number of sampled $e$ 's having a large $R_{\min }$ and $R_{\text {max }}$ when the range of an epistemic variable shrinks. But since there are only four dimensions, visualizing from the graphs directly seems sufficient for our purpose. Based on these considerations, we conclude that the rank for epistemic uncertainties is $e_{3}>e_{1}>e_{2}>e_{4}$.

### 7.9.3 Representative Realizations (C.4)

Since the distribution of $a$ in our approach is defined as an ambiguity set that depends on $e$, the failure domain would also be based on each eligible $e$. We classify an eligible $e$ to be notable if its corresponding $R_{\min }$ is relatively large. We set this threshold to define a large $R_{\text {min }}$ to 0.1 , but noting that the choice here is context-dependent in general. For convenience, we denote the


Figure 7.6: $R_{\text {min }}$ against each epistemic variable.


Figure 7.7: $R_{\max }$ against each epistemic variable.
"best-case"distribution corresponding to $R_{\min }$ as $w_{\min }$, where

$$
w_{\min }=\arg \min _{W \in U} \sum_{j=1}^{k} W_{j} I\left(g_{i}\left(a^{(j)}, e, \theta\right) \geq 0\right) .
$$

We consider the representative realizations of uncertainties as those $a$ 's with large value of $w_{\text {min }}$ (in our case we consider $>0.05$ ).

From our observation, we find that these representative realizations have a clear pattern on the scatter plot with $a_{1}$ and $a_{3}$ as the coordinates (as in Figure 7.8). We also provide some example responses of cases in each group. We observe that there is a clear similarity in the responses within each group, which can be interpreted as different failure patterns.


Figure 7.8: The four groups of representative realizations on the scatter plot with $a_{1}$ and $a_{3}$ as the coordinates. The four groups are failure cases caused by $g_{1}, g_{2}$ and $g_{3}$ (blue), $g_{1}$ (red), $g_{2}$ (yellow) and $g_{3}$ (green).

### 7.10 Reliability-Based Design (Problem D)

To find a reliability-optimal design point $\theta_{\text {new }}$, we minimize

$$
\begin{equation*}
\max _{e \in E} \min _{W \in U} \sum_{j=1}^{k} W_{j} I\left(g\left(a^{(j)}, e, \theta\right) \geq 0\right) . \tag{7.13}
\end{equation*}
$$

Here is the reason why we choose this function as the objective. For an eligible $e \in E$, if $\min _{W \in U} \sum_{j=1}^{k} W_{j} I\left(g\left(a^{(j)}, e, \theta\right) \geq 0\right)$ is large, then the true probability in which the system fails must be even larger than this "best-case" estimate, which implies that this point $e$ has a considerable failure likelihood. The objective above thus aims to find a design point to minimize this bestcase estimate, but taking the worst-case among all the eligible $e$ 's. Arguably, one can use other criteria such as minimizing $\max _{e \in E} \max _{W \in U} \sum_{j=1}^{k} W_{j} I\left(g\left(a^{(j)}, e, \theta\right) \geq 0\right)$, but this could make our procedure more conservative.

The optimization problem (7.13) is of a "black-box" nature since the function $g$ is only observed through simulation, and the problem is easily non-convex. Instead of insisting on full optimality which could be difficult to achieve in this problem, we use gradient descent to guide us toward local improvement over the baseline design. Note that we need to sample $a^{(j)}$ when we land at a new $\theta$ during our iterations, and hence our approach takes the form of a stochastic gradient descent or stochastic approximation [276, 277]. Moreover, the gradient cannot be estimated in an unbiased fashion as we only have black-box function evaluation, and thus we need to resort to the use of finite-difference. This results in a zeroth-order or the so-called Kiefer-Wolfowitz (KW) algorithm [278, 279]. As we have a nine-dimensional design variable, we choose to update $\theta$ via a coordinate descent, namely at each iteration we choose one of the dimensions and run a central finite-difference along that dimension, followed by a movement of $\theta$ guided by this gradient estimate with a suitable step size. The updates are done in a round-about fashion over the dimensions. The perturbation size in the finite-difference is chosen of order $1 / n^{1 / 4}$ here as it appears to perform well empirically (though theoretically other scaling could be better).

Algorithm 7.2 shows the details of our optimization procedure. Considering that the compo-

```
Algorithm 7.2: KW algorithm to find \(\theta_{\text {new }}\).
    Set \(x_{\text {now }}=1_{9}\) and \(n=1\).
    While \(n \leq N_{\max }\) do
        Set \(c_{n}=c_{0} / n^{1 / 4}\) and \(a_{n}=a_{0} / n\).
        For \(i\) from 1 to 9 do
            \(u=f\left(\theta_{\text {baseline }} \circ\left(x_{\text {now }}+c_{n} e_{i}\right)\right)\).
            \(l=f\left(\theta_{\text {baseline }} \circ\left(x_{\text {now }}-c_{n} e_{i}\right)\right)\).
            \(g=(u-l) /\left(2 c_{n}\right)\).
            \(x_{\text {now }}=x_{\text {now }}-a_{n} g e_{i}\).
        End
        \(n=n+1\).
    End
    Output \(\theta_{\text {baseline }} \circ x_{\text {now }}\).
    (o denotes the Hadamard product).
```

    Input: The baseline design point \(\theta_{\text {baseline }}\). The initial step size \(c_{0}\). The initial perturbation
                    size \(a_{0}\). The max iteration \(N_{\max }\). The objective function \(f(\theta)\).
    nents of $\theta_{\text {baseline }}$ are of very different magnitudes, we first perform a normalization to ease this difference. The quantity $x_{\text {now }}$ encodes the position of the normalized $\theta_{\text {now }}$, and $1_{9}$ denotes a ninedimensional vector of 1 's that is set as the initial normalized design point. We set $c_{0}=a_{0}=0.1$ and $N_{\max }=8$.

After running the algorithm, we arrive at a new design point, $\theta_{\text {new }}$. Compared with the baseline design, the objective function decreases from 0.3656 to 0.2732 . Note that this means that the best-case estimate of the failure probability, among the worst possible of all eligible $e$ 's, is 0.2732 . Compared to the final design discussed in Section 7.11, the improvement here is relatively small. We believe that the small improvement could be caused by the conservativeness of the eligible set, which is mainly due to the lack of data.

For $\theta_{\text {new }}$, the ranges of $R_{1}(\theta), R_{2}(\theta), R_{3}(\theta)$ and $R(\theta)$ (defined in Section 7.9.1) are approximately $[0,0.5935],[0,0.7469],[0,0.5465]$ and $[0,0.8205]$. We could observe from the plots of $R_{\text {min }}$ and $R_{\max }$ that $e_{2}$ has significant different patterns on high values in both plots. According to the trends shown in the plots, we rank the epistemic variables as $e_{2}>e_{3}>e_{1}>e_{4}$.


Figure 7.9: The real part and the imaginary part of $C_{k}^{(1)}$, sagainst the corresponding frequencies.


Figure 7.10: The real part and the imaginary part of $C_{k}^{(2)}$,s against the corresponding frequencies.

### 7.11 Design Tuning (Problem E)

With data sequence $D_{2}=\left\{z^{(i)}(t)\right\}$ for $i=1, \ldots, n_{2}$, we may incorporate the additional information to update our model as before. Similar to Section 7.7, we use FFT to summarize the highly seasonal responses. In particular, we represent $\left(z_{1}^{(i)}(t)\right)_{t=0, \ldots, T}$ and $\left(z_{2}^{(i)}(t)\right)_{t=0, \ldots, T}$ as $z_{1}(t)=\sum_{k=-(N-1)}^{N-1} C_{k}^{(1)} e^{-i k \omega_{0} t}$ and $z_{2}(t)=\sum_{k=-(N-1)}^{N-1} C_{k}^{(2)} e^{-i k \omega_{0} t}$ respectively. As shown in Figures 7.9 and 7.10, the responses in frequency domain have common patterns in the positive and negative peaks. Again we use the values of these peaks and their corresponding frequencies to
summarize $z_{1}$ and $z_{2}$, which leads to 20 extra parameters adding to the 12 parameters extracted from $D_{1}$.

With the extracted parameters from both $D_{1}$ and $D_{2}$, we now update our eligibility set for $E$ by computing $q_{l}^{*}$ 's. We determine eligible $e$ 's with the new threshold $q_{1-0.05 / 32}=1.89$. The values of $q_{l}^{*}$ 's are presented in Figure 7.11. Compared with Figure 7.5, we observe that the trend in $e_{2}$ changes slightly. The $q_{l}^{*}$ 's with high value in $e_{2}$ become higher after introducing the information from $D_{2}$, which indicates that $e$ with higher $e_{2}$ is less eligible. Based on the stronger trend in $e_{2}$ and the observation in Section 7.10, we determine to refine $e_{2}$ on both ends.


Figure 7.11: $q_{l}^{*}$ against each epistemic variable (after incorporating $D_{2}$ ).

In Figure 7.12, we present $q_{l}^{* \prime}$ 's of samples of $e$ for determining the final eligibility set, $E_{2}$. With these updated information, the final design $\theta_{\text {final }}$ is obtained using Algorithm 7.2, where
$\theta_{\text {final }}$. The ranges of $R_{1}\left(\theta_{\text {final }}\right), R_{2}\left(\theta_{\text {final }}\right), R_{3}\left(\theta_{\text {final }}\right)$ and $R\left(\theta_{\text {final }}\right)$ are $[0,0.1676],[0,0.1620]$, [ $0,0.046$ ] and $[0,0.2551]$ respectively. Compared to $\theta_{\text {baseline }}$ and $\theta_{\text {new }}$, the worst-case reliability performance is significantly improved.


Figure 7.12: $q_{l}^{*}$ against each epistemic variable (final refined).

### 7.12 Risk-Based Design (Problem F)

Recall that we create an eligibility set for $e$ in the form of $\left\{e^{(l)}: q_{l}^{*} \leq q_{1-\alpha / m}\right\}$, which provides us $(1-\alpha)$ confidence for covering the truth asymptotically. To reduce $r \%$ volume of the eligibility set, we remove $r \%$ number of eligible points in the set with the larger $q_{l}^{*}$ 's. Note that since the samples of $e$ 's are uniformly generated, the proportion of samples falling into the eligible set is a reasonable estimate of the relative volume of the eligible set. Because larger $q_{l}^{*}$ indicates less
similarity with the true response, the reduced eligibility set maintains more important $e^{(l)}$ 's.
In our setting for $e$, taking risks is equivalent to reducing the confidence level for covering the truth. Let us assume the $r \%$ upper quantile of $q_{l}^{*}$ is $q_{r} \%$. Then the reduced eligibility set can be represented as $\left\{e^{(l)}: q_{l}^{*} \leq q_{r} \%_{0}\right\}$. By finding the $\tilde{\alpha}$ such that $q_{r} \sigma=q_{1-\tilde{\alpha} / m}$, we can find the confidence level $1-\tilde{\alpha}$ that corresponds to each choice of $r \%$. In later discussion, the reduced eligibility set corresponding to risk level $r \%$ is denoted as $E_{r} \%$.

In our experiment, we use $E_{2}$ in Section 7.11 as the baseline. Table 7.1 shows the risk levels and their corresponding confidence levels. The relation between $r \%$ and $\tilde{\alpha}$ highly depends on the value of $q_{l}^{*}$ 's. In our case, we observe that a large portion of $q_{l}^{*}$ 's are close to $q_{1-\alpha / m}$. As a consequence, the reduction in the volume of the set does not lead to a similar extent of reduction in the confidence level. Since the confidence level is almost not changed, we can anticipate that the design results with different $r \%$ in the range of $(0,10)$ will perform similarly.

Table 7.1: The risk levels against their corresponding confidence level.

| $r \%$ | 0 | 2 | 4 | 6 | 8 | 10 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $\left\|E_{r} \sigma_{\%}\right\|$ | 114 | 113 | 110 | 108 | 106 | 104 |
| $q_{r} \%$ | 1.89 | 1.886 | 1.884 | 1.884 | 1.881 | 1.879 |
| $1-\tilde{\alpha}$ | $95 \%$ | $94.9 \%$ | $94.8 \%$ | $94.8 \%$ | $94.7 \%$ | $94.6 \%$ |

With different $r \%$ 's, we construct $E_{r}$ 's using the above approach and implement Algorithm 7.2 to obtain risk-based designs $\theta_{r}{ }^{\prime}$ 's. Then we evaluate the $\theta_{r} \%^{\prime}$ 's by computing the reliability and severity metrics based on their corresponding eligibility set $E_{r} \%$ and also $E_{2}$. The evaluation results using $E_{2}$ are shown in Figure 7.13. We observe from Figure 7.13 that both the reliability or severity metrics are insensitive to the change of $r \%$ (since the results for all considered $r \%$ 's other than $r \%=4 \%$ are similar, we believe that the disparate performance at $r \%=4 \%$ is caused jointly by the randomness in the optimization algorithm and failure probability estimation). In fact, the results are also insensitive to whether using $E_{2}$ or $E_{r \%}$ to compute the metrics. Since the difference can be neglected (the largest difference is smaller than 0.01 ), we omit the results using $E_{r \%}$. From these results, we confirm our conjecture that taking risks would not make much difference in our design approach.


Figure 7.13: The reliability and severity metrics for $\theta_{r} \%$ 's evaluated using $E_{2}$.

### 7.13 Discussion

In this UQ Challenge, we propose a methodology to calibrate model parameters and quantify calibration errors from output data under both aleatory and epistemic uncertainties. The approach utilizes a framework based on an integration of distributionally robust optimization and importance sampling, and operates computationally by solving sampled linear programs. It provides theoretical confidence guarantees on the coverage of the ground truth parameters and distributions. We apply and illustrate our approach to the model calibration and downstream risk analysis tasks in the UQ Challenge. Our approach is drastically different from established Bayesian methodologies, both in the type of guarantee (frequentist versus Bayesian) and computation method (optimization versus posterior sampling). We anticipate much further work in the future in expanding our methodology to more general problems as well as comparing it with the established approaches.

We discuss some immediate future improvements in our implementation in this UQ Challenge. Our procedure relies on several configurations that warrant further explorations. First, the eligibility set geometry is dictated by the choices of the distance metric between distributions and the summary function. Our choice of KS-distance is motivated by nonparametric hypothesis testing that provides asymptotic guarantees. However, since only a finite number of samples is available in practice, its performance can be problem dependent, and other nonparametric test statistics could
be considered. Regarding summary functions, we have chosen them based on the visualization of Fourier transform and justify their number via a balance of representativeness and conservativeness in simultaneous estimation. Our refinement results indicate that our eligibility set performs well in locating $e$, which validates our configurations. Nonetheless, a more rigorous approach to choose both the distance metric and the summary functions is desirable.

Our approach requires sampling a number of $a$ and $e$ for eligibility set and aleatory distribution construction. Since a limited size of naive (uniform) sample might miss important information in a large continuous space and cause high variance, we have used several variance reduction techniques including stratified sampling and common random numbers. We note that the samples for $a$ have a larger effect on designs, since they are used to construct the associated best- and worst-case distributions and the quality of samples can be crucial to correctly evaluating the design performances. Moreover, a good sampling scheme can also lead to higher stability of the stochastic gradient descent algorithm.

Lastly, we note that the conservative nature of our robust approach is reflected in the system design. While our robust approach performs well in locating the eligibility set and providing upper bounds on reliability, directly using these bounds as the objectives for optimizing designs appears over-conservative. Further work on improving the choice of eligibility sets and sampling on $a$ and $e$ could help improve these design performances.

### 7.14 Supplementary A: Proof of Theorem 7.2

This proof is adapted from the proof of Theorem 2 in [202]. We denote $L=d P_{\text {true }} / d P_{0}$. Let $W_{j}=\frac{L\left(a^{(j)}\right)}{\sum_{j=1}^{k} L\left(a^{(j)}\right)}$. For simplicity, we use $\mathbf{y}(a)$ to denote $\left(y\left(a, e_{t r u e}, t\right)\right)_{t=1, \ldots, T}$ and use $\mathbf{y}_{j}$ to denote $\left(y\left(a^{(j)}, e_{\text {true }}, t\right)\right)_{t=1, \ldots, T}$. Then we have that

$$
\begin{aligned}
& \sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\hat{F}_{v}(x)\right| \\
\leq & \sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)\right|+ \\
& \sup _{x \in \mathbb{R}}\left|\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-E_{P_{0}}\left(L(a) I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)\right|+ \\
& \sup _{x \in \mathbb{R}}\left|E_{P_{0}}\left(L(a) I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)-\hat{F}_{v}(x)\right| \\
= & \sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)\right|+ \\
& \sup _{x \in \mathbb{R}}\left|\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-E_{P_{0}}\left(L(a) I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)\right|+ \\
& \sup _{x \in \mathbb{R}}\left|E_{P_{\text {true }}}\left(I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)-\hat{F}_{v}(x)\right| .
\end{aligned}
$$

For the first term, we have that

$$
\begin{align*}
& \sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right) \\
= & \frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)\left(\frac{k}{\sum_{j=1}^{k} L\left(a^{(j)}\right)}-1\right) . \tag{7.14}
\end{align*}
$$

Since $\left\|d P_{\text {true }} / d P_{0}\right\|_{\infty} \leq C$, we get that $\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right) \leq C$. Moreover, we know that $E_{P_{0}}(L)=1$ and $\operatorname{var}_{P_{0}}(L)<\infty$, and thus

$$
\sqrt{k}\left(\frac{k}{\sum_{j=1}^{k} L\left(a^{(j)}\right)}-1\right) \xrightarrow{d} N\left(0, \operatorname{var}_{P_{0}}(L)\right)
$$

where $\xrightarrow{d .}$ denotes convergence in distribution or weak convergence.

Hence, we get from Eq. (7.14) that

$$
\sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)\right|=O_{p}(1 / \sqrt{k})
$$

For the second term, following the proof in [202], we know that

$$
\left\{\sqrt{k}\left(\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-E_{P_{0}}\left(L(a) I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)\right)\right\} \xrightarrow{d .}\{G(x)\}
$$

in $\ell^{\infty}\left(\left\{a \mapsto L(a) I\left(S_{v}(\mathbf{y}(a)) \leq x\right): x \in \mathbb{R}\right\}\right)$ and $G$ is a Gaussian process. Therefore, we get that

$$
\sup _{x \in \mathbb{R}}\left|\frac{1}{k} \sum_{j=1}^{k} L\left(a^{(j)}\right) I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-E_{P_{0}}\left(L(a) I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)\right|=O_{p}(1 / \sqrt{k})
$$

Finally, it is known that

$$
\sqrt{n_{1}} \sup _{x \in \mathbb{R}}\left|E_{P_{\text {true }}}\left(I\left(S_{v}(\mathbf{y}(a)) \leq x\right)\right)-\hat{F}_{v}(x)\right| \xrightarrow{d .} \sup _{x \in[0,1]}\left|B B\left(F_{\text {true }, v}(x)\right)\right| .
$$

Combining the above results, we get that for each $v=1, \ldots, m$,

$$
\limsup _{n_{1} \rightarrow \infty, k / n_{1} \rightarrow \infty} \mathbb{P}\left(\sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\hat{F}_{v}(x)\right|>\frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}\right) \leq \frac{\alpha}{m}
$$

and hence

$$
\liminf _{n_{1} \rightarrow \infty, k / n_{1} \rightarrow \infty} \mathbb{P}\left(e_{\text {true }} \in E\right) \geq 1-\alpha
$$

since $e_{\text {true }} \in E$ if and only if $\sup _{x \in \mathbb{R}}\left|\sum_{j=1}^{k} W_{j} I\left(S_{v}\left(\mathbf{y}_{j}\right) \leq x\right)-\hat{F}_{v}(x)\right| \leq \frac{q_{1-\alpha / m}}{\sqrt{n_{1}}}$ for any $v=$ $1, \ldots, m$.

## References

[1] P. Koopman and M. Wagner, "Autonomous vehicle safety: An interdisciplinary challenge," IEEE Intelligent Transportation Systems Magazine, vol. 9, no. 1, pp. 90-96, 2017.
[2] M. J. Kochenderfer, J. E. Holland, and J. P. Chryssanthacopoulos, "Next-generation airborne collision avoidance system," Massachusetts Institute of Technology-Lincoln Laboratory Lexington United States, Tech. Rep., 2012.
[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2722-2730.
[4] J. Van Brummelen, M. O’Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," Transportation Research Part C: Emerging Technologies, vol. 89, pp. 384-406, 2018.
[5] R. Glasius, A. Komoda, and S. C. Gielen, "Neural network dynamics for path planning and obstacle avoidance," Neural Networks, vol. 8, no. 1, pp. 125-133, 1995.
[6] S. X. Yang and C. Luo, "A neural network approach to complete coverage path planning," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 1, pp. 718-724, 2004.
[7] N. A. Spielberg, M. Brown, N. R. Kapania, J. C. Kegelman, and J. C. Gerdes, "Neural network vehicle models for high-performance automated driving," Science Robotics, vol. 4, no. 28, 2019.
[8] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in Advances in Neural Information Processing Systems, 2006, pp. 739-746.
[9] Z. Chen and X. Huang, "End-to-end learning for lane keeping of self-driving cars," in 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2017, pp. 1856-1860.
[10] N. T. S. Board, Preliminary report, highway, hwy18mh010, 2018.
[11] N. T. S. Board, Collision between car operating with partial driving automation and trucktractor semitrailer delray beach, florida, march 1, 2019, 2019.
[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
[13] L. Weng et al., "PROVEN: Verifying robustness of neural networks with a probabilistic approach," in International Conference on Machine Learning, PMLR, 2019, pp. 67276736.
[14] S. Webb, T. Rainforth, Y. W. Teh, and M. P. Kumar, "A statistical approach to assessing neural network robustness," arXiv preprint arXiv:1811.07209, 2018.
[15] B. Wang, S. Webb, and T. Rainforth, "Statistically robust neural network classification," in Uncertainty in Artificial Intelligence, PMLR, 2021, pp. 1735-1745.
[16] L. Fraade-Blanar, M. S. Blumenthal, J. M. Anderson, and N. Kalra, Measuring Automated Vehicle Safety: Forging a Framework. RAND Corporation, 2018.
[17] D. Zhao et al., "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 3, pp. 595-607, 2016.
[18] Z. Huang, H. Lam, D. J. LeBlanc, and D. Zhao, "Accelerated evaluation of automated vehicles using piecewise mixture models," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 9, pp. 2845-2855, 2017.
[19] M. O'Kelly, A. Sinha, H. Namkoong, R. Tedrake, and J. C. Duchi, "Scalable end-to-end autonomous vehicle testing via rare-event simulation," in Advances in Neural Information Processing Systems, 2018, pp. 9827-9838.
[20] D. Zhao, X. Huang, H. Peng, H. Lam, and D. J. LeBlanc, "Accelerated evaluation of automated vehicles in car-following maneuvers," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 3, pp. 733-744, 2017.
[21] M. Arief et al., "Deep probabilistic accelerated evaluation: A robust certifiable rare-event simulation methodology for black-box safety-critical systems," in International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 595-603.
[22] J. A. Bucklew, Introduction to Rare Event Simulation. Springer, 2004, vol. 5.
[23] S. Juneja and P. Shahabuddin, "Rare-event simulation techniques: An introduction and recent advances," Handbooks in Operations Research and Management Science, vol. 13, pp. 291-350, 2006.
[24] J. S. Sadowsky, "Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue," IEEE Transactions on Automatic Control, vol. 36, no. 12, pp. 13831394, 1991.
[25] D. P. Kroese and V. F. Nicola, "Efficient estimation of overflow probabilities in queues with breakdowns," Performance Evaluation, vol. 36, pp. 471-484, 1999.
[26] J. Blanchet, P. Glynn, and H. Lam, "Rare event simulation for a slotted time $M / G / s$ model," Queueing Systems, vol. 63, no. 1, pp. 33-57, 2009.
[27] J. Blanchet and H. Lam, "Rare-event simulation for many-server queues," Mathematics of Operations Research, vol. 39, no. 4, pp. 1142-1178, 2014.
[28] J. Blanchet and M. Mandjes, "Rare event simulation for queues," in Rare Event Simulation Using Monte Carlo Methods, Chapter 5, 2009, pp. 87-124.
[29] R. Szechtman and P. W. Glynn, "Rare-event simulation for infinite server queues," in Proceedings of the Winter Simulation Conference, IEEE, vol. 1, 2002, pp. 416-423.
[30] A. Ridder, "Importance sampling algorithms for first passage time probabilities in the infinite server queue," European Journal of Operational Research, vol. 199, no. 1, pp. 176186, 2009.
[31] P. Dupuis, K. Leder, and H. Wang, "Importance sampling for weighted-serve-the-longestqueue," Mathematics of Operations Research, vol. 34, no. 3, pp. 642-660, 2009.
[32] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," IEEE Transactions on Automatic Control, vol. 34, no. 1, pp. 54-66, 1989.
[33] B. Chen, J. Blanchet, C.-H. Rhee, and B. Zwart, "Efficient rare-event simulation for multiple jump events in regularly varying random walks and compound Poisson processes," Mathematics of Operations Research, vol. 44, no. 3, pp. 919-942, 2019.
[34] P. Glasserman, Monte Carlo Methods in Financial Engineering. Springer, 2004, vol. 53.
[35] P. Glasserman, W. Kang, and P. Shahabuddin, "Fast simulation of multifactor portfolio credit risk," Operations Research, vol. 56, no. 5, pp. 1200-1217, 2008.
[36] P. Glasserman and J. Li, "Importance sampling for portfolio credit risk," Management Science, vol. 51, no. 11, pp. 1643-1656, 2005.
[37] S. Asmussen and H. Albrecher, Ruin Probabilities. World Scientific, 2010, vol. 14.
[38] S. Asmussen, "Conjugate processes and the simulation of ruin problems," Stochastic Processes and their Applications, vol. 20, no. 2, pp. 213-229, 1985.
[39] J. F. Collamore, "Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors," The Annals of Applied Probability, vol. 12, no. 1, pp. 382-421, 2002.
[40] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 5, pp. 43-85, 1995.
[41] G. Rubino and B. Tuffin, "Markovian models for dependability analysis," in Rare Event Simulation Using Monte Carlo Methods, Chapter 6, 2009, pp. 125-144.
[42] B. Tuffin, "On numerical problems in simulations of highly reliable Markovian systems," in First International Conference on the Quantitative Evaluation of Systems, 2004. QEST 2004. Proceedings., IEEE, 2004, pp. 156-164.
[43] V. F. Nicola, M. K. Nakayama, P. Heidelberger, and A. Goyal, "Fast simulation of highly dependable systems with general failure and repair processes," IEEE Transactions on Computers, vol. 42, no. 12, pp. 1440-1452, 1993.
[44] V. F. Nicola, P. Shahabuddin, and M. K. Nakayama, "Techniques for fast simulation of models of highly dependable systems," IEEE Transactions on Reliability, vol. 50, no. 3, pp. 246-264, 2001.
[45] P. Grassberger, "Go with the winners: A general Monte Carlo strategy," Computer Physics Communications, vol. 147, no. 1-2, pp. 64-70, 2002.
[46] W. Sandmann, "Rare event simulation methodologies in systems biology," in Rare Event Simulation Using Monte Carlo Methods, Chapter 11, 2009, pp. 243-266.
[47] P. Dupuis, K. Spiliopoulos, and H. Wang, "Importance sampling for multiscale diffusions," Multiscale Modeling \& Simulation, vol. 10, no. 1, pp. 1-27, 2012.
[48] E. Vanden-Eijnden and J. Weare, "Rare event simulation of small noise diffusions," Communications on Pure and Applied Mathematics, vol. 65, no. 12, pp. 1770-1803, 2012.
[49] M. Bayati, J. H. Kim, and A. Saberi, "A sequential algorithm for generating random graphs," Algorithmica, vol. 58, pp. 860-910, 2010.
[50] J. H. Blanchet, "Efficient importance sampling for binary contingency tables," The Annals of Applied Probability, pp. 949-982, 2009.
[51] D. Siegmund, "Importance sampling in the Monte Carlo study of sequential tests," The Annals of Statistics, pp. 673-684, 1976.
[52] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," Management Science, vol. 35, no. 11, pp. 1367-1392, 1989.
[53] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient Monte Carlo estimation," IEEE transactions on Information Theory, vol. 36, no. 3, pp. 579-588, 1990.
[54] S. Asmussen and P. W. Glynn, Stochastic Simulation: Algorithms and Analysis. Springer Science \& Business Media, 2007, vol. 57.
[55] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[56] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep Learning. MIT Press Cambridge, 2016, vol. 1.
[57] A. Dieker and M. Mandjes, "On asymptotically efficient simulation of large deviation probabilities," Advances in Applied Probability, vol. 37, no. 2, pp. 539-552, 2005.
[58] V. Tjeng, K. Xiao, and R. Tedrake, "Evaluating robustness of neural networks with mixed integer programming," arXiv preprint arXiv:1711.07356, 2017.
[59] V. V. Mišić, "Optimization of tree ensembles," Operations Research, vol. 68, no. 5, pp. 16051624, 2020.
[60] H. Honnappa, R. Pasupathy, and P. Jaiswal, "Dominating points of Gaussian extremes," arXiv preprint arXiv:1810.12132, 2018.
[61] K. D. Kusano and H. C. Gabler, "Safety benefits of forward collision warning, brake assist, and autonomous braking systems in rear-end collisions," IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 4, pp. 1546-1555, 2012.
[62] E. Hashorva and J. Huesler, "On multivariate Gaussian tails," Annals of the Institute of Statistical Mathematics, vol. 55, pp. 507-522, Feb. 2003.
[63] R. Y. Rubinstein and D. P. Kroese, Simulation and the Monte Carlo Method. John Wiley \& Sons, 2016, vol. 10.
[64] J. Blanchet and H. Lam, "State-dependent importance sampling for rare-event simulation: An overview and recent advances," Surveys in Operations Research and Management Science, vol. 17, no. 1, pp. 38-59, 2012.
[65] A. B. Owen, Y. Maximov, and M. Chertkov, "Importance sampling the union of rare events with an application to power systems analysis," Electronic Journal of Statistics, vol. 13, no. 1, pp. 231-254, 2019.
[66] J. Uesato et al., "Rigorous agent evaluation: An adversarial approach to uncover catastrophic failures," arXiv preprint arXiv:1812.01647, 2018.
[67] R. Y. Rubinstein and D. P. Kroese, The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo simulation, and Machine Learning. Springer, 2004, vol. 133.
[68] P.-T. De Boer, V. F. Nicola, and R. Y. Rubinstein, "Adaptive importance sampling simulation of queueing networks," in 2000 Winter Simulation Conference Proceedings, IEEE, vol. 1, 2000, pp. 646-655.
[69] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the crossentropy method," Annals of Operations Research, vol. 134, no. 1, pp. 19-67, 2005.
[70] R. Y. Rubinstein, "Optimization of computer simulation models with rare events," European Journal of Operational Research, vol. 99, no. 1, pp. 89-112, 1997.
[71] C. Kollman, K. Baggerly, D. Cox, and R. Picard, "Adaptive importance sampling on discrete Markov chains," Annals of Applied Probability, pp. 391-412, 1999.
[72] P. Y. Desai and P. W. Glynn, "A Markov chain perspective on adaptive Monte Carlo algorithms," in Proceeding of the 2001 Winter Simulation Conference, IEEE, vol. 1, 2001, pp. 379-384.
[73] T. I. Ahamed, V. S. Borkar, and S Juneja, "Adaptive importance sampling technique for Markov chains using stochastic approximation," Operations Research, vol. 54, no. 3, pp. 489504, 2006.
[74] V. S. Borkar, S. Juneja, and A. A. Kherani, "Performance analysis conditioned on rare events: An adaptive simulation scheme," Communications in Information and Systems, vol. 3, no. 4, pp. 259-278, 2004.
[75] Z. I. Botev, P. L'Ecuyer, and B. Tuffin, "Markov chain importance sampling with applications to rare event probability estimation," Statistics and Computing, vol. 23, no. 2, pp. 271-285, 2013.
[76] A. W. Grace, D. P. Kroese, and W. Sandmann, "Automated state-dependent importance sampling for Markov jump processes via sampling from the zero-variance distribution," Journal of Applied Probability, vol. 51, no. 3, pp. 741-755, 2014.
[77] J. C. Chan and D. P. Kroese, "Improved cross-entropy method for estimation," Statistics and Computing, vol. 22, no. 5, pp. 1031-1040, 2012.
[78] J. Blanchet and P. Glynn, "Efficient rare-event simulation for the maximum of heavy-tailed random walks," The Annals of Applied Probability, pp. 1351-1378, 2008.
[79] J. H. Blanchet and J. Liu, "State-dependent importance sampling for regularly varying random walks," Advances in Applied Probability, vol. 40, no. 4, pp. 1104-1128, 2008.
[80] P. Dupuis, K. Leder, and H. Wang, "Importance sampling for sums of random variables with regularly varying tails," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 17, no. 3, 14-es, 2007.
[81] J. Blanchet, H. Lam, and B. Zwart, "Efficient rare-event simulation for perpetuities," Stochastic Processes and their Applications, vol. 122, no. 10, pp. 3361-3392, 2012.
[82] K. R. Murthy, S. Juneja, and J. Blanchet, "State-independent importance sampling for random walks with regularly varying increments," Stochastic Systems, vol. 4, no. 2, pp. 321374, 2015.
[83] H. Hult and J. Svensson, "On importance sampling with mixtures for random walks with heavy tails," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 22, no. 2, pp. 1-21, 2012.
[84] S. Asmussen and K. Binswanger, "Simulation of ruin probabilities for subexponential claims," ASTIN Bulletin: The Journal of the IAA, vol. 27, no. 2, pp. 297-318, 1997.
[85] S. Asmussen and D. P. Kroese, "Improved algorithms for rare event simulation with heavy tails," Advances in Applied Probability, vol. 38, no. 2, pp. 545-558, 2006.
[86] P. L'Ecuyer, F. L. Gland, P. Lezaud, and B. Tuffin, "Splitting techniques," in Rare Event Simulation Using Monte Carlo Methods, Chapter 3, 2009, pp. 39-62.
[87] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic, "Multilevel splitting for estimating rare event probabilities," Operations Research, vol. 47, no. 4, pp. 585-600, 1999.
[88] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic, "A large deviations perspective on the efficiency of multilevel splitting," IEEE Transactions on Automatic Control, vol. 43, no. 12, pp. 1666-1679, 1998.
[89] T. Dean and P. Dupuis, "Splitting for rare event simulation: A large deviation approach to design and analysis," Stochastic Processes and their Applications, vol. 119, no. 2, pp. 562 -587, 2009.
[90] Z. I. Botev and P. L'Ecuyer, "Sampling conditionally on a rare event via generalized splitting," INFORMS Journal on Computing, 2020.
[91] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Artificial Intelligence Safety and Security, Chapman and Hall/CRC, 2018, pp. 99-112.
[92] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," arXiv preprint arXiv:1611.01236, 2016.
[93] H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li, "Object classification using CNNbased fusion of vision and LIDAR in autonomous vehicle environment," IEEE Transactions on Industrial Informatics, vol. 14, no. 9, pp. 4224-4231, 2018.
[94] T. Dreossi, A. Donzé, and S. A. Seshia, "Compositional falsification of cyber-physical systems with machine learning components," Journal of Automated Reasoning, vol. 63, no. 4, pp. 1031-1053, 2019.
[95] M. Rajchl et al., "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," IEEE Transactions on Medical Imaging, vol. 36, no. 2, pp. 674-683, 2016.
[96] A. Deo and K. Murthy, "Efficient black-box importance sampling for VaR and CVaR estimation," in 2021 Winter Simulation Conference (WSC), IEEE, 2021, pp. 1-12.
[97] F. Bravo, C. Rudin, Y. Shaposhnik, and Y. Yuan, "Simple rules for predicting congestion risk in queueing systems: Application to ICUs," Available at SSRN 3384148, 2019.
[98] M. S. Hagen, J. K. Jopling, T. G. Buchman, and E. K. Lee, "Priority queuing models for hospital intensive care units and impacts to severe case patients," in AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol. 2013, 2013, p. 841.
[99] C. Sawade, N. Landwehr, and T. Scheffer, "Active estimation of f-measures," Advances in Neural Information Processing Systems, vol. 23, pp. 2083-2091, 2010.
[100] C Van Rijsbergen, "Information retrieval: Theory and practice," in Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems, 1979, pp. 1-14.
[101] F. Poms et al., "Low-shot validation: Active importance sampling for estimating classifier performance on rare categories," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10705-10714.
[102] P. L'ecuyer, J. H. Blanchet, B. Tuffin, and P. W. Glynn, "Asymptotic robustness of estimators in rare-event simulation," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 20, no. 1, pp. 1-41, 2010.
[103] P. Glasserman and Y. Wang, "Counterexamples in importance sampling for large deviations probabilities," The Annals of Applied Probability, vol. 7, no. 3, pp. 731-746, 1997.
[104] S. A. Vavasis, "Complexity theory: Quadratic programming," in Encyclopedia of Optimization, Springer, 2006, pp. 304-307.
[105] P. Bonami, M. Kilinç, and J. Linderoth, "Algorithms and software for convex mixed integer nonlinear programs," in Mixed Integer Nonlinear Programming, Springer, 2012, pp. 1-39.
[106] H. Y. He and A. B. Owen, "Optimal mixture weights in multiple importance sampling," arXiv preprint arXiv:1411.3954, 2014.
[107] R. J. Adler, J. H. Blanchet, and J. Liu, "Efficient Monte Carlo for high excursions of Gaussian random fields," The Annals of Applied Probability, vol. 22, no. 3, pp. 1167-1214, 2012.
[108] T. Hastie, R. Tibshirani, and J. Friedman, "Random forests," in The Elements of Statistical Learning, Springer, 2009, pp. 587-604.
[109] D. Bertsimas and J. N. Tsitsiklis, Introduction to Linear Optimization. Athena Scientific Belmont, MA, 1997, vol. 6.
[110] L. O. Chua and L. Yang, "Cellular neural networks: Theory," IEEE Transactions on Circuits and Systems, vol. 35, no. 10, pp. 1257-1272, 1988.
[111] Z. Huang, H. Lam, and D. Zhao, "Rare-event simulation without structural information: A learning-based approach," in 2018 Winter Simulation Conference (WSC), IEEE, 2018, pp. 1826-1837.
[112] A. Asuncion and D. Newman, UCI machine learning repository, 2007.
[113] R. Bock et al., "Methods for multidimensional event classification: A case study using images from a Cherenkov gamma-ray telescope," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 516, no. 2-3, pp. 511-528, 2004.
[114] P. Savickỳ and E. Kotrc, "Experimental study of leaf confidences for random forest," in Proceedings of the 16th Symposium on Computational Statistics, Prague, Czech Republic, 2004, pp. 1767-1774.
[115] J. Dvořák and P. Savickỳ, "Softening splits in decision trees using simulated annealing," in International Conference on Adaptive and Natural Computing Algorithms, Springer, 2007, pp. 721-729.
[116] G. Rubino and B. Tuffin, Rare Event Simulation Using Monte Carlo Methods. John Wiley \& Sons, 2009.
[117] P. Dupuis, A. D. Sezer, and H. Wang, "Dynamic importance sampling for queueing networks," The Annals of Applied Probability, vol. 17, no. 4, pp. 1306-1346, 2007.
[118] J. Blanchet and M. Mandjes, "Editorial: Rare-event simulation for queues," Queueing Systems, vol. 57, no. 2-3, pp. 57-59, 2007.
[119] E. E. Lewis and F. Böhm, "Monte Carlo simulation of Markov unreliability models," Nuclear Engineering and Design, vol. 77, no. 1, pp. 49-62, 1984.
[120] A. Goyal, P. Shahabuddin, P. Heidelberger, V. F. Nicola, and P. W. Glynn, "A unified framework for simulating Markovian models of highly dependable systems," IEEE Transactions on Computers, vol. 41, no. 1, pp. 36-51, 1992.
[121] J. A. Carrasco, "Failure distance-based simulation of repairable fault-tolerant systems," in Computer Performance Evaluation: Modelling Techniques and Tools: proceedings of the Fifth International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, Torino, Italy, 13-15 February 1991, Elsevier, 1992, pp. 351-365.
[122] P. Shahabuddin, "Importance sampling for the simulation of highly reliable Markovian systems," Management Science, vol. 40, no. 3, pp. 333-352, 1994.
[123] W. Ding, B. Chen, B. Li, K. J. Eun, and D. Zhao, "Multimodal safety-critical scenarios generation for decision-making algorithms evaluation," IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 1551-1558, 2021.
[124] M. Kratsios, The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. National Science and Technology Council (US), 2019.
[125] Y. Bai, Z. Huang, H. Lam, and D. Zhao, "Rare-event simulation for neural network and random forest predictors," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 32, no. 3, pp. 1-33, 2022.
[126] P. J. Smith, M. Shafi, and H. Gao, "Quick simulation: A review of importance sampling techniques in communications systems," IEEE Journal on Selected Areas in Communications, vol. 15, no. 4, pp. 597-613, 1997.
[127] J.-C. Chen, D. Lu, J. S. Sadowsky, and K. Yao, "On importance sampling in digital communications. I. fundamentals," IEEE Journal on Selected Areas in Communications, vol. 11, no. 3, pp. 289-299, 1993.
[128] X.-W. Zhang, P. W. Glynn, K. Giesecke, and J. Blanchet, "Rare event simulation for a generalized Hawkes process," in Proceedings of the 2009 Winter Simulation Conference (WSC), IEEE, 2009, pp. 1291-1298.
[129] J. Blanchet and P. Glynn, "Efficient rare event simulation of continuous time Markovian perpetuities," in Proceedings of the 2009 Winter Simulation Conference (WSC), IEEE, 2009, pp. 444-451.
[130] A. Dembo and O. Zeitouni, Large Deviations Techniques and Applications. Springer Science \& Business Media, 2009, vol. 38.
[131] Z. Huang, H. Lam, and D. Zhao, "Designing importance samplers to simulate machine learning predictors via optimization," in 2018 Winter Simulation Conference (WSC), IEEE, 2018, pp. 1730-1741.
[132] B. Tuffin and A. Ridder, "Probabilistic bounded relative error for rare event simulation learning techniques," in Proceedings of the 2012 Winter Simulation Conference (WSC), IEEE, 2012, pp. 1-12.
[133] A. Maurer and M. Pontil, "Empirical Bernstein bounds and sample variance penalization," arXiv preprint arXiv:0907.3740, 2009.
[134] A. J. McNeil, R. Frey, and P. Embrechts, Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition. Princeton University Press, 2015.
[135] P. Embrechts, T. Mikosch, and C. Klüppelberg, Modelling Extremal Events: for Insurance and Finance. Springer-Verlag, 1997.
[136] A. C. Davison and R. L. Smith, "Models for exceedances over high thresholds," Journal of the Royal Statistical Society: Series B (Methodological), vol. 52, no. 3, pp. 393-425, 1990.
[137] R. L. Smith, "Threshold methods for sample extremes," in Statistical Extremes and Applications, Springer, 1984, pp. 621-638.
[138] M. Villén-Altamirano and J. Villén-Altamirano, "RESTART: A straightforward method for fast simulation of rare events," in Proceedings of the 1994 Winter Simulation Conference, IEEE, 1994, pp. 282-289.
[139] S.-K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," Probabilistic Engineering Mechanics, vol. 16, no. 4, pp. 263-277, 2001.
[140] T.-W. Weng et al., "Evaluating the robustness of neural networks: An extreme value theory approach," arXiv preprint arXiv:1801.10578, 2018.
[141] Q.-M. Shao and Q. Wang, "Self-normalized limit theorems: A survey," Probability Surveys, vol. 10, pp. 69-93, 2013.
[142] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," The American Statistician, vol. 52, no. 2, pp. 119-126, 1998.
[143] Q. Wang and B.-Y. Jing, "An exponential nonuniform Berry-Esseen bound for self-normalized sums," The Annals of Probability, vol. 27, no. 4, pp. 2068-2088, 1999.
[144] Q. Wang and P. Hall, "Relative errors in central limit theorems for student's $t$ statistic, with applications," Statistica Sinica, pp. 343-354, 2009.
[145] A. A. Balkema and L. De Haan, "Residual life time at great age," The Annals of Probability, vol. 2, no. 5, pp. 792-804, 1974.
[146] J. Pickands III, "Statistical inference using extreme order statistics," The Annals of Statistics, pp. 119-131, 1975.
[147] A. M. Law and W. D. Kelton, Simulation Modeling and Analysis. Mcgraw-hill New York, 2007, vol. 3.
[148] J Banks, J. Carson, B. Nelson, and D. Nicol, Discrete-Event System Simulation, 5th. Prentice Hall Englewood Cliffs, NJ, USA, 2009.
[149] J. Go, B. Browning, and M. Veloso, "Accurate and flexible simulation for dynamic, visioncentric robots," in Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, vol. 3, 2004, pp. 1388-1389.
[150] L. Žlajpah, "Simulation in robotics," Mathematics and Computers in Simulation, vol. 79, no. 4, pp. 879-897, 2008.
[151] N. Ruiz, S. Schulter, and M. Chandraker, "Learning to simulate," in International Conference on Learning Representations, 2019.
[152] R. R. Barton et al., "Panel discussion on current issues in input modeling," in Proceedings of the 2002 Winter Simulation Conference, IEEE, 2002, pp. 353-369.
[153] S. G. Henderson, "Input modeling: Input model uncertainty: Why do we care and what should we do about it?" In Proceedings of the 2003 Winter Simulation Conference, IEEE, 2003, pp. 90-100.
[154] S. E. Chick, "Bayesian ideas and discrete event simulation: Why, what and how," in Proceedings of the 2006 Winter Simulation Conference, IEEE, 2006, pp. 96-106.
[155] R. R. Barton, "Tutorial: Input uncertainty in outout analysis," in Proceedings of the 2012 Winter Simulation Conference (WSC), IEEE, 2012, pp. 1-12.
[156] E. Song, B. L. Nelson, and C. D. Pegden, "Advanced tutorial: Input uncertainty quantification," in Proceedings of the 2014 Winter Simulation Conference, IEEE, 2014, pp. 162176.
[157] H. Lam, "Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation," in Proceedings of the 2016 Winter Simulation Conference, IEEE, 2016, pp. 178192.
[158] B. Nelson, Foundations and Methods of Stochastic Simulation: A First Course. Springer Science \& Business Media, 2013.
[159] C. G. Corlu, A. Akcay, and W. Xie, "Stochastic simulation under input uncertainty: A review," Operations Research Perspectives, p. 100 162, 2020.
[160] R. R. Barton, H. Lam, and E. Song, "Input uncertainty in stochastic simulation," in The Palgrave Handbook of Operations Research, Springer, 2022, pp. 573-620.
[161] R. R. Barton and L. W. Schruben, "Uniform and bootstrap resampling of empirical distributions," in Proceedings of the 25th conference on Winter simulation, 1993, pp. 503508.
[162] R. C. Cheng and W. Holland, "Sensitivity of computer simulation experiments to errors in input data," Journal of Statistical Computation and Simulation, vol. 57, no. 1-4, pp. 219241, 1997.
[163] E. Song and B. L. Nelson, "Quickly assessing contributions to input uncertainty," IIE Transactions, vol. 47, no. 9, pp. 893-909, 2015.
[164] H. Lam and H. Qian, "Subsampling to enhance efficiency in input uncertainty quantification," Operations Research, vol. 70, no. 3, pp. 1891-1913, 2022.
[165] H. Lam, "Cheap bootstrap for input uncertainty quantification," in 2022 Winter Simulation Conference (WSC), IEEE, 2022, pp. 2318-2329.
[166] R. C. Cheng and W. Holland, "Two-point methods for assessing variability in simulation output," Journal of Statistical Computation Simulation, vol. 60, no. 3, pp. 183-205, 1998.
[167] Y Lin, E Song, and B. Nelson, "Single-experiment input uncertainty," Journal of Simulation, vol. 9, no. 3, pp. 249-259, 2015.
[168] E. Song and B. L. Nelson, "Input-output uncertainty comparisons for discrete optimization via simulation," Operations Research, vol. 67, no. 2, pp. 562-576, 2019.
[169] H. Lam and H. Qian, "Random perturbation and bagging to quantify input uncertainty," in 2019 Winter Simulation Conference (WSC), IEEE, 2019, pp. 320-331.
[170] P. Glasserman and X. Xu, "Robust risk measurement and model risk," Quantitative Finance, vol. 14, no. 1, pp. 29-58, 2014.
[171] Z. Hu, J. Cao, and L. J. Hong, "Robust simulation of global warming policies using the DICE model," Management Science, vol. 58, no. 12, pp. 2190-2206, 2012.
[172] S. Ghosh and H. Lam, "Robust analysis in stochastic simulation: Computation and performance guarantees," Operations Research, vol. 67, no. 1, pp. 232-249, 2019.
[173] H. Lam, "Robust sensitivity analysis for stochastic systems," Mathematics of Operations Research, vol. 41, no. 4, pp. 1248-1275, 2016.
[174] H. Lam, "Sensitivity to serial dependency of input processes: A robust approach," Management Science, vol. 64, no. 3, pp. 1311-1327, 2018.
[175] H. Lam and H. Qian, "The empirical likelihood approach to simulation input uncertainty," in Proceedings of the 2016 Winter Simulation Conference, IEEE, 2016, pp. 791-802.
[176] P. W. Glynn and H. Lam, "Constructing simulation output intervals under input uncertainty via data sectioning," in Proceedings of the 2018 Winter Simulation Conference, IEEE, 2018, pp. 1551-1562.
[177] S. E. Chick, "Input distribution selection for simulation experiments: Accounting for input uncertainty," Operations Research, vol. 49, no. 5, pp. 744-758, 2001.
[178] F. Zouaoui and J. R. Wilson, "Accounting for parameter uncertainty in simulation input modeling," IIE Transactions, vol. 35, no. 9, pp. 781-792, 2003.
[179] R. R. Barton, B. L. Nelson, and W. Xie, "Quantifying input uncertainty via simulation confidence intervals," INFORMS Journal on Computing, vol. 26, no. 1, pp. 74-87, 2013.
[180] H. Zhu, T. Liu, and E. Zhou, "Risk quantification in stochastic simulation under input uncertainty," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 30, no. 1, pp. 1-24, 2020.
[181] B. L. Nelson, "'Some tactical problems in digital simulation' for the next 10 years," Journal of Simulation, vol. 10, no. 1, pp. 2-11, 2016.
[182] R. G. Sargent, "Verification and validation of simulation models," in Proceedings of the 37th Winter Simulation Conference, IEEE, 2005, pp. 130-143.
[183] J. P. Kleijnen, "Verification and validation of simulation models," European Journal of Operational Research, vol. 82, no. 1, pp. 145-162, 1995.
[184] O. Balci and R. G. Sargent, "Some examples of simulation model validation using hypothesis testing," in Proceedings of the 14th Winter Simulation Conference, vol. 2, 1982, pp. 621-629.
[185] L. W. Schruben, "Establishing the credibility of simulations," Simulation, vol. 34, no. 3, pp. 101-105, 1980.
[186] A. Tarantola, Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, 2005.
[187] M. C. Kennedy and A. O’Hagan, "Bayesian calibration of computer models,"Journal of the Royal Statistical Society. Series B, Statistical Methodology, pp. 425-464, 2001.
[188] S. Conti and A. O'Hagan, "Bayesian emulation of complex multi-output and dynamic computer models,"Journal of Statistical Planning and Inference, vol. 140, no. 3, pp. 640651, 2010.
[189] M. Plumlee and H. Lam, "Learning stochastic model discrepancy," in Proceedings of the 2016 Winter Simulation Conference, IEEE, 2016, pp. 413-424.
[190] H. Lam, X. Zhang, and M. Plumlee, "Improving prediction from stochastic simulation via model discrepancy learning," in Proceedings of the 2017 Winter Simulation Conference, IEEE, 2017, pp. 1808-1819.
[191] R. Tuo and C. J. Wu, "Efficient calibration for imperfect computer models," The Annals of Statistics, vol. 43, no. 6, pp. 2331-2352, 2015.
[192] R. Tuo and C. Jeff Wu, "A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties," SIAM/ASA Journal on Uncertainty Quantification, vol. 4, no. 1, pp. 767-795, 2016.
[193] R. K. W. Wong, C. B. Storlie, and T. C. M. Lee, "A frequentist approach to computer model calibration," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 79, no. 2, pp. 635-648, 2017.
[194] D. Byrd, M. Hybinette, and T. H. Balch, "ABIDES: Towards high-fidelity multi-agent market simulation," in Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation, 2020, pp. 11-22.
[195] S. Vyetrenko et al., "Get real: Realism metrics for robust limit order book market simulations," in International Conference on Artificial Intelligence in Finance, 2020.
[196] R. C. Larson, "The queue inference engine: Deducing queue statistics from transactional data," Management Science, vol. 36, no. 5, pp. 586-601, 1990.
[197] W. Whitt, "Approximating a point process by a renewal process: The view through a queue, an indirect approach," Management Science, vol. 27, no. 6, pp. 619-636, 1981.
[198] A. Mandelbaum and S. Zeltyn, "Estimating characteristics of queueing networks using transactional data," Queueing Systems, vol. 29, no. 1, pp. 75-127, 1998.
[199] N. Bingham and S. M. Pitts, "Non-parametric estimation for the M/G/ $\infty$ queue," Annals of the Institute of Statistical Mathematics, vol. 51, no. 1, pp. 71-97, 1999.
[200] P. Hall and J. Park, "Nonparametric inference about service time distribution from indirect measurements," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 66, no. 4, pp. 861-875, 2004.
[201] E. Moulines, F. Roueff, A. Souloumiac, and T. Trigano, "Nonparametric inference of photon energy distribution from indirect measurement," Bernoulli, vol. 13, no. 2, pp. 365-388, 2007.
[202] A. Goeva, H. Lam, H. Qian, and B. Zhang, "Optimization-based calibration of simulation input models," Operations Research, vol. 67, no. 5, pp. 1362-1382, 2019.
[203] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," Operations Research, vol. 58, no. 4-Part-1, pp. 902-917, 2010.
[204] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," Operations Research, vol. 58, no. 3, pp. 595-612, 2010.
[205] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," Operations Research, vol. 62, no. 6, pp. 1358-1376, 2014.
[206] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," Management Science, vol. 59, no. 2, pp. 341-357, 2013.
[207] Y. Bai and H. Lam, "Calibrating input parameters via eligibility sets," in Proceedings of the 2020 Winter Simulation Conference, IEEE, 2020, pp. 2114-2125.
[208] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," Proceedings of the National Academy of Sciences, vol. 117, no. 48, pp. 30 055-30 062, 2020.
[209] C. Wunsch, The Ocean Circulation Inverse Problem. Cambridge University Press, 1996.
[210] M. G. Shirangi, "History matching production data and uncertainty assessment with an efficient TSVD parameterization algorithm," Journal of Petroleum Science and Engineering, vol. 113, pp. 54-71, 2014.
[211] S. Mak et al., "An efficient surrogate model for emulation and physics extraction of large eddy simulations," Journal of the American Statistical Association, vol. 113, no. 524, pp. 1443-1456, 2018.
[212] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker, "Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments," Journal of the American Statistical Association, vol. 86, no. 416, pp. 953-963, 1991.
[213] P. S. Craig, M. Goldstein, J. C. Rougier, and A. H. Seheult, "Bayesian forecasting for complex systems using computer simulators," Journal of the American Statistical Association, vol. 96, no. 454, pp. 717-729, 2001.
[214] M. J. Bayarri et al., "A framework for validation of computer models," Technometrics, vol. 49, no. 2, pp. 138-154, 2007.
[215] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, "Computer model calibration using high-dimensional output," Journal of the American Statistical Association, vol. 103, no. 482, pp. 570-583, 2008.
[216] K. Kisamori, M. Kanagawa, and K. Yamazaki, "Simulator calibration under covariate shift with kernels," in International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1244-1253.
[217] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate Bayesian computational methods," Statistics and Computing, vol. 22, no. 6, pp. 1167-1180, 2012.
[218] C. P. Robert, "Approximate Bayesian computation: A survey on recent results," in Monte Carlo and Quasi-Monte Carlo Methods, Springer, 2016, pp. 185-205.
[219] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré, "Markov chain Monte Carlo without likelihoods," Proceedings of the National Academy of Sciences, vol. 100, no. 26, pp. 15 324$15328,2003$.
[220] C. P. Robert, J.-M. Cornuet, J.-M. Marin, and N. S. Pillai, "Lack of confidence in approximate Bayesian computation model choice," Proceedings of the National Academy of Sciences, vol. 108, no. 37, pp. 15 112-15 117, 2011.
[221] E. Tamer, "Partial identification in econometrics," Annual Review of Economics, vol. 2, no. 1, pp. 167-195, 2010.
[222] V. Chernozhukov, H. Hong, and E. Tamer, "Estimation and confidence regions for parameter sets in econometric models," Econometrica, vol. 75, no. 5, pp. 1243-1284, 2007.
[223] C. F. Manski, Partial Identification of Probability Distributions. Springer, 2003, vol. 5.
[224] A. Lewbel, "The identification zoo: Meanings of identification in econometrics," Journal of Economic Literature, vol. 57, no. 4, pp. 835-903, 2019.
[225] S. Alonso-Monsalve and L. H. Whitehead, "Image-based model parameter optimization using model-assisted generative adversarial networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 12, pp. 5645-5650, 2020.
[226] F. Lamperti, A. Roventini, and A. Sani, "Agent-based model calibration using machine learning surrogates," Journal of Economic Dynamics and Control, vol. 90, pp. 366-389, 2018.
[227] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, Robust Optimization. Princeton University Press, 2009.
[228] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," SIAM Review, vol. 53, no. 3, pp. 464-501, 2011.
[229] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," Mathematical Programming, vol. 171, no. 1-2, pp. 217-282, 2018.
[230] I. Petersen, M. James, and P. Dupuis, "Minimax optimal control of stochastic uncertain systems with relative entropy constraints," IEEE Transactions on Automatic Control, vol. 45, no. 3, pp. 398-412, 2000.
[231] G. Bayraksan and D. K. Love, "Data-driven stochastic programming using phi-divergences," in Tutorials in Operations Research, INFORMS, 2015, pp. 1-19.
[232] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," Mathematical Programming, vol. 171, no. 1-2, pp. 115-166, 2018.
[233] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," Mathematics of Operations Research, vol. 44, no. 2, pp. 565-600, 2019.
[234] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," Mathematics of Operations Research, 2022.
[235] L. E. Ghaoui, M. Oks, and F. Oustry, "Worst-case value-at-risk and robust portfolio optimization: A conic programming approach," Operations research, vol. 51, no. 4, pp. 543556, 2003.
[236] I. Popescu, "A semidefinite programming approach to optimal-moment bounds for convex classes of distributions," Mathematics of Operations Research, vol. 30, no. 3, pp. 632-657, 2005.
[237] B. Li, R. Jiang, and J. L. Mathieu, "Ambiguous risk constraints with moment and unimodality information," Mathematical Programming, pp. 1-42, 2017.
[238] B. P. Van Parys, P. J. Goulart, and D. Kuhn, "Generalized Gauss inequalities via semidefinite programming," Mathematical Programming, vol. 156, no. 1-2, pp. 271-302, 2016.
$[239]$ B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, "Nonparametric estimation of integral probability metrics," in 2010 IEEE International Symposium on Information Theory, IEEE, 2010, pp. 1428-1432.
[240] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings, 2012, pp. 37-49.
[241] I. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014.
[242] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Proceedings of the 34th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 214-223.
[243] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," arXiv preprint arXiv:1701.04862, 2017.
[244] M. Zhang, M. Gong, Y. Mao, J. Li, and Y. Wu, "Unsupervised feature extraction in hyperspectral images based on Wasserstein generative adversarial network," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 5, pp. 2669-2688, 2019.
[245] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 11, pp. 2092-2096, 2017.
[246] M. Naaman, "On the tight constant in the multivariate Dvoretzky-Kiefer-Wolfowitz inequality," Statistics \& Probability Letters, vol. 173, p. 109 088, 2021.
[247] W. Feller, "On the Kolmogorov-Smirnov limit theorems for empirical distributions," The annals of mathematical statistics, vol. 19, no. 2, pp. 177-189, 1948.
[248] D. Byrd, "Explaining agent-based financial market simulation," arXiv preprint arXiv:1909.11650, 2019.
[249] J.-P. Bouchaud, M. Mezard, and M. Potters, "Statistical properties of stock order books: Empirical results and models," Quantitative Finance, vol. 2, p. 301, Apr. 2002.
[250] R. Cont, "Empirical properties of asset returns: Stylized facts and statistical issues," Quantitative Finance, vol. 1, no. 2, pp. 223-236, 2001.
[251] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, Trades, Quotes and Prices: Financial Markets under the Microscope. Cambridge University Press, 2018.
[252] M. R. Kosorok, Introduction to Empirical Processes and Semiparametric Inference. Springer, 2008.
[253] F. Wei and R. M. Dudley, "Two-sample Dvoretzky-Kiefer-Wolfowitz inequalities," Statistics \& Probability Letters, vol. 82, no. 3, pp. 636-644, 2012.
[254] L. Crespo and S. Kenny, "The NASA Langley Challenge on Optimization under Uncertainty," ESREL, 2020.
[255] A. Ben-Tal and A. Nemirovski, "Robust optimization-methodology and applications," Mathematical Programming, vol. 92, no. 3, pp. 453-480, 2002.
[256] D. Bertsimas, V. Gupta, and N. Kallus, "Data-driven robust optimization," Mathematical Programming, vol. 167, no. 2, pp. 235-292, 2018.
[257] L. J. Hong, Z. Huang, and H. Lam, "Learning-based robust optimization: Procedures and statistical guarantees," Management Science, 2020.
[258] J. L. Doob, "Heuristic approach to the Kolmogorov-Smirnov theorems," The Annals of Mathematical Statistics, pp. 393-403, 1949.
[259] M. Kac, "On deviations between theoretical and empirical distributions," Proceedings of the National Academy of Sciences of the United States of America, vol. 35, no. 5, p. 252, 1949.
[260] A. K. Ramdas, R. F. Barber, M. J. Wainwright, and M. I. Jordan, "A unified treatment of multiple testing with prior knowledge using the p-filter," Annals of Statistics, vol. 47, no. 5, pp. 2790-2821, 2019.
[261] V. Vovk and R. Wang, "Combining p-values via averaging," Biometrika, vol. 107, no. 4, pp. 791-808, 2020.
[262] J. Blanchet, H. Lam, Y. Liu, and R. Wang, "Convolution bounds on quantile aggregation," arXiv preprint arXiv:2007.09320, 2020.
[263] B. Kraan and T. Bedford, "Probabilistic inversion of expert judgments in the quantification of model uncertainty," Management Science, vol. 51, no. 6, pp. 995-1006, 2005.
[264] R. G. Sargent, "Verification and validation of simulation models," in Proceedings of the 2010 winter simulation conference, IEEE, 2010, pp. 166-183.
[265] H. Lam and E. Zhou, "The empirical likelihood approach to quantifying uncertainty in sample average approximation," Operations Research Letters, vol. 45, no. 4, pp. 301-307, 2017.
[266] H. Lam, "Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization," Operations Research, vol. 67, no. 4, pp. 1090-1105, 2019.
[267] J. C. Duchi, P. W. Glynn, and H. Namkoong, "Statistics of robust optimization: A generalized empirical likelihood approach," Mathematics of Operations Research, vol. 46, no. 3, pp. 946-969, 2021.
[268] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," Journal of Applied Probability, vol. 56, no. 3, pp. 830-857, 2019.
[269] V. Gupta, "Near-optimal Bayesian ambiguity sets for distributionally robust optimization," Management Science, vol. 65, no. 9, pp. 4242-4260, 2019.
[270] H. Lam and H. Qian, "Combating conservativeness in data-driven optimization under uncertainty: A solution path approach," arXiv preprint arXiv:1909.06477, 2019.
[271] J. S. Liu, Monte Carlo Strategies in Scientific Computing. Springer Science \& Business Media, 2008.
[272] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2009.
[273] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," Advances in neural information processing systems, vol. 20, 2007.
[274] D. Precup, "Eligibility traces for off-policy policy evaluation," Computer Science Department Faculty Publication Series, p. 80, 2000.
[275] M. Schlegel, W. Chung, D. Graves, J. Qian, and M. White, "Importance resampling for offpolicy prediction," in Advances in Neural Information Processing Systems, 2019, pp. 17991809.
[276] M. C. Fu, Handbook of Simulation Optimization. Springer, 2015, vol. 216.
[277] N. Jian and S. G. Henderson, "An introduction to simulation optimization," in 2015 winter simulation conference (WSC), IEEE, 2015, pp. 1780-1794.
[278] H. Kushner and G. G. Yin, Stochastic Approximation and Recursive Algorithms and Applications. Springer Science \& Business Media, 2003, vol. 35.
[279] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," SIAM Journal on Optimization, vol. 23, no. 4, pp. 2341-2368, 2013.

